THE HONG KONG
POLYTECHNIC UNIVERSITY
香港理工大學
Pao Yue-kong Library
包玉剛圖書館

Copyright Undertaking

This thesis is protected by copyright, with all rights reserved.

By reading and using the thesis, the reader understands and agrees to the following terms:

1. The reader will abide by the rules and legal ordinances governing copyright regarding the use of the thesis.
2. The reader will use the thesis for the purpose of research or private study only and not for distribution or further reproduction or any other purpose.
3. The reader agrees to indemnify and hold the University harmless from and against any loss, damage, cost, liability or expenses arising from copyright infringement or unauthorized usage.

IMPORTANT

If you have reasons to believe that any materials in this thesis are deemed not suitable to be distributed in this form, or a copyright owner having difficulty with the material being included in our database, please contact lbsys@polyu.edu.hk providing details. The Library will look into your claim and consider taking remedial action upon receipt of the written requests.

Pao Yue-kong Library, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

http://www.lib.polyu.edu.hk

MODEL FOR ZERO-INFLATED PROPORTION DATA ANALYSIS

YANGZI ZHENG

PhD

The Hong Kong Polytechnic University

2024

The Hong Kong Polytechnic University

Department of Applied Mathematics

Model for Zero-inflated Proportion Data Analysis

Yangzi Zheng

A thesis submitted in partial fulfilment of the requirements

for the degree of Doctor of Philosophy

JUNE 2024

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

______________________________(Signed)

_________Yangzi Zheng_________(Name of student)

Abstract

The examination and interpretation of datasets containing a substantial number of zeros have become increasingly relevant across various disciplines, including ecology and sociological studies. While there has been extensive research on zero-inflated count data, models specifically designed for proportion data with a high occurrence of zeros remain relatively limited. This thesis addresses this gap by focusing on zero-inflated proportion data and proposing a novel modeling approach to distinguish between two types of zeros present in the dataset. The primary objective is to develop a regression model that can effectively capture and differentiate these two types of zeros. The first type of zero, which corresponds to random absence, is modeled using a binomial sampling approach. This accounts for instances where the proportion value is zero due to random factors or chance. The second type of zero, arising from unsuitability, is handled using a general classification indicator. This indicator helps identify situations where the proportion value is zero due to the unsuitability of certain conditions or factors. To achieve our objective, we propose both parametric and semi-parametric models, providing flexibility and robustness in capturing the characteristics of the zero-inflated proportion data. By introducing these innovative models, we aim to enhance the understanding and analysis of datasets with a high occurrence of zeros. This research contributes to the development of methodologies specifically tailored for zero-inflated proportion data, addressing a significant gap in the existing literature.

In the first section of our study, we focus on investigating a semi-parametric model. This model comprises two components: a regression component that incorporates weighted least squares to account for heterogeneity, and a classification component that benefits from an optimal decision rule derived from our model. To estimate the parameters based on the optimal decision rule, we employ the Nadaraya-Watson estimator. This estimator ensures the accuracy of our classification and contributes to the overall robustness of the model. The results of our investigation reveal that environmental features play a crucial role in understanding both types of zeros: those related to perfection and those resulting from random absence. By utilizing our proposed modeling approach, researchers can gain deeper insights into the factors that contribute to these different types of zeros, thereby improving their understanding of the underlying processes. Furthermore, our model demonstrates superior performance in both simulated and real-world scenarios when compared to traditional methods such as the Tobit model and the zero-inflated beta regression model. By significantly reducing prediction errors, our model is proven to be a valuable tool for accurate estimation and prediction in various applications. By presenting these findings, we highlight the effectiveness and practicality of our semi-parametric model, enabling researchers to make more informed decisions and gain a comprehensive understanding of the factors influencing both types of zeros and the positive percent rate.

In the second section, our main objective is to provide a precise interpretation of the factors that influence the defective rate. Particularly, we focus on the indicator part, which was left undefined in the first part but has garnered more attention due to its exploration of the covariates that distinguish the zero part from the non-zero part. In the original model assumption, the presence of the indicator part creates complexity in inferring the parameters. Taking inspiration from the smoothed maximum score estimator, we introduce a parametric model by replacing the indicator part with a smoothed kernel estimator. This substitution yields a continuously differentiable loss function, which greatly facilitates further analysis. Similar to the previous section, we take into account heterogeneity and utilize the weighted least square method to estimate both parameters. Subsequently, we establish the consistency and asymptotically normal properties for both the regression and indicator estimators. These properties assure the reliability and validity of our estimators in capturing the underlying relationships and distinguishing between the zero and non-zero parts effectively.

Keywords: Semiparametric model; Semiparametric estimation; Weighted least squares; Parametric inference; Zero-inflated proportion data

Acknowledgements

First and foremost, I would like to express my profound gratitude and sincere appreciation to my esteemed supervisor, Professor Zhao Xingqiu. Her unwavering guidance, invaluable insights, and continuous support have been the beacons that illuminated my path throughout this doctoral journey. Professor Zhao's profound expertise, unwavering patience, and steadfast dedication have been instrumental in shaping my research endeavors and helping me navigate the numerous challenges encountered along the way. I am truly fortunate and privileged to have had the opportunity to work under her outstanding mentorship.

I am also deeply indebted to my co-supervisor, Dr. Jiang Binyan, whose insightful feedback, critical perspectives, and astute observations have greatly enriched my research pursuits. Dr. Jiang's meticulous attention to detail and unwavering commitment to excellence have challenged me to think beyond conventional boundaries and continuously strive for higher standards, pushing the boundaries of my intellectual capabilities.

Furthermore, I would like to express my heartfelt gratitude to my fellow doctoral colleagues and cherished friends, whose camaraderie, encouragement, and stimulating intellectual discourse have made this journey more enjoyable and fulfilling. Our shared experiences, thought-provoking exchanges, and mutual support have been instrumental in overcoming obstacles and celebrating milestones, fostering an environment conducive to personal and professional growth.

A special note of appreciation goes to my beloved family, whose unconditional love, unwavering belief, and constant encouragement have been the driving force behind my perseverance. Their understanding and support during the demanding times have been invaluable, and I am forever grateful for their sacrifices and steadfast faith in my abilities. Their presence has been a constant source of strength, motivating me to strive for excellence and achieve my aspirations. Finally, I would like to quote the words from Once Upon a Time in America to express my gratitude to a special person Dear, who lit me up in the darkness, ‘There were times I couldn’t stand it any more. I used to think of you. Thinking on you in a place to live in the world, exist, and that would get me through it all. You are that important to me.’ Wish one day we can meet at the summit.

Contents

List of Figures

List of Tables

List of Notations

$\mathcal{R}$	set of real numbers
$\mathcal{R}^n$	set of n-dimensional real vectors
x	scale
$\mathbf{X}$	$(p-1)$-dimensional vector
$\mathbf{X}^\top$	the transpose of vector $\mathbf{X}$
$K(\cdot)$	standard Gaussian kernel function
$\|x\|_2$	the Euclidean norm of x
$\|Z_n\|_{\mathcal{F}}(x)$	the maximum value when $x \in \mathcal{F}$
$\text{sign}(x)$	the signum function of x
$\mathbf{I}(\cdot)$	the indicator function taking values (0,1).
$\text{card}(A)$	number of elements in A
$B(x, \delta)$	the ball centered at x with radius δ
$d(x, X)$	smallest distance between the vector x and the set X
$\lambda_{\min}(A)$	the smallest eigenvalue of symmetric matrix A
$\lambda_{\max}(A)$	the largest eigenvalue of symmetric matrix A
$\longrightarrow_p$	convergence in probability
$a_n = o_p(1)$	$a_n \xrightarrow{p} 0$
$a_n = o(1)$	a_n converges to 0

$a_n = O(1)$	a_n is bounded
$\Rightarrow$	asymptotic distribution

Chapter 1

Literature Review and Introduction

1.1 Literature Review and Background

Zero-inflated nonnegative data analysis has been extensively studied for many years due to its widespread applications in biomedical, economic, and ecological domains. Its applications encompass substance abuse, medical costs, medical care utilization, single-cell gene expression rates, and relative abundance of microbiomes. In all these cases, a substantial portion of the data sets comprises zero values combined with positive continuous values, which cannot be adequatel y explained by simple parametric distributions. To precisely capture the characteristics of the variables of interest in different situations, researchers have developed specific models. For instance, in the microbiome field, some zero observations do not necessarily represent actual zero values. In other words, there may exist a small $y_{\min}$ value, below which observations cannot be detected. In such cases, researchers have employed censoring models to describe the occurrence of zeros. However, in cases like alcohol consumption or medical costs, zero-valued observations represent clinically significant actual zeros, prompting researchers to introduce an additional indicator variable for the zero part. Two classical methods are most commonly used to deal with inflated data, we will briefly introduce them and their extensions on high-dimensional data sets.

1.1.1 Zero-Inflated Nonnegative Data Analysis

Tobit Model

The Tobit model, first introduced by Tobin (1958), is a crucial model in economic analysis, particularly in the study of labor market outcomes and wage determination. In this model, zero values are considered as "censored" observations, accounting for the fact that some individuals may have unobserved or unreported wages due to factors such as unemployment or non-participation in the labor market. Utilizing the Tobit model, the influential study conducted by Heckman (1979) has significantly contributed to our understanding of wage determination. While the Tobit model has been proven to be an invaluable tool in econometrics, it has also been applied to the statistical modeling of zero-inflated nonnegative data collected in other fields (Liu et al., 2019).

Denote Y^* as the latent real outcome which is continuous and positive. Given a detection limit y_{min}, the Tobit model first assumed the actual observation $Y = Y^* \mathrm{I}_{\{\mathrm{Y}^* > \mathrm{y}_{\min}\}}$. An observation of $Y = 0$ is then an indicator of left censoring. Given the covariates $\mathbf{X}$, the Tobit model is defined by modeling $\log Y^*|\mathbf{X} = \mathbf{X}^\top \alpha + \epsilon$, where ϵ is a Gaussian noise term, and inference approaches were well developed, including the maximum likelihood method proposed by Amemiya (1973), the Bayesian method brought up by Chib (1992) and the maximum entropy method presented by Golan et al. (1997).

Despite the well-construction of the framework of the Tobit model, it is important to note that the Tobit model does not account for heteroscedasticity, meaning that the variance of ϵ is assumed to be independent of $\mathbf{X}$. This assumption may not hold in many practical situations, leading to potential issues in model fitting and inference.

To address this limitation, the heteroscedastic Tobit model could be regarded as one promising extension, which allows the variance of ϵ to depend on the covariates $\mathbf{X}$. Specifically, the model assumes $\log Y^*|\mathbf{X} = \mathbf{X}^\top \alpha + \epsilon(\mathbf{X})$, where $\epsilon(\mathbf{X})$ is a function that captures the heteroscedasticity in the error term. This extension provides a more flexible framework for modeling zero-inflated data with varying variances across different covariate values.

Another extension is the censored quantile regression model, which aims to estimate the conditional quantiles of the response variable rather than just the conditional mean. This approach is particularly useful when the distribution of the errors is heavy-tailed or when robustness to outliers is desired.

By accounting for heteroscedasticity and leveraging more flexible modeling frameworks, these extensions of the Tobit model offer improved accuracy and robustness in analyzing zero-inflated nonnegative data, especially in scenarios where the assumptions of the standard Tobit model are violated.

In addition, Jacobson and Zou (2022) proposed penalized Tobit models for high-dimensional censored regression problems. They employed a convex reparameterization of the negative log-likelihood function, as introduced by Olsen (1978), to leverage convex optimization techniques.In theoretical results, they derived a bound for the l_2 estimation error of the Tobit Lasso estimator, which provides valuable insights into the behavior of the Tobit Lasso estimator in high dimensions and facilitates the development of robust and accurate estimation procedures for censored regression problems with a large number of predictors.

Two Part Model

Unlike the Tobit model, which considered the data set to be left-censored, two-part model took zero values as true observations as it introduced in Liu et al. (2019), Manning et al. (1981). In other words, the observations were separated into two parts, the zero part and the positive part. Let Y_0 be the Bernoulli random variable such that

$$\mathrm{logitP(Y_0 = 1|}\mathbf{X}) = \mathbf{X}^\top \alpha.$$

The let $Y_+ > 0$ be the continuous random variable such that

$$\mathrm{logY_+|}\mathbf{X} \sim \mathbf{X}^\top \beta + \mathrm{e},$$

where $\mathbf{X}$ denoted as the covariates and it is independent with e. Then the observation data set with a large portion of zero Y could be modeled as $Y = Y_0 Y_+$.

Analysis of this model could also be specified in two parts:

1. $P(Y > 0|\mathbf{X}) = p(\mathbf{X})$ and $P(Y = 0|\mathbf{X}) = 1 - p(\mathbf{X})$, where $p(\mathbf{X}) = \frac{\exp(\mathbf{X}^\top \alpha)}{1+\exp(\mathbf{X}^\top \alpha)}$.

2. For the positive part, $\mathbf{E}(\mathrm{logY|Y} > 0, \mathbf{X}) = \mathbf{E}(\mathrm{logY_+|}\mathbf{X}) = \mathbf{X}^\top \beta$. We could also conclude the cumulative distribution function under two-part model as

$$P(Y \leq y|\mathbf{X}) = 1 - p(\mathbf{X}) + \mathrm{I(y > 0)p(}\mathbf{X})\mathrm{F_e(logy} - \mathbf{X}^\top \beta),$$

where $F_e(v) = P(e \leq v)$.

Since the interpretation of zero part is quite different between two part model and Tobit model, it is not appropriate to decide which is better in general. They were more like answers for same problem occurred in distinct areas.

Zero-inflated Tobit Model

Since the Tobit model is designed to handle left-censored data, it encounters challenges when the data consists of a substantial proportion of zeros, resulting in increased variance and a poor fit of the model. To overcome this obstacle, the zero-inflated Tobit model is proposed to address this issue by adding an additional point mass at zero (Moulton and Halsey, 1995; Liu et al., 2019). Specifically, the zero-inflated Tobit model is given as:

$$Y = \begin{cases} Y^*\mathrm{I}(Y^* > y_{min}) & \text{with probability } p(\mathbf{X}), \\ 0 & \text{with probability } 1 - p(\mathbf{X}), \end{cases}$$

where $\mathrm{I}(\cdot)$ is the indicator function and the parameter y_{min} refers to the latent minimal detection limit. A natural extension of the zero-inflated Tobit model to deal with the case where the nonzero component of the response Y in within $(0, 1)$ is to adopt a different link function that extends the domain of Y to the real line, and as a comparison, in this paper we consider the model $\mathrm{logitY}^*|\mathbf{X} = \mathbf{X}^\top\alpha + \epsilon$.

1.1.2 Zero-inflated Proportional Data Analysis

Proportional data is a common type of observation in various fields such as animal studies, economic studies, environmental studies, and industrial manufacturing. According to Warton and Hui (2011), over a third of publications in ecology analyzed some form of proportional data. This data type arises in diverse contexts within these fields. For instance, in ecology, researchers study the proportional cover of specific plant functional types in vegetation quadrat surveys (Defries et al. (2000)), the proportion of time animals spend engaged in certain activities (Clayton and Cotgreave (1994)), and the percentages of biomass allocated to different plant organs (Poorter et al. (2012)).

Denote the n independent observations as $(\mathbf{X}_\mathbf{i}^*, Y_i), i = 1, \dots, n$ where the $Y_i \in [0, 1]$ is the proportional response and $\mathbf{X}_i^* = (1, \mathbf{X}_i^\top)$, $\mathbf{X}^\top = (x_{11}, \dots, x_{1,p-1})^\mathrm{T}$ is the $(p - 1)$-dimensional covariate of observation i. For proportional data without zero inflation, binomial regression is a standard model which is frequently used in practice (Quinn and Keough (2002)). Alternatively, we can also apply appropriate transformations to the proportions and proceed with classical ordinary linear models (Crawley (2012)). However, in many applications, the proportional responses are also characterized by a significant presence of zero values, i.e., some of the Y_i's are exactly equal to zero, and classical methods that ignore the inflation of zeros are no longer appropriate. To address the zero inflation in proportional data, a popular approach is to combine the β-distribution with a point mass (Swearingen et al., 2012), and one of the most popular methods in fitting zero-inflated proportional data is the zero-inflated β-regression model introduced by Ospina and Ferrari (2012). Specifically, Swearingen et al. (2012) assumes that the zero-inflated response Y follows:

$$Y = \begin{cases} 0 & \text{with probability p}, \\ \text{Beta}(\alpha_1, \alpha_2) & \textit{otherwise}, \end{cases}$$

where $\text{Beta}(\alpha_1, \alpha_2)$ is the density of the β-distribution with parameters (α_1, α_2). Given the covariate $\mathbf{X}^*$, the zero-inflated β-regression model (Ospina and Ferrari, 2012) formulates $\text{logit}(u) = \mathbf{X}^{*\top}\beta_1$, $\text{logit}(v) = \mathbf{X}^{*\top}\beta_2$, $\text{logit}(p) = \mathbf{X}^{*\top}\beta_3$, where $u = \frac{\alpha_1}{\alpha_1+\alpha_2}$ and $v = \frac{\alpha_1\alpha_2}{(\alpha_1+\alpha_2)^2(\alpha_1+\alpha_2+1)}$ refer to the mean and dispersion parameter of the β-distribution, respectively. Without loss of generality, the β-distribution can also be replaced by other distributions defined on the (0,1) interval such as the simplex distribution (Kieschnick and McCullough, 2003) and the generalized Johnson SB distribution (Queiroz and Lemonte, 2021). Bayesian versions of this framework have also been studied in Wieczorek and Hawala (2012), Santos and Bolfarine (2015a) and Liu and Eugenio (2018), among others. In particular, Liu and Eugenio (2018) compared the performance of the zero-inflated β-regression model among the frequentist likelihood-based method and Bayesian-based method, and found that while the two approaches are comparable, the likelihood-based approach was computationally more efficient, and the Bayesian inferences were less biased when the sample size was small. Overall, existing methods primarily focused on modeling the zero-inflated proportions via generalized linear models which rely on the distributional assumption for the zero-inflated proportional responses. While there are many application papers which use the above-mentioned approaches to analyze data in different areas, new methods for the statistical modeling, especially those directly addresses prediction of zero-inflated proportional data, have been relatively sparse.

1.1.3 Latest Research and Applications

There are many applications involving the zero-inflated proportion data such as the alcohol consumption among California students, the electricity accessibility in Brazil, and the infant mortality data obtained from the Parana State. We will illustrate them carefully the numerical study in chapter 2. Besides, a very classical example is mortality in traffic accidents of $n = 200$ Brazilian municipal districts of the southeast region in the year 2002. The response variable (Y) is the proportion of deaths caused by traffic accidents. The explanatory variables include: lnpop—the natural logarithm of the municipality's population size, prop2029—the proportion of residents aged between 20 and 29 years, and hdie—the human development index of education for the municipal district. Ospina and Ferrari (2012) and Queiroz and Lemonte (2021) used zero-inflated beta regression model and zero-or-one inflated generalized Johnson SB (GSB) regression models accordingly to examine the influence of the young population proportion on the proportion of traffic accident fatalities, after accounting for potential confounding factors.

In the analysis of metagenomic data, which is typically represented as compositions (proportions) with an excessive number of zeros and a skewed distribution, it is crucial to identify disease-associated pathogenic bacteria characterized by differential abundances across various clinical conditions. To address this challenge, Peng et al. (2016) introduced a zero-inflated beta regression approach, termed ZIBSeq, which accounts for the sparse nature of metagenomic data and enables more efficient modeling of compositional data.

In health economics for analysis of technical efficiency in hospitals, which is a measure of how well a hospital utilizes its resources to produce outputs or services. Since the technical efficiency scores are bounded between 0 and 1 and often exhibit excessive zeros (fully inefficient hospitals) and ones (fully efficient hospitals), Ocaña-Riola et al. (2021) propose a multilevel zero-one inflated beta regression model to investigate the relationship between exogenous health variables and the technical efficiency of hospitals within the Spanish National Health System.

1.1.4 Technical Tools

Uniform Convergence

Modes of convergence in probability theory, statistics, and related fields play a crucial role in understanding the behavior of sequences of probability measures, estimators, and other stochastic processes. Two important modes of convergence are uniform convergence and weak convergence (convergence in distribution). Uniform convergence is a stronger notion of convergence compared to pointwise convergence. An estimator is said to converge uniformly to the true parameter if the maximum deviation between the estimator and the true parameter over the entire parameter space goes to zero as the sample size increases. This property is desirable because it ensures that the estimator is consistently accurate across the entire range of parameter values, rather than just converging at specific points. Weak convergence of measures, also known as convergence in distribution, is an important concept in probability theory and stochastic processes. It describes the convergence of a sequence of probability measures in terms of their behavior on continuous bounded functions. The research area described in Rao (1962) deals with studying the connections and implications between different modes of convergence for sequences of measures. Specifically, it focuses on the relationships between weak convergence and uniform convergence of measures, and how these modes of convergence interact with each other. Understanding the relationships between weak and uniform convergence is crucial because it allows for translating results and implications from one mode of convergence to the other. A representative result brought up in Rao (1962) is following theorem:

Lemma 1.1. *Let $\{\lambda_n, \lambda; n = 1, 2, \cdots\}$ be a sequence of random measures on X possessing the ergodic property. Let a be a family of continuous functions on X satisfying the following two conditions: (i) there exists a continuous function $g(x)$ on X such that $|f(x)| \leqq g(x)$ for each $f \in \mathbb{Q}$ and $x \in X; E\int g(x)\lambda(dx, \omega) < \infty$; and (iii) a is equicontinuous. Then $P\left[\eta_n \to 0\right] = 1$, where*

$$\eta_n = \sup_{fea} \left| \int f d\lambda_n - \int f d\lambda \right|.$$

A General Upper Bound for Empirical Risk Minimizers

Let us consider independent random variables $\eta_1, \ldots, \eta_n$ observed in a measurable space $\mathcal{Z}$, with a common distribution P. In the context of bounded regression, for every i, the variable $\eta_i = (X_i, Y_i)$ is a copy of a pair of random variables (X, Y), where X takes values in a measurable space $\mathcal{X}$, and Y is assumed to take values in the interval $[0, 1]$. In the classification case, the response variable Y is assumed to belong to the set $0, 1$. The regression function ξ is defined as $\xi(x) = \mathbf{E}[Y|X = x]$ for every $x \in \mathcal{X}$.

In regression problems, the goal is to estimate the regression function ξ. One commonly used method for this purpose is empirical risk minimization (Vapnik, 2006). This method involves considering a set $\mathcal{S}$, which is known to contain η. In the regression case, $\mathcal{S}$ can be the set of all measurable functions from $\mathcal{X}$ to $[0, 1]$.

The key element in empirical risk minimization is the introduction of a loss (or contrast) function γ from $\mathcal{S} \times \mathcal{Z}$ to $[0, 1]$, which is well-adapted to the problem of estimating η. The expected loss $P[\gamma(t, \cdot)]$ achieves a minimum at η when t varies in $\mathcal{S}$. In other words, the relative expected loss ℓ defined by $\ell(\xi, t) = P[\gamma(t, \cdot) - \gamma(\xi, \cdot)]$

for all $t \in \mathcal{S}$ is non-negative. In the regression case, a common choice for γ is $\gamma(t,(x,y)) = (y - t(x))^2$, since η minimizes $\mathbf{E}[(Y - t(X))^2]$ over measurable functions t taking values in $[0,1]$.

