Essays in Identification and Estimation of Entry Games with Symmetry of Unobservables

by

Yu Zhou

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Economics)
in The University of Michigan
2014

Doctoral Committee:

Professor Daniel A. Ackerberg, Chair
Associate Professor Jeremy T. Fox
Assistant Professor Yesim Orhun
Professor Jeffrey A. Smith

For my parents and my husband

ACKNOWLEDGEMENTS

I have been blessed to be surrounded by many extraordinary people during the writing of my dissertation. This dissertation has benefited immeasurably from contributions by a number of people. I am very grateful to all those who have made this dissertation possible. I value them not only as my professors and colleagues, but also as friends who have made my experience at Michigan one that I will cherish forever.

First and foremost, I am deeply indebted to the members of my dissertation committee. My committee chair, Daniel Ackerberg, not only has been a great source of professional advice, but also has inspired and encouraged me to set and achieve high standards for my research. He has guided me how to take complicated material and produce a clear and understandable analysis, while also providing solid reasoning for each step. Prof. Ackerberg has always kept me focused when I was about to digress or go in too many directions at once. I have benefited greatly from Prof. Ackerberg's advice and will continue to throughout my career. Jeremy Fox has been exceptionally generous with his time, reading my papers carefully and giving me invaluable suggestions and comments. He has taught me to be rigorous on every detail or analysis, as well as on every word, whether written or spoken. Prof. Fox's questions and critiques have reshaped my perspective on research and have helped me hone my thinking. Yesim Orhun has always given me invaluable suggestions to make the technical details more accessible to the empirical fellows. Her insightful comments have largely shaped my job talk and presentation style. Jeffrey Smith is one of the best teachers that I have ever had. He can always find a way to express

technical details in the most intuitive way. Prof. Smith's insightful comments are thought-provoking and help me sharpen my ideas.

Apart from my committee, I would like to acknowledge many faculty members who have fostered my academic achievement. I am especially thankful to Matias Cattaneo and Yoonseok Lee for their extremely valuable contributions to my econometric knowledge. I also thank Jing Cai, Ying Fan and Natalia Lazzati for their invaluable suggestions during my job search. Thank you, also, to my writing and speaking advisors, Deborah Des Jardins, Christine A. Feak and Pamela Bogart, for their continued support in helping me improve my writing and speaking skills.

I am also exceptionally lucky to be surrounded by a group of friends and colleagues who have shared their wisdom, time and friendship with me during my doctoral studies at Michigan: Tanya Byker, Reid Dorsey-Palmateer, Italo Gutierrez, Dave Knapp, Ben Niu, Eric Lewis, William Lincoln and Chris Sullivan.

None of this would have been possible without my family. I am very blessed to have parents who, in their wisdom, have supported me in getting the best education, starting with my preschool years. I thank my parents for their endless love, support and encouragement all these years. I also would like to thank my husband, Guodong Chen, for his tremendous love, care and support during the peaks and valleys of the dissertation process.

Thank you, also, to Nancy Herlocker for helping me become familiar with the high-performance computation I needed to perform for my dissertation. I would also like to thank Mary Mangum and Lauren Pulay for their support for the research fund application. Last but not least, I want to thank Mary Braun and Vinnie Vinjimoor for all of their support in the program. Finally, I gratefully acknowledge financial support from the Rackham Graduate School and the Michigan Institute for Teaching and Research in Economics.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Essays in Identification and Estimation of Entry Games with Symmetry of Unobservables

by

Yu Zhou

Chair: Daniel Ackerberg

The first chapter studies semiparametric point identification and estimation of complete information entry games and proposes a root-n consistent estimator. The proposed method focuses on a two-player entry game using an example of discount retailers, where the potential profit of one retailer depends on the actions of its competitor, and the unobserved heterogeneities of the two retailers can be correlated. These two features lead to two challenges in identification and estimation: multiple equilibria and endogeneity. To address these two challenges, this paper provides a new identification and estimation strategy under a symmetry condition on unobservables. This new identification procedure requires neither an equilibrium selection rule of multiple equilibria nor parametric distributional assumptions on unobservables to solve the endogeneity problem. It also requires a weaker support condition than that in the existing literature. Following the identification argument, this paper proposes a semiparametric two-step estimation procedure using plug-in kernel estimators. Given the symmetry assumption, this paper shows that the proposed estimator is root-n consistent, unlike existing estimators for this model.

The second chapter considers a Monte Carlo simulation study for complete information entry games. The purpose of this study is to provide evidence consistent with the root-n consistency of the semiparametric estimator proposed by Zhou (2014a) and to compare this proposed estimator with an existing parametric estimator. The results are consistent with the proposed estimator being root-n consistent, as predicted by Zhou (2014a). In addition, the parametric estimator outperforms the semiparametric estimator with lower biases and variances when the model is correctly specified. When the model is incorrectly specified, the parametric estimator is inconsistent, while the semiparametric estimator is consistent.

The third chapter applies existing parametric estimation methods and a new semiparametric estimation method by Zhou (2014a) to entry games of discount retailers. Using data on Kmart's and Walmart's entry decisions in 1997 across counties in the U.S., this paper finds that, with a caveat for the possible misspecification of the latent function, semiparametric and parametric estimators give similar estimates. This result informally suggests that normality seems to be a reasonable approximation for the distribution of unobservables in the discount retailing industry.

CHAPTER I

A Theoretical Perspective

1.1 Introduction

This paper studies identification and estimation of static entry games of complete information. Entry games have been widely applied to a variety of topics, such as airline competition, technology adoption and location choices of discount retailers.[1] The previous literature primarily assumes a parametric distributional assumption on unobservables. Very recent studies (Berry and Tamer (2006), Khan and Nekipelov (2012), Fox and Lazzati (2013), Kline (2012), and Dunker, Hoderlein, and Kaido (2013)) relax the distributional assumption on unobservables and focus on a semiparametric approach. However, relaxation of the distributional assumption poses a challenge to identification and estimation. These recent semiparametric methods use identification strategies that rely on having a set of observables with a small probability mass. A consequence of this is that estimators derived from these identification strategies have a rate of convergence that is slower than $n^{-1/2}$, provided that the observables have the finite variance.[2] This paper introduces a symmetry condition on

[1] Airlines competition (Berry (1992) and Ciliberto and Tamer (2009)); technology adoption (Manuszak and Cohen (2004), Ackerberg and Gowrisankaran (2006) and Ryan and Tucker (2012)); location choices of discount retailers (Jia (2008) and Ellickson, Houghton, and Timmins (2013)).

[2] Note that $\sqrt{n}$-consistency corresponds to the $n^{-1/2}$ rate of convergence. Loosely speaking, the rate of convergence is a measure of how faster the standard error will decline to zero when we increase the sample size.

unobservables and provides a new identification and estimation strategy from which we can derive an estimator that converges at rate $n^{-1/2}$. For illustrational purpose, we consider the entry decisions faced by Kmart and Walmart throughout this paper.

Consider a simple two-player static entry game of complete information, with markets $i = 1, ..., n$; two discount retailers $p = 1(\text{Kmart}), 2(\text{Walmart})$; and the payoffs given by

$$\begin{aligned} Y_{1i}^* &= Z_{1i}\beta + \Delta_1 Y_{2i} + \varepsilon_{1i}; \\ Y_{2i}^* &= Z_{2i}\beta + \Delta_2 Y_{1i} + \varepsilon_{2i}; \\ Y_{pi} &= \begin{cases} 1, & \text{if } Y_{pi}^* \geq 0 \\ 0, & \text{otherwise} \end{cases} \quad \text{for } p = 1, 2; \end{aligned} \tag{1.1}$$

where (Y_{1i}^*, Y_{2i}^*) is a vector of latent profitability in market i for the discount retailers; (Z_{1i}, Z_{2i}) is a vector of firm-market specific observed characteristics; and $\boldsymbol{\varepsilon}_i = (\varepsilon_{1i}, \varepsilon_{2i})$ is a vector of unobserved characteristics with an unknown distribution. We allow for the correlation between ε_{1i} and ε_{2i}. In this type of entry game, the discount retailers will enter a particular market ($Y_{pi} = 1$) only if it is profitable to do so ($Y_{pi}^* \geq 0$). Under the assumption of complete information, each discount retailer knows $(Y_{ip}, Z_{ip}, \varepsilon_{ip})$ for both firms ($p = 1, 2$), while the econometrician knows only (Y_{ip}, Z_{ip}) for $p = 1, 2$. Our objective is to recover the model parameters using data on discount retailers' entry decisions and the observed characteristics.