The heuristics of empirical risk minimization (or minimum contrast estimation) can be described as follows: Substitute the empirical loss $\gamma_n(t) = P_n[\gamma(t,\cdot)] = \frac{1}{n}\sum_{i=1}^{n}\gamma(t,\eta_i)$ for its expectation $P[\gamma(t,\cdot)]$, and minimize γ_n on a subset S of $\mathcal{S}$. This provides an estimator $\hat{s}$ of ξ, which is sensible if ξ belongs (or is close enough) to the model S. This estimation method is widely used and has been extensively studied in the asymptotic parametric setting, where S is a given parametric model, ξ belongs to S, and n is large.

The purpose is to provide a general non-asymptotic upper bound for the relative expected loss between $\hat{s}$ and η. To achieve this, we introduce the centered empirical process $\bar{\gamma}_n$ defined by $\bar{\gamma}_n(t) = \gamma_n(t) - P[\gamma(t,\cdot)]$.

In addition to the relative expected loss function ℓ, another way to measure the closeness between elements of S is needed, directly connected to the variance of the increments of $\bar{\gamma}_n$, which plays an important role in analyzing $\bar{\gamma}_n$'s fluctuations. Let d be a pseudo-distance on $\mathcal{S} \times \mathcal{S}$ (which may depend on the unknown distribution P) such that $\mathrm{Var}_P[\gamma(t,\cdot) - \gamma(\xi,\cdot)] \le d^2(\xi,t)$ for every $t \in \mathcal{S}$.

In applications, it may be more convenient to take d as a more intrinsic distance. For instance, in regression or classification, d can be chosen (up to a constant) as the $L_2(\mu)$ distance, where μ denotes the distribution of X. For regression, $[\gamma(t,(x,y)) - \gamma(\eta,(x,y))]^2 = [t(x) - \xi(x)]^2[2(y - \xi(x)) - t(x) + \xi(x)]^2$. Since $\mathbf{E}_P[Y - \xi(X)|X] = 0$ and $\mathbf{E}_P[(Y - \xi(X))^2|X] \le 1/4$, we derive that $\mathbf{E}_P[\gamma(t,(X,Y)) - \gamma(\xi,(X,Y))]^2 \le 2\mathbf{E}\mu(t(X) - \xi(X))^2$.

Our main result will crucially depend on two different modulus of uniform continuity: the stochastic modulus of uniform continuity of $\bar{\gamma}_n$ over S with respect to d, and the modulus of uniform continuity of d with respect to ℓ.

The main tool we shall use is Talagrand's inequality for empirical processes (Talagrand, 1996), which will allow us to control the oscillations of the empirical process $\bar{\gamma}_n$ by the modulus of uniform continuity of $\bar{\gamma}_n$ in expectation. More precisely, we shall use the following version of Talagrand's inequality due to Bousquet (2002), which has the advantage of providing explicit constants and dealing with one-sided suprema.

Let $\mathcal{F}$ be a countable family of measurable functions such that, for some positive constants z and b, one has, for every $f \in \mathcal{F}$, $P(f^2) \leq z$ and $|f|\infty \leq b$. Then, for every positive y, the following inequality holds for $T = \sup f \in \mathcal{F}(P_n - P)(f)$:

$$\mathbb{P}\left[T - \mathbb{E}[T] \geq \sqrt{2\frac{(z + 4b\mathbb{E}[T])y}{n}} + \frac{2by}{3n}\right] \leq e^{-y}.$$

Unlike McDiarmid's inequality (McDiarmid, 1989), which has been widely used in statistical learning theory (Lugosi, 2002), a concentration inequality like Talagrand's inequality offers the possibility of controlling the empirical process locally.

By using Talagrand's inequality, we can control the oscillations of the empirical process $\bar{\gamma}_n$ by its modulus of uniform continuity in expectation. This is crucial for our main result, which depends on the moduli of uniform continuity of $\bar{\gamma}_n$ over S with respect to d, and the modulus of uniform continuity of d with respect to ℓ.

Kernel Estimation

The concept of kernel estimation was initially proposed by Rosenblatt (1956) in 1956 for density estimation. Subsequently, Watson (1964) and Nadaraya (1964) independently introduced kernel estimation as a novel approach for nonparametric regression estimation in 1964. Over the following decades, kernel estimation techniques gained widespread recognition and were extensively studied and refined by researchers across various disciplines, including statistics, econometrics, and signal processing. Notable contributions to the advancement of kernel estimation include the seminal work of Silverman (2018) on kernel density estimation, which provided a comprehensive treatment of the theory and applications of this technique. Härdle (1990) on applied nonparametric regression played a pivotal role in popularizing kernel regression methods and their practical applications. Fan (2018) made significant contributions to the field with their work on local polynomial modeling, which encompasses kernel regression as a special case.

The basic idea behind kernel estimation is to construct an estimate of the density function by summing up kernel functions centered at each data point. The kernel function is a symmetric, non-negative function that integrates to one and satisfies certain smoothness properties.

Mathematically, given a sample of observations X_1, X_2, $\cdots$, X_n drawn from an unknown density function $f(x)$, the kernel density estimator is defined as

$$\widehat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h\left(x - x_i\right) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right),$$

Where: $K(\cdot)$ is the kernel function h is the bandwidth or smoothing parameter that controls the trade-off between bias and variance of the estimator. As one of the widely used kernel estimation methods, the Nadaraya–Watson estimator, is formed as follows.

Let the data be (y_i, X_i) where y_i is real-valued and X_i is a p-dimension vector, and assume that all are continuously distributed with a joint density $f(y, x)$. The regression function for y_i on X_i is

$$g(x) = \mathbf{E}\left(y_i | X_i = x\right).$$

To estimate this nonparametrically with fewer assumptions on $g(\cdot)$, they consider a neighborhood around the point x. If the random variable X_i has a positive density at x, observations could be found within this neighborhood. The key is to estimate $g(x)$ by averaging the y_i values of observations in this neighborhood. However, the size of the neighborhood must balance two factors: the variation in $g(x)$ within the neighborhood (bias) and the number of observations included (variance). A smaller neighborhood reduces bias but increases variance, while a larger neighborhood reduces variance but increases bias. Finding the optimal neighborhood size (h) is crucial for accurate estimation. Based on the discussions, the Nadaraya–Watson estimator is constructed as

$$\hat{g}(x)=\frac{\sum_{i=1}^{n} K_h\left(d(X_i-x)\right) y_i}{\sum_{i=1}^{n} K_h\left(d(X_i-x)\right)},$$

where $K(\cdot)$ is a standard Gaussian kernel function, $K_h(u)=K\left(\frac{u}{h}\right)$, and $d(\cdot)$ refers to the distance measure.

Early works on uniform convergence for kernel density estimation are contributed by research such as Nussbaum (1996), Hardle et al. (1988), and Giné and Guillou (2002). Newey (1994) introduced kernel-based estimators for partial means and a general variance estimator in nonparametric regression settings. They established theoretical results on the consistency and asymptotic normality of the proposed estimators under certain regularity conditions. These findings provided a solid foundation for the use and interpretation of these estimators.

Let y denote a random variable and $p_0(x)=E[y|x]$. Partition $x=(x_1,x_2)$ and let $\tilde{x}_2$ be a variable that is included in z and has the same dimension as x_2, and $\bar{x}_1$ be some fixed value for x_1. Let $\tau\left(x_2\right)$ be some weight function that keeps a denominator bounded away from zero. A partial mean is

$$\beta_0=\mathbf{E}\left[\tau\left(\tilde{x}_2\right) p_0\left(\bar{x}_1,\tilde{x}_2\right)\right].$$

This object is an average over some conditioning variables holding others fixed. It can be estimated by substituting a kernel estimator for g_0 and a sample average for the expectation. Let $Y = (1, y)$, so that $f(x) = (f_1(x),\ f_2(x))'$ where $f_{10}(x)$ is the density of x and $f_{20}(x) = f_{10}(x)\mathbf{E}[y|x]$. Also, let $\hat{p}(x) = \hat{f}_2(x)/\hat{f}_1(x) = \hat{f}_2(x)/\hat{r}(x)$, for the kernel density estimator $\hat{r}(x) = \hat{f}_1(x)$, and $\bar{x}_i = (\bar{x}_1, \tilde{x}_{2i})$. Then the estimator is

$$\hat{\beta} = n^{-1} \sum_{i=1}^{n} \tau\left(\tilde{x}_{2i}\right) \hat{p}\left(\bar{x}_i\right).$$

Assumption 1.1. *There are positive integers d and q such that $K(u)$ is differentiable of order d, the derivatives of order d are Lipschitz, $K(u)$ is zero outside a bounded set, $\int K(u)du = 1$, and for all $j < q$, $\int K(u)\left[\otimes_{\ell=1}^{j} u\right] du = 0$.*

The last condition requires that the kernel should be a higher-order (bias-reducing) kernel of order q. This property is utilized to ensure that the limiting distributions of the estimators are centered around the true values, thereby guaranteeing their asymptotic unbiased. The next condition imposes smoothness on the functions $f_0(x) := \mathbf{E}[y|x]g_0(x)$, where $g_0(x)$ refers to the density of x.

Assumption 1.2. *There is a non-negative integer t and an extension of $f_0(x)$ to all of $\mathcal{R}^k$ that is continuously differentiable to order t on $\mathcal{R}^k$*

This condition is employed in conjunction with Assumption 1.1 to ensure that the bias of the estimator is sufficiently small. It mitigates cases where the density of x and its derivatives exhibit non-zero values at the boundary of the support by imposing smoothness requirements across the entire domain.

Under these conditions and certain others, partial means will be asymptotically normal. Let the u argument of $K(u)$ be partitioned conformably with x and $\tilde{g}_0\left(\tilde{x}_2\right)$ denote the true density of $\tilde{x}_2$. The asymptotic variance of the partial mean estimator will be

$$V = \left[\int \left\{ \int K\left(u_1, u_2\right) du_2 \right\}^2 du_1 \right]$$

$$\times \int g_0\left(\bar{x}_1, t\right)^{-1} \tau(w)^2 \tilde{g}_0(w)^2 \operatorname{var}\left(y \mid x = \left(\bar{x}_1, w\right)\right) dw.$$

Lemma 1.2. *Suppose that (i)* $\mathbf{E}\left[|y|^4\right] < \infty, \mathbf{E}\left[|y|^4|x\right] g_0(x)$*, and* $g_0(x)$ *are bounded; (ii) assumption 1.1 and assumption 1.2 are satisfied for* $t \geq q$*; (iii)* $\tau\left(\tilde{x}_2\right)$ *is bounded and zero except on a compact set where* $g_0\left(\bar{x}_1, \tilde{x}_2\right)$ *is bounded away from zero; (iv)* $\tau\left(\tilde{x}_2\right)$ *and* $\tilde{f}_0\left(\tilde{x}_2\right)$ *are continuous a.e.,* $\tilde{g}_0\left(\tilde{x}_2\right)$ *is bounded,* $\mathbf{E}[y|x]$ *and* $\mathbf{E}\left[y^2|x\right]$ *are continuous, and for some* $\varepsilon > 0$*,* $\int \sup_{|\eta| \leq \varepsilon} \left\{1 + \mathbf{E}\left[y^4 | x = \left(\bar{x}_1 + \eta, x_2\right)\right]\right\} g\left(\bar{x}_1 + \eta, x_2\right) dx_2 < \infty; (v) n\sigma^{2k-k_1} \ln(n)^2 \rightarrow \infty$ *and* $n\sigma^{k_1+2q} \rightarrow 0$*. Then, for* $\hat{\beta}$ *in 1.1.4,*

$$\sqrt{n}\sigma^{k_1/2}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} N(0, V).$$

Newey (1994) also provided rates of uniform convergence in probability for kernel estimators of derivatives, measured in Sobolev norms.

Proposition 1.1. *For a closed set* X*, Denote* $\|f\|_j = \sup_{\ell \leq j} \sup_{x \in x} \left\|\partial^\ell f(x) / \partial x^\ell\right\|$*. Suppose that* $\mathbf{E}\left[\|y\|^p\right] < \infty$ *for* $p > 2, \mathbf{E}\left[\|y\|^p | x\right] f_0(x)$ *is bounded,* x *is compact, Assumption 1.1 is satisfied for* $d \geq j$*, and* $\sigma = \sigma(n)$ *such that* $\sigma(n)$ *is bounded and* $n^{1-(2/p)} \sigma(n)^k / \ln(n) \rightarrow \infty$*. Then*

$$\|\hat{f} - \mathbf{E}[\hat{f}]\|_j = O_p\left(\ln(n)^{1/2}\left(n\sigma^{k+2j}\right)^{-1/2}\right).$$

These rates quantify the convergence behavior of the estimators for higher-order derivatives of the target function, offering a more comprehensive characterization of their performance. It contributed a lot in applications where accurate estimation of higher-order derivatives is crucial, such as in the study of nonparametric regression models, density estimation, and nonlinear time series analysis.

High Dimensional M-estimators

The study of high-dimensional M-estimators has its roots in the work on sparse estimation and variable selection in the early 2000s, with Tibshirani (1996) introducing the Lasso estimator, an M-estimator with an L_1 penalty, to perform variable selection and estimation simultaneously in high-dimensional linear regression models. Following the success of the Lasso, there has been a surge of research on high-dimensional M-estimation, focusing on developing new regularized M-estimators, studying their theoretical properties, and developing efficient computational algorithms. Key developments include: studying the consistency, oracle properties, and asymptotic distributions of high-dimensional M-estimators under various sparsity and regularity conditions (e.g., Bühlmann and Van De Geer (2011); Negahban et al. (2012)); growing interest in non-convex penalties, such as SCAD and MCP, which can improve estimation accuracy and variable selection performance (e.g., Fan and Li (2001); Zhang (2010)); extending high-dimensional M-estimation to robust settings with outliers or heavy-tailed errors (e.g., Sun et al. (2020)); developing efficient algorithms like coordinate descent, proximal gradient methods, and ADMM for solving high-dimensional M-estimation problems (e.g., Friedman et al. (2008); Beck and Teboulle (2009)); and applications in fields like genomics, finance, signal processing, and machine learning with high-dimensional data. The study of high-dimensional M-estimators has been a fertile area, bridging theory and computation, leading to new insights and methodologies for analyzing high-dimensional data.

In the high-dimensional statistical regime, where the number of parameters d grows to infinity as the sample size n increases, estimating a parameter $\theta_0 \in \Theta \subset \mathbb{R}^d$ could be a critical problem. For simplicity, Θ is assumed to be convex and the parameter θ_0 is defined as the minimizer of an unknown true risk function $R : \Theta \to \mathbb{R}\geq 0$, which is estimated by an empirical risk $\hat{R} : \Theta \to \mathbb{R}\geq 0$ based on a random sample of size n. In general, We considered the estimator of θ_0 of the form

$$\hat{\theta} \in \text{argmin}_{\theta \in \Theta} \widehat{R}(\theta) + \lambda_n |\theta|_1,$$

where $\lambda_n > 0$ is a penalty level chosen concerning n. The ℓ_1-norm penalization is advantageous for two main reasons: one is that according to Bühlmann and Van De Geer (2011), it avoided overfitting by shrinking coefficients of less important variables to zero. The other is that it facilitated the use of stochastic subgradient-based algorithms for optimizing the objective function. In the case of a nonconvex objective function, parallel computing such as Agarwal and Duchi (2011) and noise injection techniques introduced in Ge et al. (2015) can be employed to escape local minima and continue the descent process.

1.2 Motivations and Outline

The essence of these methods are different due to their construction. If we take a step back to understand how the zeros are generated, we find that they typically arise from three different scenarios: (i) Unsuitability, where for certain covariates $\mathbf{X}^*$, the outcome Y is zero with probability 1; (ii) Detection error, which occurs when the outcome Y is zero with high probability, leading to zero observations due to low detection rates in random sampling. Take proportional cover data in plant surveys (Tang et al., 2023) as an example. Zeros due to unsuitability could arise when some plant species are deem unsuitable under certain types of biotic and abiotic factors. If a plant species is deemed suitable for a particular area but is not found there, it could be explained by the low probability of detection during sampling.

Most of the existing methods could explain the ‘zero’ source in the specific problem. (zero-inflated beta regression model would cover the zero from random sampling while the Tobit model covers the detection error other than the random sampling.) We intend to construct a brand new model that could take care of the zero from two source so that it could be flexible and robust in different applications.

Let $(\mathbf{X}^*, \mathbf{Y})$ be a random observation where $\mathbf{X}^* = (1, \mathbf{X})$, $\mathbf{X} = (x_1, \ldots, x_{p-1})^\top$ is $(p-1)$-dimensional covariates and the response Y is a zero-inflated proportion variable, i.e., $Y \in [0, 1)$ and $0 < \mathrm{P(Y = 0|}\mathbf{X}^*) \leq 1$. The constant 1 is incorporated into the vector $\mathbf{X}^*$ to represent the intercept term in the model. To account for the zero-inflated nature of Y, we first of all model $(\mathbf{X}^*, Y)$ via:

$$\mathbf{E}[Y|\mathbf{X}^*] = f(\mathbf{X}^*)\mathrm{I}_{\{D(\mathbf{X}^*)\geq 0\}}. \tag{1.1}$$

Here $D(\mathbf{X}^*)$ is a discriminant function that accounts for the zero inflation arising from systematic or structural unsuitability, i.e., cases where $P(Y = 0|\mathbf{X}^*) = 1$. When the probability $P(Y = 0|\mathbf{X}^*) < 1$, zeros originating from random sampling are characterized by $f(\mathbf{X}^*) : \mathcal{R}^p \mapsto [0, 1]$, i.e., the conditional mean function of Y. This separation of the two zero-generating processes enables the model to provide a more accurate representation of the data-generating mechanism, leading to improved model fit and more reliable inferences, particularly in scenarios where both sources of zeros are present in the data. However, the price we have to pay for adopting the formulation (1.1) is to handle the non-smooth component introduced by the indicator function, which brings challenges to both theoretical analysis and computation. Based on the conditional expectation formulation (1.1), we first of all introduce the following population risk function:

$$\mathcal{L} = \mathbf{E}\left[\frac{1}{2\sigma^2(\mathbf{X}^*)}(Y - f(\mathbf{X}^*)\mathrm{I}_{\{\mathrm{D}(\mathbf{X}^*)\geq 0\}})^2\right], \tag{1.2}$$

where $\sigma^2(\mathbf{X}^*) := f(\mathbf{X}^*)(1 - f(\mathbf{X}^*))$ is the variance of a Bernoulli random variable with success probability $f(\mathbf{X}^*)$. Clearly (1.2) is simply the expectation of a weighted squared loss. The weight $\sigma^2(\mathbf{X}^*)$ is introduced to take into account since the variances of $Y|\mathbf{X}^*$ for different covariates $\mathbf{X}^*$ may be different. The form $\sigma^2(\mathbf{X}^*) \propto f(\mathbf{X}^*)(1 - f(\mathbf{X}^*))$ is adopted to mimic the fact that the proportional response Y is usually calculated as the rate of "events" obtained from simple random sampling. To address heterogeneity, we propose using a weighted loss function, which is a modification of the standard least squares loss that considers the covariance structure of the data. By incorporating the covariance matrix into the model, we can mitigate the influence of outliers or noisy data points, leading to more robust and reliable parameter estimates.

Our proposed model offers several advantages over conventional methods like the Tobit or zero-inflated beta regression models. In the case of the Tobit model, attempting to address heterogeneity by replacing the fixed σ with $\sigma(\mathbf{X}^*)$ would lead to unsatisfactory results due to the non-concavity of the objective function when making general assumptions about $f(X)$. Although the zero-inflated beta regression model with the maximum likelihood method accounts for heterogeneity, it cannot overcome the global convexity problem or provide an exact prediction for the "zero" part of the data. Our model shares some similarities with the two-stage method. In two-stage method, it separates the dataset into the 'zero' part and the'non-zero' part. Then, do the regression only with the 'non-zero' part and set the 'non-zero' part to one, and do the classification part. When new data comes in, use the classification model to tell whether it is zero or positive. If it is positive, predict it with a regression model. The bright side is the two-stage model could obtain an exact 'zero' prediction, but it has a critical limitation since not all 'zero' data are due to unsuitability. With the 'zero' comes from random sampling (detection error), it results in inconsistent estimators in classification. In contrast, our model achieves a balance between the classification and regression components, allowing for adjustments that mitigate the impact of misclassification, providing a more robust and reliable estimation framework by explicitly modeling the different mechanisms (unsuitability, detection error) that lead to excess zeros and incorporating a classification component. As a result, our proposed model offers a more comprehensive and robust approach to analyzing zero-inflated proportion data, resulting in improved estimation accuracy and interpretability compared to existing methods.

Based on the model assumption 1.1 and the loss function we proposed, we could present the general estimation procedure of $f(\mathbf{X}^*)$, $D(\mathbf{X}^*)$. First, based on observations $\{(\mathbf{X}_i^*, Y_i)\}_{i=1}^n$, denote $B_1 := \{(\mathbf{X}_i^*, y_i) | Y_i > 0\}_{i=1}^n$, $n_1 = \text{card(B)}$, we estimate the regression part with

$$\hat{f}(\mathbf{X}^*) = \arg\min_f \sum_{(\mathbf{X}_i^*, Y_i) \in B_1, i=1}^{n_1} \left(\frac{1}{2\sigma^2(\mathbf{X}_i^*)} [Y_i - f(\mathbf{X}_i^*)] \right)^2.$$

Then, in the second step, we noted that there is no need to derive the precise form of $D(\mathbf{X}^*)$ since the sign of it is enough for prediction.

Proposition 1.2. *The Derivation of sign of $D(\mathbf{X}^*)$*
Given $f(\mathbf{X}^*)$, the optimal decision $D(\mathbf{X}^*)$ that minimizes $R(f, D)$ must satisfy:

$$\text{sign}(D(\mathbf{X}^*)) = \text{sign}(\mathbf{E}[-g(Y, \mathbf{X}^*; f(\mathbf{X}^*)) | \mathbf{X}]),$$

where $g(Y, \mathbf{X}^*; f(\mathbf{X}^*)) = \frac{f(\mathbf{X}^*)^2 - 2Yf(\mathbf{X}^*)}{\sigma^2(\mathbf{X}^*)}$.

With proposition 1.2 we take advantage of the Nadaraya–Watson estimator, estimating $\text{sign}(D(\mathbf{X}^*))$ as

$$\text{sign}(\hat{D}(x)) = \text{sign}\Bigg[- \frac{(nh)^{-1} \sum_i^n K_h(d(x, \mathbf{X}_i^*)) g(Y, \mathbf{X}_i^*; f(\mathbf{X}^*))}{(nh)^{-1} \sum_i^n K_h(d(x, \mathbf{X}_i^*))} \Bigg],$$

where $d(x, \mathbf{X}_i^*)$ refers to the distance function and $K_h(u) = K\left(\frac{u}{h}\right)$ is a standard Gaussian kernel function.

The estimation process we provided here is very general, in the following section, we will specify the form of our $f(\cdot)$, $\sigma(\cdot)$ and the distance $d(x, \mathbf{X})$ in the Nadaraya–Watson estimator to further derive the properties of our estimators.

Chapter 2

Semiparametric Model for Zero-inflated Proportion Data

2.1 Model Construction and Estimation Mechanism

In the first section, we briefly introduced our model constructed to capture two types of ‘zero’, we then further completed our work by adding some structure assumptions on the unknown $f(\mathbf{X})$ and $D(\mathbf{X})$. To address the problem, we make the model more specific by formulating the optimization problem as

$$\begin{aligned} L &= \mathbf{E}\left(\frac{1}{2\sigma^2(\mathbf{X}^{*\top}\alpha)}[Y - f(\mathbf{X}^{*\top}\alpha)\mathrm{I}_{\{\mathrm{D}(\mathbf{X}^*)\geq 0\}}]^2\right) \\ &= \mathbf{E}\left(\frac{1}{2}\left[\frac{Y}{\sigma(\mathbf{X}^{*\top}\alpha)} - T(\mathbf{X}^{*\top}\alpha)\mathrm{I}_{\{\mathrm{D}(\mathbf{X}^*)\geq 0\}}\right]\right)^2, \end{aligned}$$

where $T(\mathbf{X}^{*\top}\alpha) = \frac{f(\mathbf{X}^{*\top}\alpha)}{\sigma(\mathbf{X}^{*\top}\alpha)}$.