Model (1.1) captures two channels of interdependence between the entry decisions of the two discount retailers—channels that are important to consider when applying this model to empirical work. First, the model explicitly permits strategic interaction between retailers. This interaction is captured by the coefficients (Δ_1, Δ_2), which can be interpreted as a decline in the potential profitability of a discount retailer due to the entry of its competitor. These competition effects are the key parameters of interest in this paper and of great importance for understanding market structure,

market regulation, and antitrust analysis. Second, the model also allows for correlated unobserved heterogeneity. For example, unobserved shocks in the market common to both firms could exist, which leads to the correlation between the unobservables. Failure to account for this correlated unobserved heterogeneity may lead to a false inference about the competition effects. However, allowing for strategic interaction creates a methodological difficulty: multiple equilibria. This means that for some realizations of the unobservables, the entry game model predicts two different entry outcomes. At the same, allowing for the correlated unobserved heterogeneity creates another difficulty: endogeneity. Therefore, the two channels of interdependence create two difficulties in identification and estimation.

To address these two difficulties, this paper constructs a semiparametric identification strategy under an additional shape restriction on the distribution of the unobservables: a "radial symmetry" condition. With this additional assumption, we are able to identify the parameters of interest using the choice probabilities of the unique entry outcomes, while relaxing the support condition on observables used by the existing literature.

Estimation can be directly constructed from the identifying restriction, where the objective function takes a form similar to the nonlinear least square estimation. It has been shown that the leading term of the estimator derived from this objective function takes the form of a U-statistic, as discussed in Powell, Stock, and Stoker (1989), Newey (1994), Imbens and Ridder (2009). These studies show that if the U-statistic can be written as an average of the plug-in kernel component, the rate of convergence of such an estimator is determined by the components that are averaged over. We note that the U-statistic in our estimator can be written as the average of a kernel regression estimator of the choice probability, and we show that since our estimator averages over all components of the covariates, it results in $\sqrt{n}$-consistency of the model parameters. Beyond the asymptotic properties, we also derive a higher-

order mean squared error approximation for the estimator that is used to compute an optimal bandwidth choice.

1.1.1 Literature Review

This paper is related to two strands of research: the entry game literature in industrial organization and the literature on semiparametric identification and estimation of discrete choice models in econometrics.[3] Below, we present a brief overview of the relevant literature in these two fields in order to provide a more detailed comparison of their respective approaches to the method proposed in the present paper.

Entry Games. This paper is related to the broader literature on entry games of complete information.[4] Early works, including seminal papers by Bresnahan and Reiss (1990, 1991a,b) and Berry (1992), used simulation-based estimators to recover the model parameters. These estimators rely on parametric assumptions to model endogeneity directly. In addressing the multiple equilibria problem, Bresnahan and Reiss (1990, 1991a,b) and Berry (1992) focused on the fact that the number of firms in markets is unique despite the existence of the multiple Nash equilibria.

While relying on parametric assumptions, more recent work attempts to explicitly examine multiple equilibria, using the equilibrium selection mechanism to gain identifying power of model parameters. This is accomplished using three possible approaches. The first approach is to specify a particular equilibrium selection mechanism to recover the model parameters, as used in Ackerberg and Gowrisankaran (2006). The second approach is to consider two extreme equilibrium selection mechanisms to place a bound on the model parameters. This bound gives the largest

[3]To some extent, this paper is also related to peer effects literature in labor economics, including: Manski (1993), Moffitt et al. (2001), Carrell, Sacerdote, and West (2011), Card and Giuliano (2013), and Huang (2013) and others. In addition, Bjorn and Vuong (1984) and Soetevent and Kooreman (2007) study labor force participation. Brock and Durlauf (2001) discuss social interaction.

[4]In parallel, the literature on entry games with incomplete information includes Sweeting (2006), Aradillas-Lopez (2012), De Paula and Tang (2012), Wan and Xu (2012), and Lewbel and Tang (2013) and others.

possible identified parameter set that contains the parameter values for any equilibrium selection rule (e.g., Ciliberto and Tamer (2009), Chernozhukov, Hong, and Tamer (2007), Romano and Shaikh (2010), Pakes, Porter, Ho, and Ishii (2011), and Andrews and Barwick (2012)). Introduced more recently, a third way involves explicitly identifying and estimating the equilibrium selection mechanism to address the multiple equilibria (Bajari, Hong, Krainer, and Nekipelov (2010) and Bajari, Hong, and Ryan (2010)). Each of these three methods has its own merits in terms of computation and estimation, and this is still an ongoing area of research.