Instead of giving the form of $f(\mathbf{X}^{*\top}\alpha)$ directly, we focus on the relationship between $f(\mathbf{X}^{*\top}\alpha)$ and $\sigma(\mathbf{X}^{*\top}\alpha)$ first. As we mentioned before, in a product factory, the defective rate for different products is determined by counts, indicating that the observations follow the Binomial distribution $\mathrm{Bin}(\mathrm{m}, \mathrm{p})$ where m represents the sampling number, and p represents the probability of defects. Considering heterogeneity issue as we mentioned before, the following weighted loss function was brought up naturally.

$$\mathcal{L} = \mathbf{E}\bigg[\frac{1}{2f(\mathbf{X}^{*\top}\alpha)(1-f(\mathbf{X}^{*\top}\alpha))}\Big(Y - f(\mathbf{X}^{*\top}\alpha)\mathrm{I}_{\{\mathrm{D}(\mathbf{X}^{*})\geq 0\}}\Big)^2\bigg].$$

Inspired by the reparameterization proposed by Olsen (1978), we considered the form $T(\mathbf{X}^{*\top}\alpha) \propto \exp\{\mathbf{X}^{*\top}\alpha\}$, which results in the global convexity of the loss function with respect to the regression part. With the relationship between $f(\cdot)$ and $\sigma(\cdot)$, and assumption on $T(\cdot)$, we could derive

$$f(\mathbf{X}^{*\top}\alpha) = \frac{(\exp\{\mathbf{X}^{\top}\alpha\})^2}{(1+(\exp\{\mathbf{X}^{\top}\alpha\})^2)}. \tag{2.1}$$

With 2.1 we obtain, our loss function could be written as

$$\begin{aligned}
\mathcal{L} &= \mathbf{E}\bigg[\bigg(\frac{(1+(\exp\{\mathbf{X}^{*\top}\alpha\})^2)Y}{\exp\{\mathbf{X}^{*\top}\alpha\}}\bigg) - (\exp\{\mathbf{X}^{*\top}\alpha\})\mathrm{I}_{\{\mathrm{D}(\mathbf{X}^{*})\geq 0\}}\bigg]^2 \\
&= \mathbf{E}\bigg[\bigg(\frac{(1+(\exp\{\mathbf{X}^{*\top}\alpha\})^2)Y}{\exp\{\mathbf{X}^{*\top}\alpha\}}\bigg)^2 \\
&\quad + \exp\{\mathbf{X}^{*\top}\alpha\}\bigg(\exp\{\mathbf{X}^{*\top}\alpha\} - 2\bigg(\frac{(1+(\exp\{\mathbf{X}^{*\top}\alpha\})^2)Y}{\exp\{\mathbf{X}^{*\top}\alpha\}}\bigg)\bigg)\mathrm{I}_{\{\mathrm{D}(\mathbf{X}^{*})\geq 0\}}\bigg].
\end{aligned}$$

As we illustrated with general construction, denote $B_1 := \{(\mathbf{X}_i^*, y_i)|Y_i > 0\}_{i=1}^n$, $n_1 = \mathrm{card(B)}$. We first obtain the estimator with

$$\hat{\alpha} = \arg\min_{\alpha} \sum_{(\mathbf{X}_i^{*\top}, Y_i)\in B_1, i=1}^{n_1} \frac{(1+(\exp\{\mathbf{X}_i^{*\top}\alpha\})^2)^2}{2(\exp\{\mathbf{X}_i^{*\top}\alpha\})^2}\bigg(Y_i - \frac{(\exp\{\mathbf{X}_i^{*\top}\alpha\})^2}{1+(\exp\{\mathbf{X}_i\alpha\})^2}\bigg)^2,$$

Then, due to proposition 1.2, we have

$$\mathrm{sign}(D(\mathbf{X}^*)) = \mathrm{sign}(\mathbf{E}[-g(Y, \mathbf{X}^*; \alpha)|\mathbf{X}^*]). \tag{2.2}$$

where $g(y, \mathbf{X}^*; \alpha) = \exp\{\mathbf{X}^{*\top}\alpha\}\left(\exp\{\mathbf{X}^{*\top}\alpha\} - 2\left(\frac{(1+(\exp\{\mathbf{X}^{*\top}\alpha\})^2)Y}{\exp\{\mathbf{X}^{*\top}\alpha\}}\right)\right)$. So we construct the Nadaraya–Watson estimator as

$$\hat{D}(x) = \frac{(nh)^{-1}\sum_i^n K_h((x-\mathbf{X}_i^*)^\top\hat{\alpha})g(Y, \mathbf{X}_i^*; \hat{\alpha})}{(nh)^{-1}\sum_i^n K_h((x-\mathbf{X}_i^*)^\top\hat{\alpha})}, \tag{2.3}$$

where $K(\cdot)$ is a standard Gaussian kernel function and $K_h(u) = K\left(\frac{u}{h}\right)$.

Remarks: Our classification estimation primarily hinges on the observation stated in 2.2. There exists a close relationship between our regression step and classification step. To further emphasize and elucidate the intrinsic connection between these two components, we employ the following distance metric: $d(x, \mathbf{X}_i^*) = (x - \mathbf{X}_i^*)^\top\hat{\alpha}$, where $\hat{\alpha}$ represents the estimated coefficient vector obtained from the regression stage. The distance construction facilitates a seamless transition between the regression and classification tasks, enhancing the overall interpretability and coherence of the proposed estimation method.

2.2 Main Theorems

The theoretical analysis of our proposed model comprises three main components. The first part delves into the properties of the loss function employed. It is well-established that a globally convex population loss function offers numerous advantages, including the guarantee of a unique solution, efficient optimization convergence, and robust and stable algorithmic behavior. These desirable characteristics facilitate the optimization process and enhance the reliability of the obtained results.

The second part of our theoretical investigation focuses on establishing the consistency of the estimators derived from the regression stage. In particular, we provide theoretical guarantees for the consistency of our estimators and further complement these results by deriving tail bounds on the estimation error of the regression coefficient vector $\hat{\alpha}$.

In the third part, we turn our attention to the kernel estimation component of our model. After obtaining the estimated regression coefficient vector $\hat{\alpha}$ from the regression stage, we analyze the uniform convergence rate of our kernel-based estimator, which provides theoretical guarantees for the accuracy of our proposed approach.

By providing a comprehensive theoretical analysis encompassing the loss function properties, tail bounds, and uniform convergence rates, we establish a solid theoretical foundation for our model. The theoretical analysis serves as a crucial component in understanding the strengths and capabilities of our proposed model, facilitating future research and development in this domain.

Firstly, we brought up the global convexity property of our loss function.

Proposition 2.1. *Given $D(\mathbf{X}^*)$, denote $n_1 = \text{card}(\{\mathbf{X}^* | D(\mathbf{X}^* \geq 0)\})$, then the empirical loss function*

$$\mathcal{L}_n(\alpha) = \frac{1}{n}\sum_{i=1}^{n} \frac{(1 + (\exp\{\mathbf{X}_i^{*\top}\alpha\})^2)^2}{2(\exp\{\mathbf{X}_i^{*\top}\alpha\})^2}\left(Y_i - \frac{(\exp\{\mathbf{X}_i^{*\top}\alpha\})^2}{1 + (\exp\{\mathbf{X}_i^{*\top}\alpha\})^2}\mathrm{I}_{\{D(\mathbf{X}_i^*)\geq 0\}}\right)^2$$

is strongly convex concerning α in probability as $n_1 \to \infty$.

The proof of 2.1 is straightforward. Since $D(\mathbf{X}^*)$ is given, we separate the loss into two parts by $D(\mathbf{X}^* \geq 0)$ and $D(\mathbf{X}^* < 0)$. Then, calculate the empirical Hessian matrix accordingly, which could be proven to be convex in probability. Details are given in the last section of Chapter 2.

Now, we proceed to prove the uniform consistency of our estimators. Firstly, we conclude a preliminary uniform consistency result based on theorem 4.1.1 of Amemiya (1985), and then, we will further discuss the tail bound of our estimator $\hat{\alpha}$.

Definition 2.1. *Ergodic Property*

An arbitrary sequence of random measures $\{\lambda_n(A, \omega),\ \lambda(A, \omega); n = 1, 2, \cdots\}$ on X

is said to possess the ergodic property if for each real-valued function $g(x)$ on X, for which $E \int |g(x)| \lambda(dx, \omega) < \infty$

$$\lim_{n \to \infty} \int g(x) d\lambda_n = \int g(x) d\lambda$$

almost everywhere.

Rao (1962) introduced the definition while exploring the relationships between weak convergence (convergence in distribution) and uniform convergence (uniform convergence in total variation) of measures. The work provides conditions under which weak convergence implies uniform convergence and demonstrates how the developed theory can be used to establish uniform consistency of empirical processes and derive uniform convergence rates for estimators.

Assumption 2.1. *Assume that $\mathbf{X}^*$ belongs to a compact set $\mathcal{X}$, and there exists a finite M such that $\|\mathbf{X}^*\|_2^2 < M$.*

Assumption 2.2. *There exists $0 < V, e < \infty$ such that $\|\alpha\|_2^2 \leq V$, $\|\mathcal{L}(\mathbf{X}^*, \alpha)\|_2^2 \leq e$.*

Assumption 2.3. *Assume that $\{(X, Y) | Y > 0\}$ possess the ergodic property, i.e. for each real-valued function $f(x, y)$ on $\{(X, Y) | Y > 0\}$, for which (i) $\mathbf{E} \int |f(x, y)| d\lambda < \infty$; (ii) $\int f(x, y) d\lambda_n = \int f(x, y) d\lambda$ almost everywhere.*

Assumption 2.4. $\mathbf{E}[X^* X^{*\top} | Y > 0]$ *is positive definite whose eigenvalues are bounded by $0 < \sigma_0^2 < \infty$.*

Assumption 2.1 and 2.2 impose general conditions on the covariates and parameters, respectively. Specifically, these assumptions ensure that the covariates and parameters are uniformly bounded, which, in turn, guarantees that the loss function $\mathcal{L}$ satisfies the properties of uniform bound and equicontinuity for any available value of α. These properties are essential for establishing uniform convergence results of the empirical loss function.

Assumption 2.3 is a crucial condition that links the convergence in distribution with the uniform convergence of measures. By satisfying this assumption, we can construct the uniform convergence of the empirical loss to the population loss under certain regularity conditions imposed on the loss function.

Assumption 2.4 ensures that the population loss function in the first step is strongly convex, which guarantees the existence and uniqueness of its minimizer. Besides, it facilitates the convergence of optimization algorithms and allows for the establishment of convergence rates.

Lemma 2.1. *Suppose assumption 2.1 $\sim$ 2.4 hold, then given the subset* $B_1 := \{(\mathbf{X}_i^*, Y_i)|Y_i > 0\}$, *denote* $n_1 = \mathrm{card}(B_1)$, $\hat{\alpha} \to_p \alpha_0$ as $n_1 \to \infty$,

where $\hat{\alpha} = \arg\min_\alpha \sum_{(\mathbf{X}_i^*, Y_i)\in B_1, i=1}^{n_1} \frac{(1+(\exp\{\mathbf{X}_i^{*\top}\alpha\})^2)^2}{2(\exp\{\mathbf{X}_i^{*\top}\alpha\})^2}\left(Y_i - \frac{(\exp\{\mathbf{X}_i^{*\top}\alpha\})^2}{1+(\exp\{\mathbf{X}_i^{*\top}\alpha\})^2}\right)^2,$

$\alpha_0 = \arg\min_\alpha \mathbf{E}\left[\frac{(1+(\exp\{\mathbf{X}^{*\top}\alpha\})^2)^2}{2(\exp\{\mathbf{X}^{*\top}\alpha\})^2}\left(Y - \frac{(\exp\{\mathbf{X}^{*\top}\alpha\})^2}{1+(\exp\{\mathbf{X}^{*\top}\alpha\})^2}\right)^2 \middle| D_0(\mathbf{X}^*) > 0\right].$

Proof. With our assumption 2.1 $\sim$ 2.3, according to the theorem 6.2 of Rao (1962), we could obtain

$$\lim_{n\to\infty} \mathbf{P}(\eta_n \to 0) = 1,$$

where $\eta_n = \sup_{l\in\mathcal{F}} \int |l(x,y)d\lambda_n - \int l(x,y)d\lambda|$, $l = \frac{(1+(\exp\{\mathbf{X}^{*\top}\alpha\})^2)^2}{2(\exp\{\mathbf{X}^{*\top}\alpha\})^2}\left(Y - \frac{(\exp\{\mathbf{X}^{*\top}\alpha\})^2}{1+(\exp\{\mathbf{X}^{*\top}\alpha\})^2}\right)^2$,

and λ_n refers to the empirical random measure of $(\mathbf{X}_n^*, Y_n)$ condition on the subset $\{Y_i > 0\}$ while λ indicates the population measure of $(\mathbf{X}^*, Y)$ condition on the subset $\{Y > 0\}$. Then, according to theorem 4.1.1 of Amemiya (1985), we concluded that $\hat{\alpha} \to \alpha_0$ in probability. □

2.2.1 Tail Bound

We intend to derive the tail bound on $P(\|\hat{\alpha} - \alpha_0\|_2 > \tau)$ for any given $\tau > 0$. To establish the tail bound, our first step is to relate $\|\hat{\alpha} - \alpha_0\|_2 > \tau$ with $\mathcal{L}_n(\hat{\alpha}) - \mathcal{L}_n(\alpha_0)$.

Denote $B^c(\alpha_0, \tau)$ as $\{\alpha : \; 2V > \|\alpha - \alpha_0\|_2 > \tau\}$. Since $\mathcal{L}_n(\hat{\alpha}) \leq \mathcal{L}_n(\alpha_0)$, then if $\inf_{\{\alpha \in B^c(\alpha_0,\tau)\}} \mathcal{L}_n(\alpha) > \mathcal{L}_n(\alpha_0)$ holds, we could conclude that $\hat{\alpha} \in B(\alpha_0, \tau)$. Based on this, given τ, we could establish that $P(\inf_{\{\alpha \in B^c(\alpha_0,\tau)\}} \mathcal{L}_n(\alpha) > \mathcal{L}(\alpha_0)) > P(\|\hat{\alpha} - \alpha_0\|_2 > \tau)$. To calculate $\inf_{\alpha \in B^c(\alpha_0,\tau)} \mathcal{L}_n(\alpha) - \mathcal{L}_n(\alpha_0)$, we need a uniform bound on α in $B^c(\alpha_0, \tau)$. However, the classical Talagrand's inequality could only be applied to the set with countable elements. We form a new set $\mathcal{F}$ by partitioning $\mathcal{G}$ into countably many balls of radius ϵ, and subsequently incorporating each ball's center into $\mathcal{F}$. Noted that from the covering number property, the covering number of $\mathcal{G}(\alpha, \epsilon) := \{\alpha : \; \|\alpha\|_2^2 < V\}$ denoted as $N(\epsilon)_{\mathcal{F}}$ satisfied that $N(\epsilon)_{\mathcal{F}} \leq (\frac{2V\sqrt{d}}{\epsilon})^d$, where d refers to the dimension of α.

Our proof is organized into two distinct segments, each addressing a specific aspect of the problem. In the initial segment, we establish a uniform bound for the quantity $\inf_{\alpha \in \mathcal{J}} \mathcal{L}_n(\alpha) - \mathcal{L}_n(\alpha_0)$, where $\mathcal{L}_n$ denotes the empirical loss function, α_0 represents the true parameter value, and $\mathcal{J}$ is a suitably defined parameter space. This bound plays a crucial role in quantifying the deviation of the empirical loss from its population counterpart, laying the foundation for subsequent analysis.

In the second segment, we investigate the relationship between the Euclidean norms $|\hat{\alpha} - \alpha_0|_2$ and $|\bar{\alpha} - \alpha_0|_2$, where $\hat{\alpha}$ is the estimated parameter vector, and $\bar{\alpha}$ denotes the projection of $\hat{\alpha}$ onto the parameter space $\mathcal{J}$, effectively representing the point in $\mathcal{J}$ closest to $\hat{\alpha}$. This analysis is essential for bridging the gap between the empirical risk minimizer $\hat{\alpha}$ and the true parameter value α_0, enabling us to quantify the estimation error. To clarify the notation used in the upcoming equation, we define the following:

Definition 2.2. $\mathcal{J} := \{\alpha | \alpha \in B^c(\alpha_0, \tau) \wedge \alpha \in \mathcal{F}\}$, *where* $B^c(\alpha_0, \tau)$ *denotes the complement of the ball centered at* α_0 *with radius* τ*, and* $\mathcal{F}$ *is a relevant function class. This set* $\mathcal{J}$ *represents the collection of parameters* α *that lie outside the ball*

$B(\alpha_0, \tau)$ and belong to the class $\mathcal{F}$.

Definition 2.3.

$$\mathcal{L}_1(x, \alpha, y) := \left[\frac{Y(1 + (\exp \mathbf{X}^{*\top}\alpha)^2)}{\exp\{\mathbf{X}^{*\top}\alpha\}} - \exp\{\mathbf{X}^{*\top}\alpha\} \right]^2,$$

which is a specific loss function involving the covariates $\mathbf{X}^*$, the parameter α, and the response variable Y. The loss function $\mathcal{L}_1(x, \alpha, y)$ is assumed to be bounded, such that $|\mathcal{L}_1(x, \alpha, y)| < v$, $|\mathcal{L}_1(x, \alpha, Y)^2| < S$ where S, v are some positive constants.

The bound assumption on $\mathcal{L}_1$ ensures that the subsequent analysis and theoretical results are valid and well-defined within the specified parameter space.

I· Evaluating $\sup_{\{\alpha \in \mathcal{J}\}} \mathcal{L}_{1n}(\alpha_0) - \mathcal{L}_{1n}(\alpha)$.

Denote $\mathbb{L}(\cdot) = \mathbf{E}[\mathcal{L}_{1n}(\cdot)]$, and then separate $\mathcal{L}_{1n}(\alpha_0) - \mathcal{L}_{1n}(\alpha)$ into three parts:

$$\begin{aligned}
&\sup_{\alpha \in \mathcal{J}} \mathcal{L}_{1n}(\alpha_0) - \mathcal{L}_{1n}(\alpha) = \sup_{\{\alpha \in \mathcal{J}\}} \mathcal{L}_{1n}(\alpha_0) - \mathbb{L}(\alpha_0) + \mathbb{L}(\alpha_0) - \mathbb{L}(\alpha) + \mathbb{L}(\alpha) - \mathcal{L}_{1n}(\alpha) \\
&\leq (\mathcal{L}_{1n}(\alpha_0) - \mathbb{L}(\alpha_0)) + \sup_{\{\alpha \in \mathcal{J}\}} [(\alpha) - \mathcal{L}_{1n}(\alpha)] + \sup_{\{\alpha \in \mathcal{J}\}} [\mathbb{L}(\alpha_0) - \mathcal{L}_{1n}(\alpha)] \\
&\leq (\mathcal{L}_{1n}(\alpha_0) - \mathbf{E}[\mathcal{L}_{1n}(\alpha_0)]) + \sup_{\alpha \in \mathcal{F}} [\mathbb{L}(\alpha) - \mathcal{L}_{1n}(\alpha)] + \sup_{\{\alpha \in \mathcal{J}\}} (\mathbb{L}(\alpha_0) - \mathbb{L}(\alpha)).
\end{aligned} \tag{2.4}$$

For the first and second part, we denote $Z_n(X_i, \alpha) = \frac{1}{n}\big\{ \sum_{i=1}^n \mathcal{L}_{1n}(\mathbf{X}_i^*, \alpha) - \mathbb{L}(\mathbf{X}_i^*, \alpha)\big\}$, $\|Z_n\|_{\mathcal{F}} := \sup_{\alpha \in \mathcal{F}} |Z_n(\mathbf{X}^*, \alpha)|$, and $\mathcal{F}$ indicates a compact set we define above. Then applying the equation (18) in Massart and Élodie Nédélec (2006), we have

$$\Pr\left\{ ||Z_n||_{\mathcal{F}} - \mathbf{E}\,\|Z_n\|_{\mathcal{F}} \geq \sqrt{2\frac{(v + 4S\mathbf{E}\,\|Z_n\|_{\mathcal{F}})y}{n}} + \frac{2Sy}{3n} \right\} \leq e^{-y}, \tag{2.5}$$

for some positive constants S and v.

In the context of this inequality, the key objective is to derive a bound for the tail probability of the supremum norm $\|Z_n\|_{\mathcal{F}}$ of the empirical process Z_n, where $\mathcal{F}$ is a countable function class. To achieve this, it is crucial to obtain an upper bound for the expected value $\mathbf{E}\|Z_n\|_{\mathcal{F}}$, which represents the mean or average behavior of the supremum norm.

Proposition 2.2. *Derivation of boundary of the empirical process* $\mathbf{E}\left\|Z_n\right\|_{\mathcal{F}}$.

$$\mathbf{E} \max_{\alpha\in\mathcal{F}} \{Z_n(X_i,\alpha)\} \leq \sqrt{\frac{8S_1^2 \log|\mathcal{F}|}{n}},$$

where $\|\mathcal{L}_1(\mathbf{X}^*,\alpha)\|_2^2 \leq S_1$, *and* $|\mathcal{F}|$ *refers to the cardinality of* $\mathcal{F}$.

The proposition here is an application of Hoeffding's lemma, which was originally introduced by Hoeffding (1994), providing a powerful concentration inequality for the supremum of an empirical process indexed by a class of functions. Specifically, it establishes an upper bound on the probability that the supremum of the empirical process deviates from its expectation by a certain amount, given that the underlying functions are bounded.

As $\alpha \in \mathcal{F}$, which is a countable set, $N(\epsilon)_{\mathcal{F}} \leq (\frac{2V\sqrt{d}}{\epsilon})^d$ could be obtained. Then we take $\epsilon = n^{-c_0}$, where $0 < c_0$, according to proposition 2.2 , it is not hard to derive $\mathbf{E}\left\|Z_n\right\|_{\mathcal{F}} = c_1\sqrt{\frac{\log(n)}{n}}$. Gather the derivation together, we have

$$\Pr\left\{ ||Z_n||_{\mathcal{F}} - c_1\sqrt{\frac{\log(n)}{n}} \geq \sqrt{2\frac{(v+4Sc_1\sqrt{\frac{\log(n)}{n}})y}{n} + \frac{2Sy}{3n}} \right\} \leq e^{-y}.$$

In other words, with probability e^{-y}, we have

$$\sup_{\alpha\in\mathcal{F}}[\mathbb{L}(\alpha) - \mathcal{L}_{1n}(\alpha)] \leq c_2 \max\left\{\frac{2Sy}{3n}, \sqrt{2\frac{(v+4Sc_1\sqrt{\frac{\log(n)}{n}})y}{n}}, c_1\sqrt{\frac{\log(n)}{n}}\right\},$$

where c_2 is some bounded constant.

Then, for the last part in 2.4, we have

$$\begin{aligned}\mathbb{L}(\alpha_0) - \mathbb{L}(\alpha) &= -(\alpha - \alpha_0)^{\mathrm{T}}\mathbb{L}'(\alpha_0) - (\alpha - \alpha_0)^{\mathrm{T}}\mathbb{L}''(\alpha_0)(\alpha - \alpha_0) + o_p(\|\alpha - \alpha_0\|_2) \\ &= -(\alpha - \alpha_0)^{\mathrm{T}}\mathbb{L}''(\alpha_0)(\alpha - \alpha_0) + o_p(\|\alpha - \alpha_0\|_2) \\ &\leq -c_3\|\alpha - \alpha_0\|_2^2 \\ &\leq -c_3\tau^2,\end{aligned}$$

where $c_3 > 0$ indicates the minimal eigenvalue of $\mathbb{L}''(\alpha_0)$. It's important to note that because $\mathbb{L}''(\alpha_0)$ exhibits strong convexity within the subset $\{(\mathbf{X}_i^*, Y_i)|Y_i > 0\}$, c_3 is guaranteed to be a positive value distinct from zero.

Note that the results we obtained are based on $\alpha \in \mathcal{F}$, which is the countable set. Now, we need to extend it to $\alpha \in \mathcal{J}$, which require us to balance the value of τ and y to ensure

$$\sup_{\{\alpha \in \mathcal{J}\}} \mathcal{L}_n(\alpha_0) - \mathcal{L}_n(\alpha) \leq 2c_2\frac{y}{n} - c_3\tau^2 \leq 0.$$

This inequality is used in the choose of the order of y and τ in our theorem.

II· Evaluating $P(\alpha \in B(\alpha_0, \tau))$.

Since

$$\|\hat{\alpha} - \alpha_0\|_2 < \|\hat{\alpha} - \bar{\alpha}\|_2 + \|\bar{\alpha} - \alpha_0\|_2,$$

where $\bar{\alpha}$ refers to the nearest center to $\hat{\alpha}$ with respect to radius ϵ, α_0^* refers to the nearest center to α_0 with respect to radius ϵ. we have $P(\|\hat{\alpha} - \alpha_0\|_2 > \tau) < P(\|\bar{\alpha} - \alpha_0\|_2 > \tau - \epsilon)$, where $\epsilon = o_p(\tau)$ refers to the radius of the balls.

Proposition 2.3. *Suppose $\hat{\alpha} \in B(\bar{\alpha}, \epsilon)$, $\hat{\alpha} \in B^c(\alpha_0, \tau)$ $\epsilon = o_p(\tau)$ hold, then $\mathcal{L}_{1n}(\bar{\alpha}) \leq \mathcal{L}_{1n}(\alpha_0)$ with probability one.*

The derivation is straightforward, denote $b_1 :=$ the ball contained $\hat{\alpha}$. Since $\|\hat{\alpha} - \alpha_0\|_2 < \|\hat{\alpha} - \bar{\alpha}\|_2 + \|\bar{\alpha} - \alpha_0\|_2 < \|\bar{\alpha} - \alpha_0^*\|_2 + \max_{\alpha \in b_1} \|\alpha - \hat{\alpha}\|_2$, $\|\hat{\alpha} - \alpha_0\|_2 > \tau \quad \text{implies} \quad \|\bar{\alpha} - \alpha_0\|_2 > \tau - \epsilon$.

According to proposition 2.3 the loss function $\mathcal{L}_{1n}$ is strongly convex in probability, $\mathcal{L}_{1n}(\bar{\alpha}) \leq \mathcal{L}_{1n}(\alpha_0)$ holds in probability.

Then, according to our discussion, as we take $\epsilon = n^{-c_0}$, $\tau = n^{-d_1}$ where $0 < d_1 < c_0 < 1/2$, then according to (2.2.1), set $y = n^{1-2d_1}$, we have

$$P(\|\bar{\alpha} - \alpha_0\|_2 < \tau - \epsilon) \geq P\left(\inf_{\mathcal{J}^*} \mathcal{L}_{1n}(\alpha) > \mathcal{L}_{1n}(\alpha_0)\right),$$

where $\mathcal{J}^* = \{\alpha \in B^c(\alpha_0, (\tau - 2\epsilon)) \wedge \alpha \in \mathcal{F}\}$ which leads to

$$\begin{aligned} P(\|\hat{\alpha} - \alpha_0\|_2 > \tau) &\leq P(\|\bar{\alpha} - \alpha_0\|_2 > \tau - \epsilon) \\ &< 1 - P\left(\inf_{\alpha \in \mathcal{J}^*} \mathcal{L}_{1n}(\alpha) > \mathcal{L}_{1n}(\alpha_0)\right) \qquad (2.6) \\ &= e^{-y} \end{aligned}$$

The right hand side of (2.6) tends to zero. Gather all the discussion, we present our theorem of our tail bound as follows

Theorem 2.1. *Suppose Assumptions 2.2 $\sim$ 2.4 hold, for any* $0 < d_1 < 1/2$, *we have*

$$P(\|\hat{\alpha} - \alpha_0\|_2 > n^{-d_1}) \leq e^{-n^{(1-2d_1)}}.$$

2.2.2 Consistency in Classification Part

For the theoretical analysis of classification part, it mainly relies on the Lemma B.1 of Newey (1994). However due to the construction of our Nadaraya-Watson estimator, the uniform convergence rate not only depends on the kernel approximation but also depends on the convergence rate of $\hat{\alpha}$ to α_0. So, we separate the proof into two parts.

Denote

$$\hat{g}^*(\mathbf{x}^*, Y_i, \hat{\alpha}) = \frac{(nh)^{-1}\sum_i^n K_h(d(\mathbf{x}^*, \mathbf{X}_i^*))g(\mathbf{X}_i^*; Y_i, \hat{\alpha})}{(nh)^{-1}\sum_i^n K_h(d(\mathbf{x}^*, \mathbf{X}_i^*))},$$

$g^*(\mathbf{X}^*, \alpha) = \mathbf{E}[g(\mathbf{X}^*, Y, \alpha)|\mathbf{X}^*]$, where

$$g(\mathbf{X}^*, Y, \alpha) = \exp\{\mathbf{X}^{*\top}\alpha\}\bigg(\exp\{\mathbf{X}^{*\top}\alpha\} - 2\bigg(\frac{(1 + (\exp\{\mathbf{X}^{*\top}\alpha\})^2)Y}{\exp\{\mathbf{X}^{*\top}\alpha\}}\bigg)\bigg).$$

In the first part, we establish a connection between the bound of $\hat{g}^*(\mathbf{X}^*, Y, \hat{\alpha}) - \hat{g}^*(\mathbf{X}^*, Y, \alpha_0)$ and $\hat{\alpha} - \alpha_0$. This step is crucial as it links the error in the estimated parameter $\hat{\alpha}$ to the error in the estimated function $\hat{g}^*$.

In the second part, we further analyze the error bound of $g^*(\mathbf{X}^*, \alpha_0) - \hat{g}^*(\mathbf{X}^*, Y_i, \alpha_0)$ by leveraging the properties of the Nadaraya-Watson estimator and kernel approximation.

Lemma 2.2. *Suppose Assumption 2.1 $\sim$ 2.4 holds, we have*

$$\sup_{\{\mathbf{x}^* | \|\mathbf{x}^*\|_2^2 \leq M,\}} |\hat{g}(Y, \mathbf{x}^*; \hat{\alpha}_0) - g^*(\mathbf{x}^*, \alpha_0)| = O_p\bigg(\max\bigg(\Big(\frac{\log n}{n}\Big)^{1/3}, n^{-d_1}\bigg)\bigg),$$

where $0 < d_1 < 1/2$.

Note that, as we mentioned before, we did not need the estimation of $D(X^*)$, we only need to ensure the estimation of the sign is consistent in probability. Thus, we present the consistency estimation of our classification result as follows.

Theorem 2.2. *If assumption1 $\sim$ 3 holds, we have*

$$\sup_{\{\mathbf{x}^* | \|\mathbf{x}^*\|_2 \leq M\}} |\mathrm{sgn}(\hat{g}(Y, \mathbf{x}^*; \hat{\alpha}_0)) - \mathrm{sgn}(g^*(\mathbf{x}^*, \alpha_0))| = 0,$$

with probability tending to one as $n \to \infty$.

Proof. According to Lemma 2.2, We could conclude the theorem as long as $g(x, y, \alpha)$ is bounded away from zero with any $\mathbf{X}^*$ and α. □

Through theorem 2.1, we establish the consistency property of the parameter α, while theorem 2.2 addresses the consistency of the sign estimation for $D(\mathbf{X}^*)$. Together, these two theorems provide a comprehensive theoretical framework for our model.

2.3 Proof

2.3.1 Proof of proposition 2.1

Proof. For the original problem, in the case of $D(\mathbf{X}^*) \geq 0$, the weighted loss function could be written as

$$L_1 = \frac{1}{n_1}\sum_{i=1}^{n_1}\left[\frac{y_i(1+(\exp\{\mathbf{X}_i^{*\top}\alpha\})^2)}{\exp\{\mathbf{X}_i^{*\top}\alpha\}} - \exp\{\mathbf{X}_i^{*\top}\alpha\}\right]^2,$$

where we denote n_1 here to be the number of $\{\mathbf{X}^*|D(\mathbf{X}^* \geq 0)\}$. So, we get the derivative of L_1 with respect of α_j is

$$\frac{\partial L_1}{\partial \alpha_j} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{ij}\frac{(\exp\{\mathbf{X}_i^{*\top}\alpha\})^4(y_i-1)^2 - y_i^2}{(\exp\{\mathbf{X}_i^{*\top}\alpha\})^2}.$$

Further, we get the Hessian matrix with respect to α is

$$\frac{\partial^2 L_1}{\partial \alpha_j \partial \alpha_k} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{ij}X_{ik}2(\exp\{\mathbf{X}_i^{*\top}\alpha\})^2(y_i-1)^2 + 2y_i^2(\exp\{\mathbf{X}_i^{*\top}\alpha\})^{-2}.$$

We transform the above equation to matrix form, denote

$$T_i = \mathbf{X}_i^*\sqrt{2(\exp\{\mathbf{X}_i^{*\top}\alpha\})^2(y_i-1)^2 + 2y_i^2(\exp\{\mathbf{X}_i^{*\top}\alpha\})^{-2}},$$

for every i $\in (1, ..., n_1)$, then we have

$$\frac{\partial^2 L_1}{\partial \alpha^2} = \frac{1}{n_1}T^\top T,$$

where T is $n_1 \times p$ matrix. As long as the full column rank of T is guaranteed with probability one, we could obtain the positive-definite property of the Hessian matrix.

For another part $D(\mathbf{X}^*) < 0$, denote n_2 here to be the number of $\{\mathbf{X}^* | D(\mathbf{X}^* < 0)\}$, The loss function is

$$L_2 = \frac{1}{2n_2} \sum_{i=1}^{n_2} \left[\frac{y_i(1 + (\exp\{\mathbf{X}_i^{*\top}\alpha\})^2)}{\exp\{\mathbf{X}_i^{*\top}\alpha\}} \right]^2,$$

so, we obtain the derivative of $\mathcal{L}$ with respect of α_j is

$$\frac{\partial L_2}{\partial \alpha_j} = \frac{1}{n_2} \sum_{i=1}^{n_2} X_{ij} y_i^2 \frac{(\exp\{\mathbf{X}_i \alpha\})^4 - 1}{(\exp\{\mathbf{X}_i^{*\top}\alpha\})^2}.$$

Then, the Hessian matrix with respect to α is

$$\frac{\partial^2 L_2}{\partial \alpha_j \partial \alpha_k} = \frac{1}{n_2} \sum_{i=1}^{n_2} X_{ij} X_{ik} y_i^2 [2(\exp\{\mathbf{X}_i^{*\top}\alpha\})^2 + 2(\exp\{\mathbf{X}_i^{*\top}\alpha\})^{-2}],$$

which is non-negative.

By combining the Hessian matrix for $D(\mathbf{X}^*) \geq 0$ and $D(\mathbf{X}^* < 0)$ together, we can ensure that the resulting loss function is positive definite in probability.

□

Discussions: the weighted loss function serves multiple purposes in our problem. It addresses the heterogeneity in the data and helps overcome the convexity problem commonly encountered with the regular least square loss.

If we insist with

$$\mathcal{R}(\theta) = \mathbf{E}\left[Y - f(\mathbf{X}^{*\top}\alpha) \mathrm{I}_{\{\mathrm{D}(\mathbf{X}^*) \geq 0\}} \right]^2,$$

then, we could still discuss the convexity of our loss function for the case $\{D(\mathbf{X}^*) \geq 0\}$ and $\{D(\mathbf{X}^*) < 0\}$.

When $D(\mathbf{X}^*) \geq 0$, we have

$$\mathcal{R}(\theta) = \mathbf{E}\left[Y - f(\mathbf{X}^{*\top}\alpha)\right]^2,$$

then, we have its second derivative equal to

$$\frac{\partial^2 R}{\partial \alpha^2} = 2\mathbf{X}^{*\top}\mathbf{X}^*\mathbf{E}[f'(\mathbf{X}^{*\top}\alpha)^2 - f''(\mathbf{X}^{*\top}\alpha)(Y - f(\mathbf{X}^{*\top}\alpha))].$$

Since $\mathbf{E}[Y|\mathbf{X}^*] = f(\mathbf{X}^{*\top}\alpha_0)$, so we have

$$\frac{\partial^2 R}{\partial \alpha^2} = 2\mathbf{X}^{*\top}\mathbf{X}^*\mathbf{E}[f'(\mathbf{X}^{*\top}\alpha)^2 - f''(\mathbf{X}^{*\top}\alpha)f(\mathbf{X}^{*\top}\alpha_0) - f(\mathbf{X}^{*\top}\alpha_0))].$$

With the variation of α, $f''(\mathbf{X}^{*\top}\alpha)(f(\mathbf{X}^{*\top}\alpha_0) - f(\mathbf{X}^{*\top}\alpha)) < 0$ could not be guaranteed under the condition that $\mathbf{X}^{*\top}\alpha$ changes between $(-\infty, \infty)$ while $f(\mathbf{X}^{*\top}\alpha)$ ranges among (0,1). As a result, the convexity of the regular least squares loss cannot be ensured, which may lead to stability issues in the algorithm.

2.3.2 Proof of proposition 2.2

Proof. Set $T = \max_{a\in\mathcal{F}} T_a = \max_{a\in\mathcal{F}}\{\mathcal{L}_1(X_i,\alpha) - \mathbf{E}(\mathcal{L}_1(X_i,\alpha))\}$. $0 \leq \mathcal{L}_1(X_i,\alpha) \leq S_1$ and $0 < \mathbf{E}(\mathcal{L}_1(X_i,\alpha)) \leq S_1$, so that $0 \leq \mathcal{L}_1(X_i,\alpha) - \mathbf{E}(\mathcal{L}_1(X_i,\alpha)) \leq S_1$, by Hoeffding's lemma we get

$$\begin{aligned}\mathbf{E}e^{\lambda T_a} &= \mathbf{E}e^{\lambda\sum_{i=1}^n\{\mathcal{L}_1(X_i,\alpha)-\mathbf{E}(\mathcal{L}_1(X_i,\alpha))\}/n} \\ &= \prod_{i=1}^n \mathbf{E}e^{\lambda\{\mathcal{L}_1(X_i,\alpha)-\mathbf{E}(\mathcal{L}_1(X_i,\alpha))\}/n} \leq \left(e^{\lambda^2(S_1)^2/\left(8n^2\right)}\right)^n = e^{L_0^2\lambda^2/(8n)},\end{aligned}$$

Putting everything together, we get

$$\begin{aligned}\mathbf{E}\max_{a\in\mathcal{F}}\{Z_n(X_i,\alpha)\} &\leq \frac{1}{\lambda}\log\sum_{a\in\mathcal{F}} e^{S_1^2\lambda^2/(8n)} \\ &= \frac{1}{\lambda}\log\left(|\mathcal{F}|e^{S_1^2\lambda^2/(8n)}\right) \qquad (2.7) \\ &= \frac{1}{\lambda}\log|\mathcal{F}| + \frac{S_1^2\lambda}{8n}.\end{aligned}$$

The optimal value of (2.7) is $\sqrt{\frac{8S_1^2 \log|\mathcal{F}|}{n}}$. □

2.3.3 Proof of proposition 2.3

Proof.

$$\begin{aligned}\mathcal{L}_n(\bar{\alpha}) - \mathcal{L}_n(\alpha_0) &= \mathcal{L}_n(\bar{\alpha}) - \mathcal{L}_n(\hat{\alpha}) + \mathcal{L}_n(\hat{\alpha}) - \mathcal{L}_n(\alpha_0) \\ &= (\bar{\alpha} - \hat{\alpha})^{\mathrm{T}} \mathcal{L}_n''(\hat{\alpha})(\bar{\alpha} - \hat{\alpha}) - (\alpha_0 - \hat{\alpha})^{\mathrm{T}} \mathcal{L}_n''(\alpha_0)(\alpha_0 - \hat{\alpha}) \\ &+ o_p(\|\bar{\alpha} - \hat{\alpha}\|_2) + o_p(\|\hat{\alpha} - \alpha_0\|_2).\end{aligned}$$

From proposition 2.1 and assumption 2.1, it is not hard to conclude that the eigenvalue of $\mathcal{L}_n''\, \mathcal{L}_n(\alpha)$ is bounded in $[e_1, e_2]$ in probability , where e_1, e_2 are some constants satisfied $e_1 > 0$ is bounded away from zero and $e_2 < \infty$. Then we have

$$\mathcal{L}_n(\bar{\alpha}) - \mathcal{L}_n(\alpha_0) < e_2\|\bar{\alpha} - \hat{\alpha}\|_2 - e_1\|\hat{\alpha} - \alpha_0\|_2 < e_2\epsilon^2 - e_1\tau^2 < 0$$

□

2.3.4 Proof of lemma 2.2

Proof. For the consistency estimation of the classification part, denote $g^*(\mathbf{X}, \alpha) = \mathbf{E}[g(\mathbf{X}, Y, \alpha)|\mathbf{X}]$, where $g(\mathbf{X}, y, \alpha) = \exp\{\mathbf{X}^\top\alpha\}\Bigg(\exp\{\mathbf{X}^\top\alpha\} - 2\bigg(\frac{(1+(\exp\{\mathbf{X}^\top\alpha\})^2)Y}{\exp\{\alpha^{\mathrm{T}}x\}}\bigg)\Bigg)$. We intend to take advantage of the Nadaraya-Watson estimator, construct

$$\hat{g}^*(\mathbf{x}^*, Y_i, \hat{\alpha}) = \frac{(nh)^{-1}\sum_i^n K_h(d(\mathbf{x}^*, \mathbf{X}_i^*))g(\mathbf{X}_i^*; Y_i, \hat{\alpha})}{(nh)^{-1}\sum_i^n K_h(d(\mathbf{x}^*, \mathbf{X}_i^*))},$$

where $d(\mathbf{x}^*, \mathbf{X}_i^*) := (\mathbf{x}^* - \mathbf{X}_i^*)^{\mathrm{T}}\alpha$.

In the algorithm, $\hat{\alpha}$ is what we obtained and used to estimate the sign of $D_1(\mathbf{X}*)$. Since $\sup_{\|\mathbf{x}^*\|_2^2 \geq M} |g(\mathbf{x}^*, Y, \hat{\alpha}) - g(\mathbf{x}^*, Y, \alpha_0)| \leq C(\|\hat{\alpha} - \alpha_0\|_2)$, where C refers to some constant. So, we conclude $\|\hat{g}^*(\mathbf{x}^*, Y, \hat{\alpha})$ - $\hat{g}^*(\mathbf{x}^*, Y, \alpha_0)\| = O_p(\|\hat{\alpha} - \alpha_0\|_2) = O_p(n^{-d_1})$.

Then, we consider $m(\mathbf{x}^*, \alpha_0) = g^*(\mathbf{x}^*, \alpha_0) f(\mathbf{x}^{*\top}\alpha_0)$, where $f(\mathbf{X}^{*\top}\alpha_0)$ refers to the density of $\mathbf{x}^{*\top}\alpha_0$. Then denote
$\hat{m}(x, \alpha_0) = (nh)^{-1}\sum_i^n K_h(d(x, \mathbf{X}_i^*))g(\mathbf{X}_i^*; Y_i, \alpha_0)$, $\hat{f}(x) = (nh)^{-1}\sum_i^n K_h(d(x, \mathbf{X}_i^*))$.
Define $t := \mathbf{x}^{*\top}\alpha_0$, $T := \mathbf{X}^{*\top}\alpha_0$. From Lemma B.1 of Newey (1994), we have

$$\sup_{\{\|t\|_2 \le C_1\}} |\hat{f}(t) - f(t)| = O_p\Big(\Big(\frac{\log n}{nh}\Big)^{1/2} + h^2\Big),$$

where C_1 refers to some positive constant. Further, we have

$$\begin{aligned}
&\mathbf{E}\hat{m}(t) = \frac{1}{h}\mathbf{E}(\mathbf{E}(Y|T)k(\frac{t-T}{h}))\\
&= \frac{1}{h}\int_{\mathcal{R}} k(\frac{t-T}{h})g^*(T, \alpha_0)f(T)dT\\
&= \int_{\mathcal{R}} k(u)m(t-hu, \alpha_0)du\\
&= \int_{\mathcal{R}/B} k(u)m(t-hu, \alpha_0)du + \int_B k(u)m(t-hu, \alpha_0)du\\
&= m(t)\int_{\mathcal{R}/B} k(u)du - hm_1(t)\int_{\mathcal{R}/B} uk(u)du\\
&+ m(t)\int_B k(u)du - hm_2(t)\int_{\mathcal{R}/B} uk(u)du\\
&= m(t) + O(h),
\end{aligned}$$

where $k(\cdot)$ refers to the standard normal distribution, B refers to the area corresponding to $D(\mathbf{X}^*) \ge 0$. and $m_1(\cdot)$ refer to the derivative of $m(\cdot)$ on $D(\mathbf{X}^*) \ge 0$ while $m_2(\cdot)$ refer to the derivative of $m(\cdot)$ on $D(\mathbf{X}^*) < 0$. So, we conclude

$$\sup_{\{\mathbf{x}^* | \|\mathbf{x}^*\|_2 \le M\}} |\hat{m}(x, \alpha_0) - m(x, \alpha_0)| = O_p\Big(\Big(\frac{\log n}{nh}\Big)^{1/2} + h\Big).$$

Since $m(x, \alpha_0) = g^*(x, \alpha_0)f(x)$, imply that

$$\sup_{\{\mathbf{x}^* | \|\mathbf{x}^*\|_2 \leq M\}} \left| \frac{\hat{m}(\alpha_0, x)}{f(x)} - g^*(\mathbf{x}^*, \alpha_0) \right| = O_p\Bigg(\Big(\frac{\log n}{n} \Big)^{1/3} \Bigg).$$

□

2.4 Algorithm

2.4.1 Two-stage Estimation Process

Having established the consistency of our estimation procedure, we were able to devise an efficient algorithm to implement our proposed method as follows:

Algorithm 1 Two-stage Algorithm

Initialization: Based on observations $\{(\mathbf{X}_i^*, Y_i)\}_{i=1}^n$, denote $B_1 := \{(\mathbf{X}_i^*, Y_i) | Y_i > 0\}_{i=1}^n$, $n_1 = \text{card}(\mathrm{B}_1)$,. Take the bandwidth of the kernel function $h = 0.1$.
Solve

$$\hat{\alpha} = \arg\min_\alpha \frac{1}{n_1} \sum_{i=1, i \in b_1}^{n_1} \Big(\frac{(1 + (\exp\{\mathbf{X}_i^{*\top}\alpha\})^2)Y_i}{\exp\{\mathbf{X}_i^{*\top}\alpha\}} \Big)^2 + \exp\{\mathbf{X}_i^{*\top}\alpha\} \Big(\exp\{\mathbf{X}_i^{*\top}\alpha\} - 2\Big(\frac{(1 + (\exp\{\mathbf{X}_i^{*\top}\alpha\})^2)Y_i}{\exp\{\mathbf{X}_i^{*\top}\alpha\}} \Big) \Big).$$

Classification: The $\{\hat{D}(\mathbf{X}_i^*)\}_{i=1}^n$ is estimated by (2.3).
Output: $\{\hat{\alpha}\}$ and the classifier$\{\hat{D}(\mathbf{X}_i^*)\}_{i=1}^n$.

We denote the estimators obtained in Algorithm 1 as WLSRF. In order to enhance the accuracy of our subsequent response, we plan to repeat two stages that are interconnected, unlike the conventional two-stage model. However, it is important to note that errors may accumulate with each iteration. Although the expected loss is guaranteed to decrease, there is a possibility that the empirical loss could increase. Therefore, it becomes necessary to evaluate the loss at every step.

2.4.2 Further Improvement in Algorithm

Theorem 2.3.

$$\mathcal{L}(\alpha^t, D^t(\mathbf{X}^*)) \geq \mathcal{L}(\alpha^{t+1}, D^{t+1}(\mathbf{X}^*)),$$

where $(\alpha^t, D^t(\mathbf{X}^*))$ *refers to the optimal solution of population loss function in* t *step of the Algorithm 2.*

Proof. In the t step, we have

$$\alpha^t = \arg\min_{\alpha} \mathbf{E}\bigg[\frac{1}{2f(\mathbf{X}^{*\top}\alpha)(1-f(\mathbf{X}^{*\top}\alpha))}\Big(Y - f(\mathbf{X}^{*\top}\alpha)\mathrm{I}_{\{\mathrm{D_{t-1}}(\mathbf{X}^*)\geq 0\}}\Big)^2\bigg].$$

Then according to equation (2.2), we have:

$$S_t(\mathbf{X}^*) = \mathrm{sign}(\mathbf{E}[-g(Y, \mathbf{X}^*; \alpha^t)|\mathbf{X}^*]).$$

Since (2.2) is the optimal condition for minimizing the expected loss function, we have $R(\alpha^t, S_{t-1}(\mathbf{X}^*)) \geq R(\alpha^t, S_t(\mathbf{X}^*))$ and the equation holds if and only if $S_{t-1} = S_t$.

Then the estimation of α_{t+1} was from

$$\alpha^{t+1} = \arg\min_{\alpha} \mathbf{E}\bigg[\frac{1}{2f(\mathbf{X}^{*\top}\alpha)(1-f(\mathbf{X}^{*\top}\alpha))}\Big(Y - f(\mathbf{X}^{*\top}\alpha)\mathrm{I}_{\{\mathrm{D_{t}}(\mathbf{X}^*)\geq 0\}}\Big)^2\bigg].$$

The global convexity of above optimization ensures that

$$\mathcal{L}(\alpha^{t+1}, S_t(\mathbf{X}^*)) \geq \mathcal{L}(\alpha^t, S_t(\mathbf{X}^*)),$$

and the equation holds if and only if $\alpha^{t-1} = \alpha^t$.

□

We further improve the algorithm as follows:

We denote the estimators obtained in Algorithm 2 as WLSR.

2.5 Numerical Study

2.5.1 Simulation

In the first part of the simulation study, we compared our method (including the results from the first step and the results from the iteration algorithm) with the following method:

```
Algorithm 2 Improved Algorithm
Folder generalization: Divide the data set to 5 equal folders randomly.
For k in range(0,5),
1. Initialization: Based on $\{(\mathbf{X}_i^*, Y_i)\}_{i=1}^n$, denote $B_1 := \{(\mathbf{X}_i^*, Y_i) | Y_i > 0\}_{i=1}^n$, $n_1 = \mathrm{card(B_1)}$, $b_1$ refers to the index set of $B_1$. Take the bandwidth of the kernel function $h = 0.1$, $T = 10$.
2. Solve
   $$\begin{aligned} \alpha_1 &= \arg\min_{\alpha} \frac{1}{n_1} \sum_{i=1, i \in b_1}^{n_1} \left( \frac{(1 + (\exp\{\mathbf{X}_i^{*\top}\alpha\})^2) Y_i}{\exp\{\mathbf{X}_i^{*\top}\alpha\}} \right)^2 \\ & \quad + \exp\{\mathbf{X}_i^{*\top}\alpha\} \left( \exp\{\mathbf{X}_i^{*\top}\alpha\} - 2 \left( \frac{(1 + (\exp\{\mathbf{X}_i^{*\top}\alpha\})^2) Y_i}{\exp\{\mathbf{X}_i^{*\top}\alpha\}} \right) \right). \end{aligned}$$
3. Classification: The $\{D_1(\mathbf{X}_i^*)\}_{i=1}^n$ is estimated by (2.3).
REPEAT: $t \to t + 1$
   $B_{t+1} := \{(\mathbf{X}_i^*, Y_i) | \hat{D}_t(\mathbf{X}^*) \geq 0\}_{i=1}^n$, $n_{t+1} = \mathrm{card(B_{t+1})}$,
   $b_{t+1}$ refers to the index set of $B_{t+1}$.
   Update: $\alpha_{t+1}$ based on Equation
   $$\begin{aligned} \alpha_{t+1} &= \arg\min_{\alpha} \frac{1}{n_{t+1}} \sum_{i=1, i \in b_{t+1}}^{n_{t+1}} \left( \frac{(1 + (\exp\{\mathbf{X}_i^{*\top}\alpha\})^2) Y_i}{\exp\{\mathbf{X}_i^{*\top}\alpha\}} \right)^2 \\ & \quad + \exp\{\mathbf{X}_i^{*\top}\alpha\} \left( \exp\{\mathbf{X}_i^{*\top}\alpha\} - 2 \left( \frac{(1 + (\exp\{\mathbf{X}_i^{*\top}\alpha\})^2) Y_i}{\exp\{\mathbf{X}_i^{*\top}\alpha\}} \right) \right). \end{aligned}$$
   Update: $\{D_{t+1}(\mathbf{X}_i^*)\}_{i=1}^n$ based on (2.3).
UntiL $t = T$.
End For.
Evaluation the prediction error of T iterations, select $\hat{\alpha}$ and $\hat{D}(\mathbf{X}^*)$ with minimum prediction error.
Output: $\hat{\alpha}$ and $\hat{D}(\mathbf{X}^*)$.
```

- WLSR: weighted least square regression method(our method)

- WLSRF: weighted least square regression method with one-time iteration.

- Two-stage method: use the non-zero data to fit a regression model and set all non-zero data as one then fit the classification part. The method is introduced in the case study [1]

[1] Improving fabric manufacturing efficiency through a hybrid quality rate prediction model

- Tobit model: transform non-zero Y into $\log(\sqrt{\frac{Y}{1-Y}})$, then fit it as regular Tobit model.
- MLE model: zero-inflated beta regression model based on the maximum likelihood method.
- ZIT model: zero-inflated Tobit model.

The data sets are generated as follows: set $\mathbf{X}^* = (1, x_1, x_2, ..., x_{p-1})$ from Gaussian distribution $N(\mu, \Sigma)$ where μ takes 0 value and Σ is the identity matrix. After obtaining the $\mathbf{X}^*$, Y came from $\mathrm{Bin}(m, p)$ with probability p and equals to zero with probability $(1-p)$, and $p = \frac{(\exp(\mathbf{X}^{*\top}\alpha))^2}{1+(\exp(\mathbf{X}^{*\top}\alpha))^2}\mathrm{I}_{\{\mathbf{X}^{*\top}\alpha \geq c\}}$. Denote $J = \mathrm{Bin}(m, p)$, then $Y = J/m$. We set three cut-off points in the simulation, which is $c = -1, -2, -3$, and we denote them as cases 1-3. For the training set, we take $n = 400$, and for the testing set, we take $n_1 = 200$. The simulation was repeated 100 times for each setting. The results in tables include the prediction error $pr := n_1^{-1}\sum_{i=1}^{n_1} \|Y_i - \hat{Y}_i\|_2$, where Y_i comes from the testing set, and the standard derivation of the prediction error.

- Model 1: we set $m = 30$, $p = 6$, $\alpha_0 = (-2, 2, -2.5, 1, -1, 2)$.
- Model 2, we set $m = 30$, $p = 4$, $\alpha_0 = (-2, 2, -2.5, 1)$.
- Model 3, we set $m = 30$, $p = 2$, $\alpha_0 = (-2, 2.5)$.
- Model 4: we set $m = 50$, $p = 6$, $\alpha_0 = (-2, 2, -2.5, 1, -1, 2)$.
- Model 5, we set $m = 50$, $p = 4$, $\alpha_0 = (-2, 2, -2.5, 1)$.
- Model 6, we set $m = 50$, $p = 2$, $\alpha_0 = (-2, 2.5)$.

- Model 7: we set $m = 100$, $p = 6$, $\alpha_0 = (-2, 2, -2.5, 1, -1, 2)$.
- Model 8, we set $m = 100$, $p = 4$, $\alpha_0 = (-2, 2, -2.5, 1)$.
- Model 9, we set $m = 100$, $p = 2$, $\alpha_0 = (-2, 2.5)$.

The data we present in Tables 2.1- 2.9 mainly simulates situations where the sampling size (m) is relatively small. In such cases, the occurrence of 'zero' comes from random chance increases, aligning more closely with the model assumption of MLE. As the sampling size gradually increases, it implies that if the sampling probability is non-zero, the observed value y is highly likely to be non-zero. The zero case becomes more aligned with our model assumption. From the final prediction error values, we can see that our model performs the best in these various scenarios.

In the second part, our focus shifts towards the convergence behavior of our method, specifically how the prediction error decreases as the value of n increases.

- In the first figure, we take $p = 2$, $n_1 = 5000$, $c = -1$, $m = 100$, $\alpha_0 = (-1.2, 0.8)$. The simulation was repeated 100 times.
- In the second figure, we take $p = 4$, $n_1 = 5000$, $c = -1$, $m = 100$, $\alpha_0 = (-0.8, 1, 0.6, -0.4)$. The simulation was repeated 100 times.
- In the third figure, we take $p = 6$, $n_1 = 5000$, $c = -1$, $m = 100$, $\alpha_0 = (0.7, 2.8, -0.7, 2.1, -0.7, 2.1)$. The simulation was repeated 100 times.
- In the fourth figure, we take $p = 8$, $n_1 = 5000$, $c = -1$, $m = 100$, $\alpha_0 = (0.3, 1.2, -0.3, 0.9, -0.3, 0.9, -0.3, 0.9)$. The simulation was repeated 100 times.

In this simulation, we could show that as n goes up, the prediction errors of each model decrease.

Table 2.1: Prediction error under model setting with $m = 30$, $p = 6$

Model 1	$c = -1$	$c = -2$	$c = -3$
WLSR	0.0123 (0.0015)	0.0127(0.0015)	0.0127(0.0015)
WLSRF	0.0129(0.0017)	0.0129(0.0016)	0.0128(0.0016)
Two-stage	0.0192(0.0019)	0.0188(0.0020)	0.0182(0.0019)
Tobit model	0.0385(0.0029)	0.0350(0.0028)	0.0347(0.0028)
MLE model	0.0857(0.0035)	0.0977(0.0036)	0.0996(0.0036)
ZIT model	0.0178(0.0017)	0.0172(0.0018)	0.0157(0.0016)

Table 2.2: Prediction error under model setting with $m = 30$, $p = 4$

Model 2	$c = -1$	$c = -2$	$c = -3$
WLSR	0.0109(0.0013)	0.0111(0.0014)	0.0111(0.0014)
WLSRF	0.0111(0.0014)	0.0111(0.0014)	0.0111(0.0014)
Two-stage	0.0172(0.0017)	0.0156(0.0017)	0.0150(0.0017)
Tobit model	0.0348(0.0025)	0.0312(0.0025)	0.0309(0.0025)
MLE model	0.0872(0.0033)	0.1007(0.0033)	0.1028(0.0032)
ZIT model	0.0161(0.0014)	0.0141(0.0015)	0.0130(0.0014)

Table 2.3: Prediction error under model setting with $m = 30$, $p = 2$

Model 3	$c = -1$	$c = -2$	$c = -3$
WLSR	0.0092(0.0012)	0.0094(0.0012)	0.0094(0.0012)
WLSRF	0.0093(0.0012)	0.0094(0.0012)	0.0094(0.0012)
Two-stage	0.0136(0.0013)	0.0116(0.0011)	0.0114(0.0011)
Tobit model	0.0293(0.0023)	0.0255(0.0022)	0.0252(0.0022)
MLE model	0.0828(0.0036)	0.0978(0.0040)	0.0999(0.0038)
ZIT model	0.0130(0.0013)	0.0107(0.0010)	0.0102(0.0010)

2.5.2 Applications on Real Datasets

We used three datasets for our analysis: AlcoholUse, Access to electricity in Brazil dataset, and InfMort.

- **The AlcoholUse dataset:** This dataset contains information on alcohol consumption among California students between 2008 and 2010, Our analysis of the AlcoholUse dataset provides insights into alcohol consumption patterns among public school students in grades 7, 9, and 11. It includes information on the percentage of students who reported the number of days they con-

Table 2.4: Prediction error under model setting with $m = 50$, $p = 6$

Model 4	$c = -1$	$c = -2$	$c = -3$
WLSR	0.0034 (0.0005)	0.0109(0.0014)	0.0109(0.0015)
WLSRF	0.0034(0.0005)	0.0110(0.0016)	0.0109(0.0015)
Two-stage	0.0041(0.0005)	0.0170(0.0022)	0.0153(0.0018)
Tobit model	0.0109(0.0016)	0.0309(0.0027)	0.0306(0.0027)
MLE model	0.0807(0.0034)	0.1055(0.0038)	0.1091(0.0038)
ZIT model	0.0040(0.0005)	0.0161(0.0021)	0.0138(0.0016)

Table 2.5: Prediction error under model setting with $m = 50$, $p = 4$

Model 5	$c = -1$	$c = -2$	$c = -3$
WLSR	0.0092(0.0013)	0.0093(0.0013)	0.0092(0.0013)
WLSRF	0.0093(0.0014)	0.0093(0.0014)	0.0092(0.0013)
Two-stage	0.0172(0.0018)	0.0130(0.0018)	0.0119(0.0015)
Tobit model	0.0307(0.0025)	0.0271(0.0024)	0.0267(0.0024)
MLE model	0.0917(0.0034)	0.1097(0.0033)	0.1135(0.0033)
ZIT model	0.0162(0.0016)	0.0124(0.0017)	0.0110(0.0013)

Table 2.6: Prediction error under model setting with $m = 50$, $p = 2$

Model 6	$c = -1$	$c = -2$	$c = -3$
WLSR	0.0074(0.0011)	0.0075(0.0011)	0.0075(0.0011)
WLSRF	0.0075(0.0011)	0.0075(0.0011)	0.0075(0.0011)
Two-stage	0.0129(0.0014)	0.0089(0.0010)	0.0089(0.0010)
Tobit model	0.0250(0.0022)	0.0213(0.0021)	0.0210(0.0020)
MLE model	0.0879(0.0038)	0.1073(0.0039)	0.1110(0.0038)
ZIT model	0.0123(0.0014)	0.0086(0.0010)	0.0081(0.0009)

sumed alcohol in the past 30 days, categorized by gender, grade level, and MedDays. We exclude students in Community Day Schools or Continuation Education. Our objective is to understand alcohol use trends among California students during this period. We assess the effectiveness of our method in predicting alcohol consumption patterns based on factors like gender, grade level, and MedDays. Through this analysis, we contribute insights to alcohol use research and provide evidence for the reliability of our method in predicting alcohol consumption behaviors among California students.

Table 2.7: Prediction error under model setting with $m = 100$, $p = 6$

Model 7	$c = -1$	$c = -2$	$c = -3$
WLSR	0.0017 (0.0004)	0.0089(0.0014)	0.0087(0.0014)
WLSRF	0.0017 (0.0004)	0.0089(0.0015)	0.0088(0.0014)
Two-stage	0.0026(0.0004)	0.0149(0.0020)	0.0115(0.0015)
Tobit model	0.0076(0.0011)	0.0265(0.0024)	0.0262(0.0023)
MLE model	0.0842(0.0037)	0.1122(0.0046)	0.1202(0.0046)
ZIT model	0.0025(0.0004)	0.0146(0.0020)	0.0110(0.0014)

Table 2.8: Prediction error under model setting with $m = 100$, $p = 4$

Model 8	$c = -1$	$c = -2$	$c = -3$
WLSR	0.0073(0.0012)	0.0074(0.0013)	0.0073(0.0013)
WLSRF	0.0074(0.0013)	0.0074(0.0013)	0.0073(0.0013)
Two-stage	0.0169(0.0020)	0.0106(0.0019)	0.0084(0.0012)
Tobit model	0.0261(0.0025)	0.0226(0.0023)	0.0223(0.0023)
MLE model	0.0962(0.0039)	0.1176(0.0041)	0.1255(0.0041)
ZIT model	0.0159(0.0018)	0.0104(0.0018)	0.0082(0.0012)

Table 2.9: Prediction error under model setting with $m = 100$, $p = 2$

Model 9	$c = -1$	$c = -2$	$c = -3$
WLSR	0.0052(0.0010)	0.0053(0.0010)	0.0053(0.0010)
WLSRF	0.0052(0.0010)	0.0053(0.0010)	0.0053(0.0010)
Two-stage	0.0116(0.0016)	0.0063(0.0009)	0.0058(0.0008)
Tobit model	0.0205(0.0020)	0.0170(0.0018)	0.0168(0.0018)
MLE model	0.0929(0.0036)	0.1151(0.0040)	0.1227(0.0042)
ZIT model	0.0111(0.0015)	0.0062(0.0010)	0.0056(0.0008)

- **Access to electricity in Brazil dataset:** The dataset pertained to electricity accessibility in Brazil, specifically focusing on cities within the Southeast and Northeast regions. The data, accessible at http://www.atlasbrasil.org.br/2013 /en/download/, reveals the correlation between the proportion of households with electricity and various socio-demographic variables of these cities which are studied in Santos and Bolfarine (2015b). The dataset comprises 500 cities, with all variables measured during the 2000 national census. According to the United Nations, as of 2009, 1.5 billion people worldwide lacked access to electricity. In developing nations, access to energy services can significantly al-

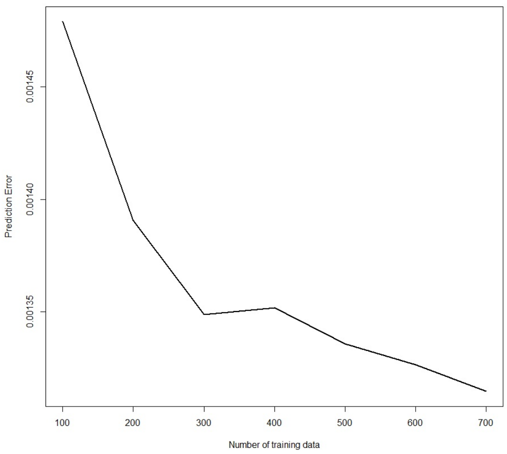

Figure 2.1: Setting with $p = 2$

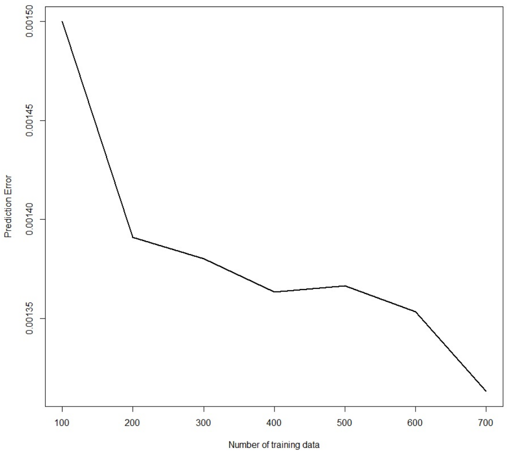

Figure 2.2: Setting with $p = 4$

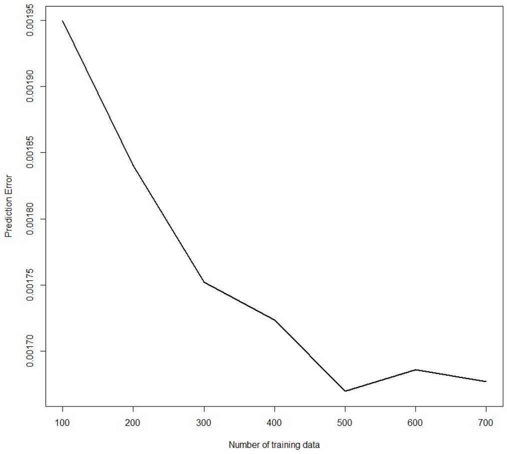

Figure 2.3: Setting with $p = 6$

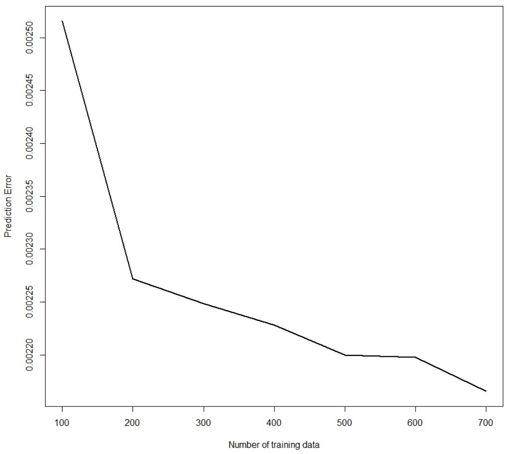

Figure 2.4: Setting with $p = 8$

leviate poverty, enhance public health, and stimulate economic growth, among other benefits. Given this context, our analysis aims to shed light on the relationship between electricity accessibility and socio-demographic factors. The response variable, the proportion of households in a city with electricity access ($PROP_ELEC$), was slightly adjusted to enhance the model's ability to estimate the probability of this proportion equalling one. Values nearing 1, specifically those exceeding 0.995, were rounded up to 1. Then we translate the response variable by $1 - PROP_ELEC$. While the Southeast region is among Brazil's most developed areas, the Northeast region remains one of the least developed. In our sample, the proportion of households with electricity access in the Southeast region ranges between 0.14 and 1, and between 0.54 and 1 in the Northeast region. The sample includes 81 cities where the response variable equals one. For covariates, we incorporated region ($REG = 0$ for Southwest, $REG = 1$ for Northeast), population (POP), per capita income (INCPC), human development index (HDI), and population density (DENS). All continuous variables were standardized before model fitting.

- **The InfMort dataset:** The dataset is based on the real data obtained from the Parana State in Brazil in 2010, which provided us with valuable information on the relationship between infant mortality and socio-economic factors in the Parana State. We explored indicators such as the FIRJAN city development index, illiteracy index, and income disparities to uncover the factors contributing to infant mortality rates. Our findings highlight the importance of addressing issues like city development, illiteracy, and income disparities to reduce infant mortality rates effectively. These results have implications for policymakers and healthcare professionals working towards improving infant health outcomes in the Parana State and beyond. By incorporating both datasets into our study, we were able to expand our understanding of alcohol use among California students and the factors influencing infant mortality rates in the Parana State. Our findings contribute to the existing body of knowledge in these fields and provide valuable insights for future research and policy-making efforts.

In the following applications, we normalized each covariate to have a better prediction on the response. For the AlcoholUse example, the estimator of the coefficients is $\hat{\alpha} = (-1.8386, 0.0993, -0.0421, 0.0297)$, which indicates that with higher grades, and shorter MedDays, students are more likely to drink.

In the access to electricity in Brazil example, we have $\hat{\alpha} = (1.6002, 0.0833, 0.0705, -0.0029, -5.4930, -0.7766)$. The findings suggest that the larger the population, the greater the proportion of insufficient power supply. In areas with higher income, HDI, and population density, the proportion of insufficient power supply is smaller. Northeast has a more severe shortage of power supply than Southwest.

In the infant mortality dataset, $\alpha = (-1.9559, 0.0596, -0.0134, -0.0116, -0.0202, 0.0356, -0.1477, -0.0286, -0.0162)$ indicated that higher FIRJAN index of city development, higher gross national product and higher proportion of covered by family health program results in lower infant mortality. As it is shown in the table, our model outperforms the remaining models in terms of prediction accuracy, and it can provide more valuable insights for industrial production as well as sociological research.

Table 2.10: Prediction Error of Real Applications

Datasets	WLSR	WLSRF	Two stage	ZIT
AlcoholUse	$1e-03$	$1e-03$	$1.37e-03$	$1.38e-03$
Access to ELEC	0.0122	0.0122	0.0130	0.0131
Infant Mortality	$1.5e-04$	$2.2e-04$	$2e-04$	$9.4e-04$

Chapter 3

Parametric Model of Zero-inflated Proportion Data Analysis

3.1 Model Construction

In the previous chapter, we employed a semiparametric model to handle our dataset, circumventing the discontinuity issue by estimating the sign of $D(\mathbf{X}^*)$ using the Nadaraya-Watson estimator. In this chapter, we address the indicator component by smoothing it with a continuous function. The key advantage of this approach is that it enables us to gain deeper insights into the covariates that determine the conditions under which the 'zero' value arises. To address the problem, we make the model more specific by formulating the optimization problem as

$$\mathcal{L} = \mathbf{E}\left\{ \frac{(1 + (\exp\{\mathbf{X}^{*\top}\alpha\})^2)^2}{2(\exp\{\mathbf{X}^{*\top}\alpha\})^2} \left(Y - \frac{(\exp\{\mathbf{X}^{*\top}\alpha\})^2}{1 + (\exp\{\mathbf{X}^{*\top}\alpha\})^2} \mathrm{I}_{\{\mathbf{X}^{*\top}\beta \geq 0\}} \right)^2 \right\}. \tag{3.1}$$

Here we use a linear classification condition to help us identify the key factors affecting the unsuitability i.e. zero. One problem of our formulation is the identification of β, but it could be easily solved by putting a restriction on the L_2 norm of β_0. We will show the assumptions later. Another obvious problem is that the loss function is not continuously differentiable and does not possess an explicit solution for α and β. Inspired by Horowitz (1992) we introduce the kernel smoothing function to replace the indicator function.

$$\mathcal{L}_1 = \mathbf{E}\left\{\frac{(1+(\exp\{\mathbf{X}^{*\top}\alpha\})^2)^2}{2(\exp\{\mathbf{X}^{*\top}\alpha\})^2}\left(Y - \frac{(\exp\{\mathbf{X}^{*\top}\alpha\})^2}{1+(\exp\{\mathbf{X}^{*\top}\alpha\})^2}\Phi\left\{\frac{\mathbf{X}^{*\top}\beta}{h}\right\}\right)^2\right\}. \tag{3.2}$$

The coordinate descent method is considered to be applied in the optimization of $\hat{\mathcal{L}}_1$. For α, the global convexity is proven when β is known. For β, we could also obtain the local convexity around the true β_0 when α is known.

3.2 Main Theorems

3.2.1 Consistency Property

Note that there is an indicator function in the oracle loss (3.1). Direct analysis of the 0-1 loss (which is not continuous) is analytically challenging, and one of the popular treatment in the literature of classification is to replace it by a smooth surrogate loss; see for example Bartlett et al. (2006) and the references therein. One of the popular technique, known as the smoothed "maximum score estimator", is to approximate the 0-1 by a smoothed loss based on kernel smoothing. (Horowitz, 1992). In the maximum score estimation, one aims to obtain the estimated classifiers by maximizing $S_N(b) = \frac{1}{N}\sum_{n=1}^{N}[2 * \mathbf{I}(y_n = 1) - 1]\mathbf{I}(b^\top x_n \geq 0)$, which is quite similar to our estimation equation. The difficulty of the score function came from the indicator function $\mathbf{I}_{\{\mathbf{X}^*\beta\geq 0\}}$. Discontinuity of the function $\mathcal{L}$ leads to the slow rate of convergence and complexity of inference for the estimator. Similarly, we solved this problem with the kernel method replacing the original indicator function with a sufficiently smooth function $K(.)$. To guarantee the consistency of the estimator, we put a few assumptions to ensure the identifiability and consistency of the estimators when the sample size goes to infinity. We remark that our problem is very different from the classical classification problems, as our loss function (3.1) is confounded with both a 0-1 loss and a continuous component.

Assumption 3.1. *Assumption for smooth function K(x)*

$|K(x)| \leq M$ *for some fixed* M *and all* x *in* $(-\infty, \infty)$,

$\lim_{x\to -\infty} K(x) = 0$ *and* $\lim_{x\to\infty} K(x) = 1$,

$|K'(x)| \leq m$ *for some fixed* m *and all* x *in* $(-\infty, \infty)$.

Assumption 3.2. *Distribution of (y,x)*

The support of F_x *is not contained in any proper linear subspace of* R^n.

$0 < P[Y \geq 0|x]$<1, *for almost every* x.

Assumption 3.3. *Restriction on* α, β

To ensure the identifiability of β, *we assume that* β *is drawn from the unit sphere, i.e.* $\|\beta\|_2 = 1$. *For* α, *we assume that there exists a constant M, s.t.* $\|\alpha\|_2 \leq M$.

The identifiability of β could be ensured with assumptions above, as discussed in Manski (1985). Then, we propose to estimate the oracle loss function via the following smoothed empirical loss function:

$$\hat{\mathcal{L}}_1 = \frac{1}{n}\sum_{i=1}^{n} \frac{(1+(\exp\{\mathbf{X}_i^{*\top}\alpha\})^2)^2}{2(\exp\{\mathbf{X}_i^{*\top}\alpha\})^2}\Bigg(Y_i - \frac{(\exp\{\mathbf{X}_i^{*\top}\alpha\})^2}{1+(\exp\{\mathbf{X}_i^{*\top}\alpha\})^2}\Phi\bigg\{\frac{\mathbf{X}_i^{*\top}\beta}{h}\bigg\}\Bigg)^2 \qquad (3.3)$$

Note that for any given bandwidth h, the loss function (3.3) is differentiable, and as a result, consistency and asymptotic normality under some specific cases become feasible.

For the sake of concise expression, the original loss function is written as :

$$\mathcal{L} = \min_{\alpha,\beta} \mathbf{E}\left[\frac{Y^2}{2f(\mathbf{X}^*)(1-f(\mathbf{X}^*))}\right] + \mathbf{E}\left[\frac{(f(\mathbf{X}^*) - 2Y)I_{\{G(\mathbf{X}^*)\geq 0\}}}{2(1-f(\mathbf{X}^*))}\right], \qquad (3.4)$$

where $f(\mathbf{X}^*) = \frac{(\exp\{\mathbf{X}^{*\top}\alpha\})^2}{1+(\exp\{\mathbf{X}^{*\top}\alpha\})^2}$ and $G(\mathbf{X}^*) = \mathbf{X}^{*\top}\beta$. Since model satisfied equation 1.1, we could guarantee the unique solution of the equation 3.1. Denote the second part in 3.4 as $\mathcal{R}_0$, then inspired by Horowitz (1992), we replace the indicator function with the kernel function in the second part as follows:

$$\mathcal{R}_1(\alpha, \beta) = \min_{\alpha,\beta} E_{W,Z} W K(\frac{Z}{h}), \tag{3.5}$$

where $W = \frac{(f(\mathbf{X}^*)-2Y)}{f(\mathbf{X}^*)}$ and $Z = G(\mathbf{X}^*)$, $K(\cdot)$ is kernel function satisfied our general assumption 3.1, h is the bandwidth.

Since the uniform convergence of the parameters is closely tied to the uniform convergence of the loss function, and as mentioned in the previous section, Rao (1962) established a classical theorem connecting the uniform convergence of measures with weak convergence under certain conditions. We divide our proof into the following steps. First, we prove the uniform convergence between $\mathcal{R}_0$, which refers to the original loss function with the indicator function, and $\mathcal{R}_1$, which refers to the population loss function that replaces the indicator function with a certain kernel function. Then, we aim to construct the uniform convergence of the empirical loss $\mathcal{R}_{1n}$ to $\mathcal{R}_1$.

The problematic aspect lies in the equicontinuity property since the derivative involves the term h^{-1}, leading to the result that $h^{-1}K(\frac{G(\mathbf{X})}{h})G(\mathbf{X}^*) \to 0$ diverges as $h \to 0$ for some θ and $\mathbf{X}^*$. To overcome this issue, we introduce a truncated kernel function as a solution. First, we stated the following lemmas:

Lemma 3.1. $\mathcal{R}_1 \to \mathcal{R}_0$ *as* $h \to 0$.

Further, we denote $\hat{\mathcal{R}_1} = \frac{1}{n}\sum_{i=1}^{n} W_i K(\frac{G(\mathbf{X}_i^*)}{h})$, then, if we could prove that $\hat{\mathcal{R}_1}$ converges to $\mathcal{R}_1$ uniformaly as $h \to 0$, then we completed the proof. We intend to take advantage of theorem 3.1 proposed by Rao (1962), which requires the bounded and equicontinuous properties on the function space. The bounded property could be guaranteed for any $\mathbf{X}^*$ and bound assumption on α and β. We state our general ideas formally as follows:

Set the point C_h, s.t. $\lim_{h\to 0} C_h = 0$, and

$$\lim_{h\to 0}\frac{1}{h}K'(\frac{C_h}{h}) = d,$$

where K is the first derivation of K while d is a finite constant.
Then, we made the following truncation:

$$K_1(x) = \begin{cases} K(x) & x \le -C_h \\ K(-C_h) + (x + C_h)(\dfrac{1 - 2K(-C_h)}{2C_h}) & -C_h < x < C_h \\ K(x) & x \ge C_h, \end{cases}$$

In this case, we denoted

$$T_1(x) = WK_1(\frac{G(\mathbf{X}^*)}{h}),$$

then,

$$T_1'(x) = \begin{cases} [2f(\mathbf{X}^*)f'(\mathbf{X}^*)] \cdot K(\dfrac{G(\mathbf{X}^*)}{h}) + G'(\mathbf{X}^*) \cdot \dfrac{1}{h}K'(\dfrac{G(\mathbf{X}^*)}{h}) \cdot W & x \le -C_h \\ \dfrac{1 - 2K(-C_h)}{2C_h} & -C_h < x < C_h \\ [2f(\mathbf{X}^*)f'(\mathbf{X}^*)] \cdot K(\dfrac{G(\mathbf{X}^*)}{h}) + G'(\mathbf{X}^*) \cdot \dfrac{1}{h}K'(\dfrac{G(\mathbf{X}^*)}{h}) \cdot W & x \ge C_h. \end{cases}$$

Under our assumptions, it is easy to verify that $T_1'(x)$ is bounded for any $\mathbf{X}^*$ and bounded α, β.

Theorem 3.1. (Consistency) *Let assumption 1-3 hold, then,*

$$\lim_{n\to\infty}\sup_{\alpha,\beta}|\hat{\mathcal{L}}_1 - \mathcal{L}| = 0.$$

Since the solution (α_0, β_0) *to* $\arg\min \mathcal{L}$ *is unique, we further obtained* $(\hat{\alpha}, \hat{\beta}) \xrightarrow{p} (\alpha_0, \beta_0)$*, as* $n \to \infty$*.*

Notations:, the $\hat{\mathcal{L}_1}$ is not just the empirical loss function that replace the indicator part with any kernel function satisfied assumption 3.1, but the modification vision of the kernel function as we illustrated above.

Having established the consistency property, we proceed to examine the solution process for this loss function. Within our algorithm, we maintain the approach of initially fixing one set of parameters and solving for the other, subsequently iterating this process until converging to the optimal solution that minimizes the loss. During this iterative procedure, the following two propositions can be leveraged.

Proposition 3.1. *Given* α*, the expectation of the loss function*

$$\hat{\mathcal{L}_1} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{(1+(\exp\{\mathbf{X}_i^{*\top}\alpha\})^2)^2}{2(\exp\{\mathbf{X}_i^{*\top}\alpha\})^2}\left(Y_i - \frac{(\exp\{\mathbf{X}_i^{*\top}\alpha\})^2}{1+(\exp\{\mathbf{X}_i^{*\top}\alpha\})^2}\Phi\left\{\frac{\mathbf{X}_i^{*\top}\beta}{h}\right\}\right)^2\right\}$$

is locally convex concerning β *in probability.*

Proposition 3.2. *Given* β*, loss function*

$$\hat{\mathcal{L}_1} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{(1+(\exp\{\mathbf{X}_i^{*\top}\alpha\})^2)^2}{2(\exp\{\mathbf{X}_i^{*\top}\alpha\})^2}\left(Y_i - \frac{(\exp\{\mathbf{X}_i^{*\top}\alpha\})^2}{1+(\exp\{\mathbf{X}_i^{*\top}\alpha\})^2}\Phi\left\{\frac{\mathbf{X}_i^{*\top}\beta}{h}\right\}\right)^2\right\}$$

is global convex concerning α *in probability.*

Along with proposition 3.1 and 3.2, the iterative algorithm could be applied to solve our model. Besides, we could further deduce the inference results of $(\hat{\alpha}, \hat{\beta})$.

3.2.2 Inference Results

Denote $\hat{J}(\theta) = \frac{\partial \hat{\mathcal{L}}_1}{\partial \theta}$, where $\theta = (\alpha, \beta)$. Then according to the asymptotic distribution developed based on Taylor expansion of $\hat{J}(\theta)$, for large n,

$$\hat{J}(\theta_n) = \hat{J}(\theta_0) + Q_n(\bar{\theta})(\theta_n - \theta_0) = 0,$$

where $\bar{\theta} \in \Theta := \{\theta \text{ lies between } \theta_0 \text{ and } \theta_\text{n}\}$, θ_0 referring to the true value and θ_n referring to the optimal solution to empirical loss function. Define a $2p$ dimensional diagonal matrix K_n whose first p elements are 1 and the others are $\sqrt{h}$. Then, we have

$$\sqrt{n}K_n^{-1}(\theta_n - \theta_0) = \begin{pmatrix} \frac{\partial \hat{\mathcal{L}}_1^2(\theta)}{\partial \alpha^2}|_{\theta=\bar{\theta}} & \sqrt{h}\frac{\partial \hat{\mathcal{L}}_1^2(\theta)}{\partial \alpha \partial \beta}|_{\theta=\bar{\theta}} \\ \sqrt{h}\frac{\partial \hat{\mathcal{L}}_1^2(\theta)}{\partial \alpha \partial \beta}|_{\theta=\bar{\theta}} & h\frac{\partial \hat{\mathcal{L}}_1^2(\theta)}{\partial \beta^2}|_{\theta=\bar{\theta}} \end{pmatrix}^{-1} \begin{pmatrix} \sqrt{n}\hat{J}_n^{\alpha}(\theta_0) \\ \sqrt{nh}\hat{J}_n^{\beta}(\theta_0), \end{pmatrix}$$

where $\hat{J}_n^{\alpha}(\cdot)$ refers to the first derivation of $\hat{\mathcal{L}}_1(\cdot)$ with respect of α,and $\hat{J}_n^{\beta}(\cdot)$ refers to the first derivation of $\hat{\mathcal{L}}_1(\cdot)$ with respect of β.

To discuss the asymptotic property of $\theta_n - \theta_0$, we will separate it into three parts. First, we will prove the $\sqrt{n}\hat{J}_n^{\alpha}(\theta_0)$ and $\sqrt{nh}\hat{J}_n^{\beta}(\theta_0)$ are asymptotically negligible. The we will calculate the limitation of $\frac{\partial \hat{\mathcal{L}}_1^2(\theta)}{\partial \alpha^2}|_{\theta=\bar{\theta}}$, $\sqrt{h}\frac{\partial \hat{\mathcal{L}}_1^2(\theta)}{\partial \alpha \partial \beta}|_{\theta=\bar{\theta}}$, and $h\frac{\partial \hat{\mathcal{L}}_1^2(\theta)}{\partial \beta^2}|_{\theta=\bar{\theta}}$. Since the uniform convergence is proven in the previous section, it is not hard to derive the limitations equal to $\mathbf{E}\left[\frac{\partial \hat{\mathcal{L}}_1^2(\theta)}{\partial \alpha^2}|_{\theta=\theta_0}\right]$, $\sqrt{h}\mathbf{E}\left[\frac{\partial \hat{\mathcal{L}}_1^2(\theta)}{\partial \alpha \partial \beta}|_{\theta=\theta_0}\right]$, $\mathbf{E}\left[h\frac{\partial \hat{\mathcal{L}}_1^2(\theta)}{\partial \beta^2}|_{\theta=\theta_0}\right]$ respectively.

In the last part, we will derive the expectation and variance of $\sqrt{n}\hat{J}_n^{\alpha}(\theta_0)$ and $\sqrt{nh}\hat{J}_n^{\beta}(\theta_0)$.

Assumption 3.4. *Assume $\beta_0 = (b_1, \cdots, b_p)$, and b_1 is bound away from zero.*

The assumption could be achieved as long as $\beta_0 \neq 0$. Then we could exchange corresponding covariates and move the first non-zero element to b_1.

Assumption 3.5. *Assume that the expectation of covariates* $\mathbf{X}^*$ *is uniformly bounded by a constant* $E_1 < \infty$.

Assumption 3.6. *There exists a constant* $c > 0$ *satisfying that* $|\mathbf{X}^{*\top}\beta_0| \geq c$ *almost everywhere.*

Theorem 3.2. *Let* $\theta_n \to \theta_0$, *assume assumptions 3.4 and 3.6 hold,* $nh^2 \to 0$. *Then,*

$$\sqrt{n}(\hat{\alpha} - \alpha_0) \Rightarrow N(0, \mathcal{D}_1^{-\top}\mathcal{V}\mathcal{D}_1^{-1}),$$

$$\sqrt{nh}(\hat{\beta} - \beta_0) \Rightarrow N(0, \mathcal{D}_2^{-\top}\mathcal{W}\mathcal{D}_2^{-1}),$$

where

$$\mathcal{D}_1 = \mathbf{E}\Bigg[\mathbf{E}\Bigg[\mathbf{X}^*\mathbf{X}^{*\top}2(\exp\{\mathbf{X}^{*\top}\alpha\})^2\Bigg(Y - \Phi\Bigg\{\frac{\mathbf{X}^{*\top}\beta}{h}\Bigg\}\Bigg)^2 + 2Y^2(\exp\{\mathbf{X}^{*\top}\alpha\})^{-2}\Bigg]\Bigg|\mathbf{X}^*\Bigg]$$

,

$$\mathcal{D}_2 = -(b_1)^{-1}\int\int_{z\geq 0} Aa^2\mathbf{X}^*\mathbf{X}^{*\top}z\Bigg\{\Bigg(1 - \Phi(z)\Bigg)\Phi'''(z) - \Phi'(z)\Phi''(z)\Bigg\}dF(hz|\widetilde{\mathbf{X}^*})dF(\widetilde{\mathbf{X}^*})$$

$$+(b_1)^{-1}\int\int_{z<0} Aa^2\mathbf{X}^*\mathbf{X}^{*\top}z\Bigg\{\Phi(z)\Phi'''(z) + \Phi'(z)\Phi''(z)\Bigg\}dF(hz|\widetilde{\mathbf{X}^*})dF(\widetilde{\mathbf{X}^*})$$

$\mathcal{V} = \mathbf{E}\Bigg[\mathbf{X}^*\mathbf{X}^{*\top}[H(\alpha_0, \beta_0, \mathbf{X}^*)]\Bigg]$, *and*

$$H(\alpha_0, \beta_0, \mathbf{X}^*) = \mathbf{E}\Bigg[\Bigg[\frac{(\exp\{\mathbf{X}^{*\top}\alpha_0\})^4\big(Y - \Phi\{\frac{\mathbf{X}^{*\top}\beta_0}{h}\}\big)^2 - Y^2}{(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2}\Bigg]^2\Bigg|\mathbf{X}^*\Bigg]$$

is a bounded function, and

$$\mathcal{W} = b_1^{-1}\int\int \mathbf{X}^*\mathbf{X}^{*\top}A^2a^3\mathrm{I}_{\{hz\geq 0\}}\Phi'(z)^2dF(hz|\widetilde{\mathbf{X}^*})dF(\widetilde{\mathbf{X}^*}).$$

are matrices with bounded elements. The notations $A = \frac{(1+(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2)^2}{(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2}, a = \frac{(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2}{1+(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2}$.

Lemma 3.2. *Let $\theta_n \to \theta_0$, assume assumptions 3.4 and 3.5 hold, then*

$$\sqrt{h}\frac{\partial \hat{\mathcal{L}}_1^2(\theta)}{\partial\alpha\partial\beta}\bigg|_{\theta=\bar{\theta}} \xrightarrow{p} 0,$$

as $n \to \infty$ and $h \to 0$.

Lemma 3.3. *Let $\theta_n \to \theta_0$, assume assumptions 3.4 and 3.5 hold, Then,*

$$\mathbf{E}[\sqrt{n}\hat{J}_n^{\alpha}(\theta)|_{\theta=\theta_0}] = 0$$

as long as $nh^4 \to 0$.

Lemma 3.4. *Assume assumptions 3.4 and 3.5 hold, Then*

$$\frac{\partial \hat{\mathcal{L}}_1^2(\theta)}{\partial\alpha^2}\bigg|_{\theta=\bar{\theta}} \xrightarrow{p} \mathcal{D}_1,$$

where

$$\mathcal{D}_1 = \mathbf{E}\bigg[\mathbf{E}\bigg[\mathbf{X}^*\mathbf{X}^{*\top}2(\exp\{\mathbf{X}^{*\top}\alpha\})^2\bigg(Y - \Phi\bigg\{\frac{\mathbf{X}^{*\top}\beta}{h}\bigg\}\bigg)^2 + 2Y^2(\exp\{\mathbf{X}^{*\top}\alpha\})^{-2}\bigg]\bigg|\mathbf{X}^*\bigg]$$

is constant matrix.

Lemma 3.5. *Assume Assumption 3.5 hold, Then*

$$\frac{\partial \hat{\mathcal{L}}_1^2(\theta)}{\partial\beta^2}\bigg|_{\theta=\bar{\theta}} \xrightarrow{p} \mathcal{D}_2,$$

where

$$\mathcal{D}_2 = -(b_1)^{-1}\int\int_{z\geq 0} Aa^2\mathbf{X}^*\mathbf{X}^{*\top}z\bigg\{\bigg(1-\Phi(z)\bigg)\Phi'''(z) - \Phi'(z)\Phi''(z)\bigg\}dF(hz|\widetilde{\mathbf{X}^*})dF(\widetilde{\mathbf{X}^*})$$
$$+(b_1)^{-1}\int\int_{z<0} Aa^2\mathbf{X}^*\mathbf{X}^{*\top}z\bigg\{\Phi(z)\Phi'''(z) + \Phi'(z)\Phi''(z)\bigg\}dF(hz|\widetilde{\mathbf{X}^*})dF(\widetilde{\mathbf{X}^*})$$

is a constant matrix.

Lemma 3.6. *Let $\theta_n \to \theta_0$, assume assumptions 3.4 and 3.5 hold, Then*

$$\mathbf{E}[\sqrt{nh}\hat{J}_n^\beta(\theta)|_{\theta=\theta_0}] = 0,$$

as long as $nh^3 \to 0$

Lemma 3.7. *Let $\theta_n \to \theta_0$, assume assumptions 3.4 and 3.5 hold, Then covariance between $\sqrt{nh}\hat{J}_n^\beta(\theta)$ and $\sqrt{n}\hat{J}_n^\alpha(\theta)$ are asymptotically negligible as long as $nh^2 \to 0$.*

Lemma 3.8. *Assume assumptions 3.4 $\sim$ 3.6 hold,*

$$\mathrm{Var}(\sqrt{\mathrm{n}}\hat{\mathrm{J}}_{\mathrm{n}}^\alpha(\theta_0)) = \mathcal{V},$$

where $\mathcal{V} = \mathbf{E}\Big[\mathbf{X}^*\mathbf{X}^{*\top}[H(\alpha_0, \beta_0, \mathbf{X}^*)]\Big]$, and

$$H(\alpha_0, \beta_0, \mathbf{X}^*) = \mathbf{E}\left[\left[\frac{(\exp\{\mathbf{X}^{*\top}\alpha_0\})^4\big(Y - \Phi\{\frac{\mathbf{X}^{*\top}\beta_0}{h}\}\big)^2 - Y^2}{(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2}\right]^2\middle|\mathbf{X}^*\right]$$

is a bounded function.

Lemma 3.9. *Assume assumptions 3.4 $\sim$ 3.6 hold, then*

$$\mathrm{Var}(\sqrt{\mathrm{nh}}\hat{\mathrm{J}}_{\mathrm{n}}^\beta(\theta_0)) = \mathcal{W},$$

where

$$\mathcal{W} = b_1^{-1}\int\int \mathbf{X}^*\mathbf{X}^{*\top}A^2a^3\mathrm{I}_{\{hz\geq 0\}}\Phi'(z)^2 dF(hz|\widetilde{\mathbf{X}^*})dF(\widetilde{\mathbf{X}^*}).$$

3.2.3 Proof

Proof of Lemma 3.1

Proof. Denote the density of W and Z as $f_\theta(w, z)$. Let $k(x)$ be a density of kernel function. Under standard assumptions in kernel smoothing (assumptions for $K(x)$

and smoothness assumptions for $f_\theta(w,z)$)

$$\begin{aligned}
&\mathbf{E}_{W,Z}\frac{1}{h^2}k\Big(\frac{w-W}{h}\Big)k\Big(\frac{z-Z}{h}\Big)\\
=\;& \int\int \frac{1}{h^2}k\Big(\frac{w-w'}{h}\Big)k\Big(\frac{z-z'}{h}\Big)f_\theta(w',z')dw'dz'\\
=\;& \int\int k(t_1)k(t_2)f_\theta(w+ht_1,z+ht_2)dt_1dt_2\\
=\;& f_\theta(w,z)+O(h^2),
\end{aligned}$$

where $E_{W,Z}$ indicates that the expectation is taken over (W,Z). Note that W is bounded, and without loss of generality we also assume that the X_i's are bounded variables. Such an assumption can always be relaxed using truncation techniques. We thus have

$$\begin{aligned}
\mathbf{E}WI_{\{Z\geq 0\}} &= \int\int wI_{\{z\geq 0\}}f_\theta(w,z)dwdz\\
&= \int\int wI_{\{z\geq 0\}}E_{W,Z}\frac{1}{h^2}k\Big(\frac{w-W}{h}\Big)k\Big(\frac{z-Z}{h}\Big)dwdz+O(h^2)\\
&= \mathbf{E}_{W,Z}W\left(1-\Phi\left(-\frac{Z}{h}\right)\right)+O(h^2),\\
&= \mathbf{E}_{W,Z}WK\left(\frac{Z}{h}\right)+O(h^2), \qquad (3.6)
\end{aligned}$$

where $K(\cdot)$ is the CDF corresponding to $k(\cdot)$, and the last step holds as $K(\cdot)$ is symmetric.

Then, we could prove that $\mathcal{R}_1(\alpha,\beta)$ converges to $\mathcal{R}_0(\alpha,\beta)$ uniformly as $h\to 0$. □

Proof of Theorem 3.1

Combine assumptions 3.1-3.3 and uniformly convergence results, theorem 4.1.1 of Amemiya (1985) implies the uniform convergence of $\hat{\theta}=(\hat{\alpha},\hat{\beta})$ to $\theta=(\alpha,\beta)$.

Denote $\hat{\mathcal{R}}_1 = \frac{1}{n}\sum_{i=1}^n W_i K(\frac{G(\mathbf{X}_i^*)}{h})$, where $K(\cdot)$ refers to the truncated kernel function, we have $\hat{\mathcal{R}}_{1n}$ converges to $\mathcal{R}_0$ uniformly.

Further, the theorem 4.1.1 brought up by Manski (1985) could be applied to verify that $(\hat{\alpha}, \hat{\beta}) \rightarrow_p (\alpha, \beta)$ with our assumption and uniformly convergence results. then we completed the proof.

Proof of Proposition 3.1

Proof. We simplify our loss function as $\hat{\mathcal{L}}_1 = \frac{1}{n}\sum_{i=1}^n \bigg\{A_i\bigg(Y_i - a_i\Phi\bigg\{\frac{\mathbf{X}_i^{*\top}\beta}{h}\bigg\}\bigg)^2\bigg\}$, where $A_i = \frac{(1+(\exp\{\mathbf{X}_i^{*\top}\alpha\})^2)^2}{(\exp\{\mathbf{X}_i^{*\top}\alpha\})^2}$, $a_i = \frac{(\exp\{\mathbf{X}_i^{*\top}\alpha\})^2}{1+(\exp\{\mathbf{X}_i^{*\top}\alpha\})^2}$. Then, we have

$$\frac{\partial\hat{\mathcal{L}}_1}{\partial\beta} = -\frac{1}{n}\sum_{i=1}^n A_i a_i \frac{\mathbf{X}_i^*}{h}\bigg(Y_i - a_i\Phi\bigg\{\frac{\mathbf{X}_i^{*\top}\beta}{h}\bigg\}\bigg)\Phi'\bigg\{\frac{\mathbf{X}_i^{*\top}\beta}{h}\bigg\}.$$

Further, we could derive

$$\begin{aligned}
&\frac{\partial^2\hat{\mathcal{L}}_1}{\partial\beta^2}\\
= & -\frac{1}{n}\sum_{i=1}^n A_i a_i \frac{\mathbf{X}_i^*\mathbf{X}_i^{*\top}}{h^2}\bigg\{\bigg(Y_i - a_i\Phi\bigg\{\frac{\mathbf{X}_i^{*\top}\beta}{h}\bigg\}\bigg)\Phi''\bigg\{\frac{\mathbf{X}_i^{*\top}\beta}{h}\bigg\} - a_i\bigg(\Phi'\bigg\{\frac{\mathbf{X}_i^{*\top}\beta}{h}\bigg\}\bigg)^2\bigg\}\\
= & -\frac{1}{n}\sum_{i=1}^n A_i a_i \frac{\mathbf{X}_i^*\mathbf{X}_i^{*\top}}{h^2}\bigg\{\bigg(Y_i - a_i\Phi\bigg\{\frac{\mathbf{X}_i^{*\top}\beta}{h}\bigg\}\bigg)\Phi''\bigg\{\frac{\mathbf{X}_i^{*\top}\beta}{h}\bigg\}\bigg\}\\
& +\frac{1}{n}\sum_{i=1}^n A_i a_i^2 \frac{\mathbf{X}_i^*\mathbf{X}_i^{*\top}}{h^2}\bigg(\Phi'\bigg\{\frac{\mathbf{X}_i^{*\top}\beta}{h}\bigg\}\bigg)^2. \qquad (3.7)
\end{aligned}$$

Since the expectation of (3.8) equals to

$$-\frac{1}{n}\sum_{i=1}^n \mathbf{E}\bigg\{A_i a_i \frac{\mathbf{X}_i^*\mathbf{X}_i^{*\top}}{h^2}\bigg\{\bigg(a_i \mathrm{I}_{\{\mathbf{X}_i^{*\top}\beta_0 \geq 0\}} - a_i\Phi\bigg\{\frac{\mathbf{X}_i^{*\top}\beta}{h}\bigg\}\bigg)\Phi''\bigg\{\frac{\mathbf{X}_i^{*\top}\beta}{h}\bigg\}\bigg\}\bigg\}$$

the expectation of (3.7) is positive definite and for (3.2.3), we need to discuss it separately. When $\mathbf{X}^{*\top}\beta_0 \geq 0$ holds, then in the small ball $B(\beta_0, \delta)$, where $\delta > 0$ is the radius of the ball that ensures that for every $\beta \in B(\beta_0, \delta)$, $\mathbf{X}^{*\top}\beta \geq 0$. Then we have $\left(a\mathrm{I}_{\{\mathbf{X}^{*\top}\beta_0 \geq 0\}} - a\Phi\left\{\frac{\mathbf{X}^{*\top}\beta}{h}\right\}\right) > 0$ and $\Phi''\left\{\frac{\mathbf{X}^{*\top}\beta}{h}\right\} < 0$, then (3.2.3) is positive definite in probability. Similarly, we could obtain the same conclusion when $\mathbf{X}^{*\top}\beta_0 < 0$.

□

Proof of Proposition 3.2

Proof. Since

$$\hat{\mathcal{L}}_1 = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Y_i(1+(\exp\{\mathbf{X}_i^{*\top}\alpha\})^2)}{2\exp\{\mathbf{X}_i^{*\top}\alpha\}} - \frac{1}{2}\exp\{\mathbf{X}_i^{*\top}\alpha\}\Phi\left\{\frac{\mathbf{X}_i^{*\top}\beta}{h}\right\}\right]^2,$$

so, we get the derivative of $\mathcal{L}$ with respect of α_j is

$$\frac{\partial\hat{\mathcal{L}}_1}{\partial\alpha_j} = \frac{1}{n}\sum_{i=1}^{n}X_{ij}\frac{(\exp\{\mathbf{X}_i^{*\top}\alpha\})^4\left(Y_i - \Phi\left\{\frac{\mathbf{X}_i^{*\top}\beta}{h}\right\}\right)^2 - Y_i^2}{(\exp\{\mathbf{X}_i^{*\top}\alpha\})^2}.$$

Further, we get the Hessian matrix with respect to α is

$$\frac{\partial^2\hat{\mathcal{L}}_1}{\partial\alpha_j\partial\alpha_k} = \frac{1}{n}\sum_{i=1}^{n}X_{ij}X_{ik}2(\exp\{\mathbf{X}_i^{*\top}\alpha\})^2\left(Y_i - \Phi\left\{\frac{\mathbf{X}_i^{*\top}\beta}{h}\right\}\right)^2 + 2Y_i^2(\exp\{\mathbf{X}_i^{*\top}\alpha\})^{-2}.$$

We transform the above equation to matrix form, denote

$$\bar{\mathbf{X}}_i^* = \mathbf{X}_i^*\sqrt{2(\exp\{\mathbf{X}_i^{*\top}\alpha\})^2\left(Y_i - \Phi\left\{\frac{\mathbf{X}_i^{*\top}\beta}{h}\right\}\right)^2 + 2Y_i^2(\exp\{\mathbf{X}_i^{*\top}\alpha\})^{-2}},$$

for every i $\in$ (1,...,n), then we have

$$\frac{\partial^2\hat{\mathcal{L}}_1}{\partial\alpha^2} = \frac{1}{n}\bar{\mathbf{X}}^*\bar{\mathbf{X}}^{*\top},$$

where $\bar{\mathbf{X}}^*$ is $p \times n$ matrix. As long as the full column rank of $\bar{\mathbf{X}}^*$ is guaranteed with probability one, we could obtain the positive-definite property of the Hessian matrix which indicates the strong convexity of our loss function concerning α given β. □

Lemma 3.10. *Refer the property of Bernoulli distribution, for integer $k > 0$, we have* $\mathbf{E}(Y^k|\mathbf{X}^*) = (1-P(\mathbf{X}^*, \alpha_0, \beta_0))(-P(\mathbf{X}^*, \alpha_0, \beta_0))^k + P(\mathbf{X}^*, \alpha_0, \beta_0)(1-(P(\mathbf{X}^*, \alpha_0, \beta_0))^k)$, *where*

$$P(\mathbf{X}^*, \alpha_0, \beta_0) = \frac{(\exp\{\mathbf{X}^{*\top}\alpha\})^2}{1+(\exp\{\mathbf{X}^{*\top}\alpha\})^2}\mathrm{I}_{\{\mathbf{X}^{*\top}\beta\geq 0\}}.$$

The lemma 3.10 was used to illustrate the calculations of the terms like $\mathbf{E}(Y^k|\mathbf{X}^*)$ could be transformed into the function of $\mathbf{X}^*$.

Proof of Lemma 3.2

Proof. Since

$$\frac{\partial\hat{\mathcal{L}}_1^2(\theta)}{\partial\alpha\partial\beta} = \frac{1}{n}\sum_{i=1}^{n}\frac{-2\mathbf{X}_i^*(\exp\{\mathbf{X}_i^{*\top}\alpha\})^2}{h}\Phi'\Big(\frac{\mathbf{X}_i^{*\top}\beta}{h}\Big)\Big(Y_i - \Phi\Big(\frac{\mathbf{X}_i^{*\top}\beta}{h}\Big)\Big).$$

Then, denote $z = \mathbf{X}^{*\top}\beta_0$, $p_1(\cdot)$ refers to the density of $\widetilde{\mathbf{X}^*}$ and $p_2(\cdot)$ refers to the density of z. We derive that

$$\begin{aligned} & \sqrt{h}\mathbf{E}\Big[\frac{\partial\hat{\mathcal{L}}_1^2(\theta)}{\partial\alpha\partial\beta}_{\theta=\theta_0}\Big] \\ = & \sqrt{h}\mathbf{E}\Big[\frac{-2\mathbf{X}^*\mathbf{X}^{*\top}(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2}{h}\Phi'\Big(\frac{\mathbf{X}^{*\top}\beta_0}{h}\Big)\Big(Y - \Phi\Big(\frac{\mathbf{X}^{*\top}\beta_0}{h}\Big)\Big)\Big] \\ = & \frac{-2\sqrt{h}}{b_1}\int \mathbf{X}^*\mathbf{X}^{*\top}(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2\Phi'(z)\Big(\frac{(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2}{1+(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2}\mathrm{I}_{\{\mathbf{X}^{*\top}\beta_0\geq 0\}} - \Phi(z)\Big) \\ & dF(hz|\widetilde{\mathbf{X}^*})dF(\widetilde{\mathbf{X}^*}) \\ = & o(1), \end{aligned}$$

as $h \to 0$. If $\theta_n \xrightarrow{p} \theta_0$ hold, then $\frac{\partial\hat{\mathcal{L}}_1^2(\theta)}{\partial\alpha\partial\beta}|_{\theta=\bar{\theta}} \xrightarrow{p} \mathbf{E}\big[\frac{\partial\hat{\mathcal{L}}_1^2(\theta)}{\partial\alpha\partial\beta}|_{\theta=\theta_0}\big] = o(1)$.

□

Proof of Lemma 3.3

Proof. According to assumption 3.6, denote $z = \mathbf{X}^{*\top}\beta$, it is not hard to derive $\left|\mathrm{I}_{\{z\geq 0\}} - \Phi(\frac{z}{h})\right| = o(h^2)$. For the sake of conciseness, we denote

$$\begin{aligned} C(\alpha,\beta,\mathbf{X}^*): &= \Bigg[\bigg(\frac{(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2}{1+(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2}\mathrm{I}_{\{\mathbf{X}^{*\top}\beta_0\geq 0\}} - \Phi\bigg\{\frac{\mathbf{X}^{*\top}\beta_0}{h}\bigg\}\bigg)^2 \\ &\quad -\bigg(\frac{1}{(1+(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2)^2}\mathrm{I}_{\{\mathbf{X}^{*\top}\beta_0\geq 0\}}\bigg)\Bigg]. \end{aligned}$$

Then,

$$\mathbf{E}[\sqrt{n}\hat{J}_n^{\alpha}(\theta_0)] = \sqrt{n}\mathbf{E}\left[\mathbf{X}^*\frac{(\exp\{\mathbf{X}^*\alpha_0\})^4\bigg(Y-\Phi\bigg\{\frac{\mathbf{X}^{*\top}\beta_0}{h}\bigg\}\bigg)^2 - Y^2}{(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2}\right]$$

$$\begin{aligned} &= \sqrt{n}\mathbf{E}\big[\mathbf{X}^*(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2 C(\alpha,\beta,\mathbf{X}^*)\big] \\ &= \sqrt{n}\mathbf{E}\Bigg[\mathbf{X}^*(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2\Bigg[\bigg(\frac{(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2-1}{1+(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2}\mathrm{I}_{\{\mathbf{X}^{*\top}\beta_0\geq 0\}} - \Phi\bigg\{\frac{\mathbf{X}^{*\top}\beta_0}{h}\bigg\}\bigg) \\ &\quad \times\bigg(\mathrm{I}_{\{\mathbf{X}^{*\top}\beta_0\geq 0\}} - \Phi\bigg\{\frac{\mathbf{X}^{*\top}\beta_0}{h}\bigg\}\bigg)\Bigg]\Bigg] \\ &\quad \leq \sqrt{nh^4}\mathbf{E}\big[\mathbf{X}^*(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2[(\frac{(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2-1}{1+(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2}\mathrm{I}_{\{\mathbf{X}^{*\top}\beta_0\geq 0\}} - \Phi\{\frac{\mathbf{X}^{*\top}\beta_0}{h}\})]] \\ &= o(1), \end{aligned}$$

as $h \to 0$. □

Proof of Lemma 3.4

Proof. According to Lemma 3.10, we could derive

$$\mathbf{E}\left[\frac{\partial^2\hat{\mathcal{L}}_1}{\partial\alpha^2}_{\theta=\theta_0}\right] = \mathbf{E}\left[\mathbf{X}^*\mathbf{X}^{*\top}2(\exp\{\mathbf{X}^{*\top}\alpha\})^2\left(Y-\Phi\left\{\frac{\mathbf{X}^{*\top}\beta}{h}\right\}\right)^2 + 2Y^2(\exp\{\mathbf{X}^{*\top}\alpha\})^{-2}\right]$$

$$= \mathbf{E}\left[\mathbf{E}\left[\mathbf{X}^*\mathbf{X}^{*\top}2(\exp\{\mathbf{X}^{*\top}\alpha\})^2\left(Y-\Phi\left\{\frac{\mathbf{X}^{*\top}\beta}{h}\right\}\right)^2 + 2Y^2(\exp\{\mathbf{X}^{*\top}\alpha\})^{-2}\right]\middle|\mathbf{X}^*\right]$$

$$= \mathbf{E}[P_1(\mathbf{X}^*, \alpha_0, \beta_0)]$$

$$= \mathcal{D}_1$$

If $\theta_n \xrightarrow{p} \theta_0$ hold, then $\frac{\partial^2\hat{\mathcal{L}}_1}{\partial\alpha^2}|_{\theta=\theta_n} \xrightarrow{p} \mathbf{E}\big[\frac{\partial^2\hat{\mathcal{L}}_1}{\partial\alpha^2}|_{\theta=\theta_0}\big] = \mathcal{D}_1$.

□

Proof of Lemma 3.5

Proof. Since

$$
\begin{aligned}
& \mathbf{E}\Big[\frac{\partial^2 \hat{\mathcal{L}}_1}{\partial \beta^2}_{\theta=\theta_0}\Big] \\
= \;& -\mathbf{E}\Big[Aa\frac{\mathbf{X}^*\mathbf{X}^{*\top}}{h^2}\Big\{\Big(Y - a\Phi\Big\{\frac{\mathbf{X}^{*\top}\beta_0}{h}\Big\}\Big)\Phi''\Big\{\frac{\mathbf{X}^{*\top}\beta_0}{h}\Big\}\Big\} \\
& +Aa^2\frac{\mathbf{X}^*\mathbf{X}^{*\top}}{h^2}\Big(\Phi'\Big\{\frac{\mathbf{X}^{*\top}\beta_0}{h}\Big\}\Big)^2\Big] \\
= \;& -h^{-2}\mathbf{E}\Big[Aa\mathbf{X}^*\mathbf{X}^{*\top}\Big\{\Big(Y - a\Phi\Big\{\frac{\mathbf{X}^{*\top}\beta_0}{h}\Big\}\Big)\Phi''\Big\{\frac{\mathbf{X}^{*\top}\beta_0}{h}\Big\}\Big\} \\
& +Aa^2\mathbf{X}^*\mathbf{X}^{*\top}\Big(\Phi'\Big\{\frac{\mathbf{X}^{*\top}\beta_0}{h}\Big\}\Big)^2\Big] \\
= \;& -h^{-2}\mathbf{E}\Big[Aa\mathbf{X}^*\mathbf{X}^{*\top}\Big\{\Big(Y - a\Phi\Big\{\frac{\mathbf{X}^{*\top}\beta_0}{h}\Big\}\Big)\Phi''\Big\{\frac{\mathbf{X}^{*\top}\beta_0}{h}\Big\}\Big\}\Big] \\
& +\mathbf{E}\Big[Aa^2\mathbf{X}^*\mathbf{X}^{*\top}\Big(\frac{\Phi'\Big\{\frac{\mathbf{X}^{*\top}\beta_0}{h}\Big\}}{h}\Big)^2\Big] \\
= \;& -h^{-2}\mathbf{E}\Big[Aa\mathbf{X}^*\mathbf{X}^{*\top}\Big\{\Big(Y - a\Phi\Big\{\frac{\mathbf{X}^{*\top}\beta_0}{h}\Big\}\Big)\Phi''\Big\{\frac{\mathbf{X}^{*\top}\beta_0}{h}\Big\}\Big\}\Big] + o(h) \\
= \;& -h^{-2}\mathbf{E}\Big[Aa^2\mathbf{X}^*\mathbf{X}^{*\top}\Big\{\Big(\mathrm{I}_{\{\mathbf{X}^{*\top}\beta_0\geq 0\}} - \Phi\Big\{\frac{\mathbf{X}^{*\top}\beta_0}{h}\Big\}\Big)\Phi''\Big\{\frac{\mathbf{X}^{*\top}\beta_0}{h}\Big\}\Big\}\Big] + o(h) \\
= \;& -(hb_1)^{-1}\int Aa^2\mathbf{X}^*\mathbf{X}^{*\top}\Big\{\Big(\mathrm{I}_{\{hz\geq 0\}} - \Phi(z)\Big)\Phi''(z)\Big\}dF(hz|\widetilde{\mathbf{X}^*})dF(\widetilde{\mathbf{X}^*}) + o(h) \\
= \;& -(b_1)^{-1}\int_{z\geq 0} Aa^2\mathbf{X}^*\mathbf{X}^{*\top}z\Big\{\Big(1 - \Phi(z)\Big)\Phi'''(z) - \Phi'(z)\Phi''(z)\Big\}dF(hz|\widetilde{\mathbf{X}^*})dF(\widetilde{\mathbf{X}^*}) \\
& +(b_1)^{-1}\int_{z<0} Aa^2\mathbf{X}^*\mathbf{X}^{*\top}z\Big\{\Phi(z)\Phi'''(z) + \Phi'(z)\Phi''(z)\Big\}dF(hz|\widetilde{\mathbf{X}^*})dF(\widetilde{\mathbf{X}^*}) + o(h) \\
& = \mathcal{D}_2
\end{aligned}
$$

where $A = \frac{(1+(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2)^2}{(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2}$, $a = \frac{(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2}{1+(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2}$, and

$$\mathcal{D}_2 = -(b_1)^{-1}\int_{z\geq 0} Aa^2\mathbf{X}^*\mathbf{X}^{*\top}z\bigg\{\Big(1-\Phi(z)\Big)\Phi'''(z) - \Phi'(z)\Phi''(z)\bigg\}dF(hz|\widetilde{\mathbf{X}^*})dF(\widetilde{\mathbf{X}^*})$$

$$+(b_1)^{-1}\int_{z<0} Aa^2\mathbf{X}^*\mathbf{X}^{*\top}z\bigg\{\Phi(z)\Phi'''(z) + \Phi'(z)\Phi''(z)\bigg\}dF(hz|\widetilde{\mathbf{X}^*})dF(\widetilde{\mathbf{X}^*})$$

is a constant matrix. If $\theta_n \xrightarrow{p} \theta_0$ hold, then

$$\frac{\partial^2\hat{\mathcal{L}}_1}{\partial\alpha^2}\bigg|_{\theta=\bar{\theta}} \xrightarrow{p} \mathbf{E}\bigg[\frac{\partial^2\hat{\mathcal{L}}_1}{\partial\alpha^2}\bigg|_{\theta=\theta_0}\bigg].$$

□

Proof of Lemma 3.6

Proof.

$$\begin{aligned}
&\mathbf{E}[\sqrt{nh}\hat{J}_n^\beta(\theta_0)] \\
= \quad & -\sqrt{\frac{n}{h}}\mathbf{E}\bigg[\mathbf{X}^*\bigg[(1+(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2)\bigg(Y - \frac{(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2}{1+(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2}\Phi\bigg\{\frac{\mathbf{X}^{*\top}\beta_0}{h}\bigg\}\bigg) \\
& \times\Phi'\bigg\{\frac{\mathbf{X}^*\beta_0}{h}\bigg\}\bigg]\bigg] \\
= \quad & \frac{\sqrt{nh}}{b_1}\int \mathbf{X}^*\bigg[(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2\Big(\mathrm{I}_{\{z\geq 0\}} - \Phi(z)\Big)\Phi'(z)\bigg]dF(hz|\widetilde{\mathbf{X}^*})dF(\widetilde{\mathbf{X}^*}) \\
< \quad & \frac{\sqrt{nh^3}}{b_1}\int \mathbf{X}^*\bigg[(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2\Phi'(z)\bigg]dF(hz|\widetilde{\mathbf{X}^*})dF(\widetilde{\mathbf{X}^*}) \\
= \quad & o(1),
\end{aligned}$$

as $h \to 0$. □

Proof of Lemma 3.7

Proof. According to the Lemma 3.3 and Lemma 3.6, we only need to prove

$$\mathbf{E}[n\sqrt{h}\hat{J}_n^{\alpha}(\theta_0)\hat{J}_n^{\beta}(\theta_0)] \xrightarrow{p} o(1),$$

as long as $nh^2 \to 0$. Since

$$\begin{aligned}
&\mathbf{E}[n\sqrt{h}\hat{J}_n^{\alpha}(\theta_0)\hat{J}_n^{\beta}(\theta_0)] \\
= \;& \frac{n}{\sqrt{h}}\mathbf{E}\Bigg[\mathbf{X}^*\mathbf{X}^{*\top}(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2\Bigg[\Bigg(\frac{(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2-1}{1+(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2}\mathrm{I}_{\{\mathbf{X}^{*\top}\beta_0\geq 0\}} - \Phi\Bigg\{\frac{\mathbf{X}^{*\top}\beta_0}{h}\Bigg\}\Bigg) \\
&\Bigg(\mathrm{I}_{\{\mathbf{X}^{*\top}\beta_0\geq 0\}} - \Phi\Bigg\{\frac{\mathbf{X}^{*\top}\beta_0}{h}\Bigg\}\Bigg)\Bigg] \\
&*\Bigg[(1+(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2)\Bigg(Y - \frac{(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2}{1+(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2}\Phi\Bigg\{\frac{\mathbf{X}^{*\top}\beta_0}{h}\Bigg\}\Bigg)\Phi'\Bigg\{\frac{\mathbf{X}^{*\top}\beta_0}{h}\Bigg\}\Bigg]\Bigg] \\
= \;& \frac{n\sqrt{h}}{b_1}\int \mathbf{X}^*\mathbf{X}^{*\top}(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2\Bigg[\Bigg(\frac{(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2-1}{1+(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2}\mathrm{I}_{\{z\geq 0\}} - \Phi\{z\}\Bigg)\Bigg(\mathrm{I}_{\{z\geq 0\}} - \Phi\{z\}\Bigg)\Bigg] \\
&\Bigg[(1+(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2)\Bigg(Y - \frac{(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2}{1+(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2}\Phi\{z\}\Bigg)\Phi'\{z\}\Bigg]dF(hz|\widetilde{\mathbf{X}^*})dF(\widetilde{\mathbf{X}^*}) \\
= \;& \frac{nh^2\sqrt{h}}{b_1}\int \mathbf{X}^*\mathbf{X}^{*\top}(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2\Bigg[\Bigg(\frac{(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2-1}{1+(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2}\mathrm{I}_{\{z\geq 0\}} - \Phi\{z\}\Bigg)\Bigg] \\
&\Bigg[(1+(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2)\Bigg(Y - \frac{(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2}{1+(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2}\Phi\{z\}\Bigg)\Phi'\{z\}\Bigg]dF(hz|\widetilde{\mathbf{X}^*})dF(\widetilde{\mathbf{X}^*}) \\
&= o(1),
\end{aligned}$$

as $h \to 0$. □

Proof of Lemma 3.8

Proof. Since $\mathbf{E}[\hat{J}_n^{\alpha}(\theta_0)] = 0$, we derive

$$\mathrm{Var}(\hat{\mathrm{J}}_{\mathrm{n}}^{\alpha}(\theta_0)) = \mathbf{E}[(\hat{\mathrm{J}}_{\mathrm{n}}^{\alpha}(\theta_0))^2].$$

Refer the property of Bernoulli distribution, for integer $k > 0$, we have $\mathbf{E}(Y^k|\mathbf{X}^*) = (1-f(\mathbf{X}^*,\alpha_0,\beta_0)\mathrm{I}_{\{\mathbf{X}^{*\top}\beta_0\geq 0\}})(-f(\mathbf{X}^*,\alpha_0,\beta_0)\mathrm{I}_{\{\mathbf{X}^{*\top}\beta_0\geq 0\}})^k+(f(\mathbf{X}^*,\alpha_0,\beta_0)\mathrm{I}_{\{\mathbf{X}^{*\top}\beta_0\geq 0\}})((1-(f(\mathbf{X}^*,\alpha_0,\beta_0))^k)\mathrm{I}_{\{\mathbf{X}^{*\top}\beta_0\geq 0\}})$, where $f(\mathbf{X}^*,\alpha_0,\beta_0) = \frac{(\exp\{\mathbf{X}^{*\top}\alpha\})^2}{1+(\exp\{\mathbf{X}^{*\top}\alpha\})^2}\mathrm{I}_{\{\mathbf{X}^{*\top}\beta\geq 0\}}$. Then,

$$
\begin{aligned}
\mathbf{E}\bigg[\bigg(\frac{\partial\hat{\mathcal{L}}_1}{\partial\alpha}\bigg)^2\bigg] &= \frac{1}{n}\mathbf{E}\bigg[\mathbf{X}^*\mathbf{X}^{*\top}\bigg[\frac{(\exp\{\mathbf{X}^{*\top}\alpha_0\})^4\big(Y-\Phi\big\{\frac{\mathbf{X}^{*\top}\beta_0}{h}\big\}\big)^2-Y^2}{(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2}\bigg]^2\bigg] \\
&= \frac{1}{n}\mathbf{E}\bigg[\mathbf{X}^*\mathbf{X}^{*\top}\bigg[\frac{(\exp\{\mathbf{X}^{*\top}\alpha_0\})^4\big(Y-\Phi\big\{\frac{\mathbf{X}^{*\top}\beta_0}{h}\big\}\big)^2-Y^2}{(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2}\bigg]^2\bigg] \\
&= \frac{1}{n}\mathbf{E}\bigg[\mathbf{X}^*\mathbf{X}^{*\top}[H(\alpha_0,\beta_0,\mathbf{X}^*)]\bigg] \\
&= \frac{1}{n}\mathcal{V},
\end{aligned}
$$

Where $H(\alpha_0,\beta_0,\mathbf{X}^*) = \mathbf{E}\bigg[\bigg[\frac{(\exp\{\mathbf{X}^{*\top}\alpha_0\})^4\big(Y-\Phi\big\{\frac{\mathbf{X}^{*\top}\beta_0}{h}\big\}\big)^2-Y^2}{(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2}\bigg]^2\bigg|\mathbf{X}^*\bigg]$ is a bounded function. □

Proof of Lemma 3.9

Proof. Since $\mathbf{E}[\hat{J}_n^\beta(\theta_0)] = 0$, we derive

$$
\mathrm{Var}(\hat{\mathrm{J}}_\mathrm{n}^\beta(\theta_0)) = \mathbf{E}[(\hat{\mathrm{J}}_\mathrm{n}^\beta(\theta_0))^2].
$$

According to Lemma 3.10, we could derive

$$\mathbf{E}\bigg[\bigg(\frac{\partial\hat{\mathcal{L}}_1}{\partial\beta}\bigg)^2\bigg] = \frac{1}{n}\mathbf{E}\bigg[\bigg(Aa\frac{\mathbf{X}^*}{h}\bigg(Y - a\Phi\bigg\{\frac{\mathbf{X}^{*\top}\beta_0}{h}\bigg\}\bigg)\Phi'\bigg\{\frac{\mathbf{X}^{*\top}\beta_0}{h}\bigg\}\bigg)^2\bigg]$$

$$\begin{aligned}
&= \frac{1}{n}\mathbf{E}\bigg[\mathbf{X}^*\mathbf{X}^{*\top}\frac{A^2a^2}{h^2}\bigg(a\mathrm{I}_{\{\mathbf{X}^{*\top}\beta_0\geq 0\}} + a^2\bigg(\mathrm{I}_{\{\mathbf{X}^{*\top}\beta_0\geq 0\}} - \Phi\bigg\{\frac{\mathbf{X}^{*\top}\beta_0}{h}\bigg\}\bigg)^2\bigg) \\
&\quad \times\bigg(\Phi'\bigg\{\frac{\mathbf{X}^{*\top}\beta_0}{h}\bigg\}\bigg)^2\bigg] \\
&= \frac{1}{n}\mathbf{E}\bigg[\mathbf{X}^*\mathbf{X}^{*\top}\frac{A^2a^3}{h^2}\bigg(\mathrm{I}_{\{\mathbf{X}^{*\top}\beta_0\geq 0\}}\bigg)\bigg(\Phi'\bigg\{\frac{\mathbf{X}^{*\top}\beta_0}{h}\bigg\}\bigg)^2\bigg] + o(h) \\
&= \frac{(nh)^{-1}}{b_1}\int \mathbf{X}^*\mathbf{X}^{*\top}A^2a^3\mathrm{I}_{\{hz\geq 0\}}\Phi'(z)^2dF(hz|\widetilde{\mathbf{X}^*})dF(\widetilde{\mathbf{X}^*}) + o(h) \\
&= (nh)^{-1}\mathcal{W} + o(h),
\end{aligned}$$

where $A = \frac{(1+(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2)^2}{(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2}$, $a = \frac{(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2}{1+(\exp\{\mathbf{X}^{*\top}\alpha_0\})^2}$,, and

$$\mathcal{W} = b_1^{-1}\int \mathbf{X}^*\mathbf{X}^{*\top}A^2a^3\mathrm{I}_{\{hz\geq 0\}}\Phi'(z)^2dF(hz|\widetilde{\mathbf{X}^*})dF(\widetilde{\mathbf{X}^*}).$$

□

Chapter 4

Conclusion and Further Discussion

Explaining zero-inflated proportion data has long posed a statistical challenge for researchers. Despite many traditional methods, they often fall short of adequately explaining the underlying sources of zero inflation.

Our proposed semiparametric model effectively addresses this issue by explicitly accounting for the distinct sources of zero observations, as illustrated in the introduction. Specifically, it differentiates between zeros arising from unsuitability (biotic and abiotic factors) and those due to random absence or detection errors. This model provides a clearer understanding of the incidence of absence, as well as the explanatory contribution of unsuitability versus random chance. The application to real-world data demonstrates the predictive power of our model, highlighting its ability to accurately capture the underlying data-generating processes.

Furthermore, the consistency of both the regression and classification components of our model, coupled with the global convexity of the loss function, ensures the model's stability and robustness. This ensures reliable and interpretable parameter estimates, as well as accurate predictions.

In Chapter 3, we enhanced our model by replacing the unspecified classification part with a parametric model. This would allow us to derive more theoretical inferences and potentially improve the efficiency of parameter estimation.

Overall, our proposed semiparametric model offers a comprehensive and flexible approach to analyzing zero-inflated proportion data, addressing the limitations of traditional methods and providing a deeper understanding of the underlying mechanisms driving the excess zeros. Its robustness, interpretability, and potential for further enhancements make it a valuable contribution to the field of statistical modeling.

Ongoing Research and Further Extension

The extension of our models to high-dimensional data settings is highly significant, as numerous practical applications involve datasets where the number of parameters vastly exceeds the number of samples, and the covariates are inherently sparse. Such scenarios are prevalent in fields like microbial data analysis and industrial production data processing. To address these high-dimensional challenges, both our semiparametric model and the parametric model introduced in Chapter 3 can incorporate penalty terms for further analysis and regularization.

The semiparametric model could be adapted to high-dimensional extensions due to its partitioned structure. By considering the two components separately, the first part reduces to an optimization problem involving a globally convex function combined with a penalty term. This formulation has been extensively studied in the literature, as detailed in the technical tools section, and the properties of the parameters after incorporating the lasso penalty have been rigorously analyzed. Consequently, these established results can be directly applied to the analysis of our first part. For the classification component, since the first part has already identified the non-zero parameters, the second part effectively regresses to a low-dimensional data analysis problem.

However, the parametric model described in Chapter3 presents a more formidable challenge in the high-dimensional setting. This model simultaneously involves two parameters, with the β parameter embedded within the indicator function. Extending this entire parametric model to high-dimensional data while obtaining meaningful theoretical guarantees is a highly intricate task that warrants further investigation.

An intriguing extension to our model involves introducing a neural network architecture for fitting the regression component. Analogously, once we obtain accurate estimates of the parameters for the regressor $f(\cdot)$, we could leverage this knowledge to calculate the sign of $D(\cdot)$, as demonstrated in Chapter 2. An interesting avenue for future exploration lies in evaluating the practical utility of our proposed technical approach, which separates the regression and classification tasks, and subsequently utilizes the regression results to inform the classification process. This could be achieved by conducting comparative analyses between the prediction errors obtained through our sequential procedure and those obtained by fitting the entire dataset simultaneously. Such an investigation would shed light on the potential advantages and limitations of our decoupled methodology, paving the way for further refinements and optimizations.

Bibliography

Agarwal, A. and Duchi, J. C. (2011), “Distributed delayed stochastic optimization,” *Advances in neural information processing systems*, 24.

Amemiya, T. (1973), “Regression analysis when the dependent variable is truncated normal,” *Econometrica: Journal of the Econometric Society*, pp. 997–1016.

Amemiya, T. (1985), *Advanced econometrics*, Harvard university press.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006), “Convexity, classification, and risk bounds,” *Journal of the American Statistical Association*, 101, 138–156.

Beck, A. and Teboulle, M. (2009), “A fast iterative shrinkage-thresholding algorithm for linear inverse problems,” *SIAM journal on imaging sciences*, 2, 183–202.

Bühlmann, P. and Van De Geer, S. (2011), *Statistics for high-dimensional data: methods, theory and applications*, Springer Science & Business Media.

Chib, S. (1992), “Bayes inference in the Tobit censored regression model,” *Journal of Econometrics*, 51, 79–99.

Clayton, D. H. and Cotgreave, P. (1994), “Comparative analysis of time spent grooming by birds in relation to parasite load,” *Behaviour*, 131, 171–187.

Crawley, M. J. (2012), *The R book*, John Wiley & Sons.

Defries, R. S., Hansen, M. C., Townshend, J. R., Janetos, A., and Loveland, T. R. (2000), “A new global 1-km dataset of percentage tree cover derived from remote sensing,” *Global Change Biology*, 6, 247–254.

Fan, J. (2018), *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, Routledge.

Fan, J. and Li, R. (2001), “Variable selection via nonconcave penalized likelihood and its oracle properties,” *Journal of the American statistical Association*, 96, 1348–1360.

Friedman, J., Hastie, T., and Tibshirani, R. (2008), “Sparse inverse covariance estimation with the graphical lasso,” *Biostatistics*, 9, 432–441.

Ge, R., Huang, F., Jin, C., and Yuan, Y. (2015), "Escaping from saddle points—online stochastic gradient for tensor decomposition," in *Conference on learning theory*, pp. 797–842, PMLR.

Giné, E. and Guillou, A. (2002), "Rates of strong uniform consistency for multivariate kernel density estimators," in *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, vol. 38, pp. 907–921, Elsevier.

Golan, A., Judge, G., and Perloff, J. (1997), "Estimation and inference with censored and ordered multinomial response data," *Journal of Econometrics*, 79, 23–51.

Härdle, W. (1990), *Applied nonparametric regression*, no. 19, Cambridge university press.

Hardle, W., Janssen, P., and Serfling, R. (1988), "Strong uniform consistency rates for estimators of conditional functionals," *The Annals of Statistics*, pp. 1428–1449.

Heckman, J. J. (1979), "Sample selection bias as a specification error," *Econometrica: Journal of the econometric society*, pp. 153–161.

Hoeffding, W. (1994), "Probability inequalities for sums of bounded random variables," *The collected works of Wassily Hoeffding*, pp. 409–426.

Horowitz, J. L. (1992), "A smoothed maximum score estimator for the binary response model," *Econometrica: journal of the Econometric Society*, pp. 505–531.

Jacobson, T. and Zou, H. (2022), "High-dimensional Censored Regression via the Penalized Tobit Likelihood," *arXiv preprint arXiv:2203.02601*.

Kieschnick, R. and McCullough, B. D. (2003), "Regression analysis of variates observed on (0, 1): percentages, proportions and fractions," *Statistical modelling*, 3, 193–213.

Liu, F. and Eugenio, E. C. (2018), "A review and comparison of Bayesian and likelihood-based inferences in beta regression and zero-or-one-inflated beta regression," *Statistical methods in medical research*, 27, 1024–1044.

Liu, L., Shih, Y.-C. T., Strawderman, R. L., Zhang, D., Johnson, B. A., and Chai, H. (2019), "Statistical analysis of zero-inflated nonnegative continuous data: a review," *Statistical Science*, 34, 253–279.

Manning, W. G., Morris, C. N., Newhouse, J. P., Orr, L. L., Duan, N., Keeler, E. B., Leibowitz, A., Marquis, K. H., Marquis, M. S., and Phelps, C. E. (1981), "A two-part model of the demand for medical care: preliminary results from the health insurance study," *Health, economics, and health economics*, 137, 103–23.

Manski, C. F. (1985), "Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator," *Journal of econometrics*, 27, 313–333.

Massart, P. and Élodie Nédélec (2006), "Risk bounds for statistical learning," *The Annals of Statistics*, 34, 2326 – 2366.

Moulton, L. H. and Halsey, N. A. (1995), "A mixture model with detection limits for regression analyses of antibody response to vaccine," *Biometrics*, pp. 1570–1578.

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

Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012), "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers," .

Newey, W. K. (1994), "Kernel estimation of partial means and a general variance estimator," *Econometric Theory*, 10, 1–21.

Nussbaum, M. (1996), "Asymptotic equivalence of density estimation and Gaussian white noise," *The Annals of Statistics*, 24, 2399–2430.

Ocaña-Riola, R., Pérez-Romero, C., Ortega-Díaz, M. I., and Martín-Martín, J. J. (2021), "Multilevel Zero-One Inflated Beta Regression Model for the Analysis of the Relationship between Exogenous Health Variables and Technical Efficiency in the Spanish National Health System Hospitals," *International Journal of Environmental Research and Public Health*, 18, 10166.

Olsen, R. J. (1978), "Note on the uniqueness of the maximum likelihood estimator for the Tobit model," *Econometrica: Journal of the Econometric Society*, pp. 1211–1215.

Ospina, R. and Ferrari, S. L. (2012), "A general class of zero-or-one inflated beta regression models," *Computational Statistics & Data Analysis*, 56, 1609–1623.

Peng, X., Li, G., and Liu, Z. (2016), "Zero-inflated beta regression for differential abundance analysis with metagenomics data," *Journal of Computational Biology*, 23, 102–110.

Poorter, H., Niklas, K. J., Reich, P. B., Oleksyn, J., Poot, P., and Mommer, L. (2012), "Biomass allocation to leaves, stems and roots: meta-analyses of interspecific variation and environmental control," *New Phytologist*, 193, 30–50.

Queiroz, F. F. and Lemonte, A. J. (2021), "A broad class of zero-or-one inflated regression models for rates and proportions," *Canadian Journal of Statistics*, 49, 566–590.

Quinn, G. P. and Keough, M. J. (2002), *Experimental design and data analysis for biologists*, Cambridge university press.

Rao, R. R. (1962), “Relations between weak and uniform convergence of measures with applications,” *The Annals of Mathematical Statistics*, pp. 659–680.

Rosenblatt, M. (1956), “Remarks on Some Nonparametric Estimates of a Density Function,” *The Annals of Mathematical Statistics*, 27, 832 – 837.

Santos, B. and Bolfarine, H. (2015a), “Bayesian analysis for zero-or-one inflated proportion data using quantile regression,” *Journal of Statistical Computation and Simulation*, 85, 3579 – 3593.

Santos, B. and Bolfarine, H. (2015b), “Bayesian analysis for zero-or-one inflated proportion data using quantile regression,” *Journal of Statistical Computation and Simulation*, 85, 3579–3593.

Silverman, B. W. (2018), *Density estimation for statistics and data analysis*, Routledge.

Sun, Q., Zhou, W.-X., and Fan, J. (2020), “Adaptive huber regression,” *Journal of the American Statistical Association*, 115, 254–265.

Swearingen, C., Castro, M., and Bursac, Z. (2012), “Inflated Beta Regression: Zero, One, and Everything in Between,” .

Tang, B., Frye, H. A., Gelfand, A. E., and Silander, J. A. (2023), “Zero-inflated Beta distribution regression modeling,” *Journal of Agricultural, Biological and Environmental Statistics*, 28, 117–137.

Tibshirani, R. (1996), “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58, 267–288.

Tobin, J. (1958), “Estimation of relationships for limited dependent variables,” *Econometrica: journal of the Econometric Society*, pp. 24–36.

Warton, D. I. and Hui, F. K. (2011), “The arcsine is asinine: the analysis of proportions in ecology,” *Ecology*, 92, 3–10.

Watson, G. S. (1964), “Smooth regression analysis,” *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 359–372.

Wieczorek, J. and Hawala, S. (2012), “A Bayesian Zero-One Inflated Beta Model for Estimating Poverty,” .

Zhang, C.-H. (2010), “Nearly unbiased variable selection under minimax concave penalty,” .