Another stream of literature for entry games has relaxed parametric assumptions on the distribution of unobservables, that is, they adopt semiparametric identification and estimation strategies. The first to use this approach was Tamer (2003), whose results were based on an identification at infinity argument.[5] More recently, Khan and Nekipelov (2010, 2012), and Fox and Lazzati (2013) have suggested tracing the distribution of unobserved characteristics and recovering the model parameters from the identified error distribution. Kline (2012) proposes combining identification at infinity and maximum score estimation to recover the model parameters.

While these semiparametric approaches have made a significant contribution to the literature, Khan and Nekipelov (2012) show that an identification strategy built on the conditions used in these studies cannot lead to an estimator with $n^{-1/2}$ rate of convergence, which is a property that the semiparametric literature often attempts to attain. The key reason for this impossibility result is that the identification of model primitives built on these conditions relies on either extreme values of observables in identification at infinity, or the identification relies on a set of observables with possible small probability mass at the tails, referred to as "thin set identification"

[5]Early work for the simultaneous discrete choice models for other context with an identification at infinity argument is Heckman (1978).

(Khan and Tamer (2010)).[6,7]

To reconcile the problem faced in the recent entry game studies, we develop a new identification and estimation strategy based on a shape restriction, on the distribution of unobserved characteristics. This shape restriction is referred to as radial symmetry (sometimes referred to as central symmetry). Radial symmetry permits a large class of distributions, including but not limited to those commonly used in parametric approaches, such as the bivariate normal distribution, the bivariate Laplace distribution, the bivariate symmetric logistic distribution and more general elliptically contoured distribution. Importantly, our additional symmetry assumption allows us to relax the support condition of observables used in the existing literature. Intuitively, this assumption allows us to identify the model parameters by identifying the point of symmetry. The symmetric point can be identified using data "nearby" the symmetry point. More specifically, for a *given* set of parameters and error distribution, we need only bounded support for the excluded variables to achieve identification. Of course, since this bounded support depends on the parameters and error distribution, which are unknown, it would be inappropriate to call this a bounded support condition. Specifically, one could find a particular parameter vector and the distribution of unobservables such that any portion of the real line would be needed as part of the support. One might describe our necessary support condition as "bounded conditional on parameters". Using this additional assumption and identification strategy allows us to construct a new estimator that converges at the rate $n^{-1/2}$, unlike the existing literature. We show this by using a U-statistic analysis, similar to Powell, Stock, and Stoker (1989), Newey (1994), and Imbens and Ridder (2009). This result

[6]The method used in Kline (2012) is more closely related to the discussion in Khan and Tamer (2007) on the the maximum score estimation.

[7]The broader literature, initiated by Chamberlain (1986), Heckman (1990), Andrews and Schafgans (1998) and, more recently,Khan and Nekipelov (2012), has concluded that point identification based on identification at infinity or thin set identification will lead to a estimator with a slower rate of convergence than $n^{-1/2}$, provided that the observables have the finite variance. Such results can be found in single-agent (Chamberlain (1986)) and single-agent with two decisions (Heckman (1990), Andrews and Schafgans (1998)) as well as two-agent models (Khan and Nekipelov (2012)).

in our two-agent entry game context is consistent with recent findings by Jochmans (2011) (in a triangular context) and Chen, Khan, and Tang (2013) (for a single-agent model without endogeneity but with heteroskedasticity), who also used symmetry to obtain $n^{-1/2}$ convergence results.

Semiparametric Identification and Estimation in Discrete Choice Models (or more general Limited Dependent Variable Models). This paper is also related to the broader literature on semiparametric identification and estimation of discrete choice models. Econometrically, a discrete game generalizes a standard single-agent discrete choice model by allowing for the agents' decisions to be interrelated. There are five popular approaches to the semiparametric identification and estimation of single-agent models with discrete choice or limited dependent variables: Maximum Score Estimation (Manski (1985)), Rank Correlation Estimation (Han (1987)), Pairwise-difference Estimation (Honoré (1992) and Ahn and Powell (1993)), Single-index Model Estimation (Ichimura (1993)) and Special Regressor (Lewbel (1998, 2000)). These five estimation approaches should be viewed as complements rather than substitutes as these methods adopt different assumptions. In the discrete game context, Fox and Lazzati (2013) follow the special regressor approach, while Kline (2012) follows the maximum score approach. Different from these studies, the present paper combines the rank correlation and pairwise-difference approaches. To outline our approach, we will give a detailed review of only the rank correlation estimation and pairwise-difference approaches.

Rank correlation estimation began with the Maximum Rank Correlation approach proposed by Han (1987), which follows the rank correlation statistic of Kendall (1938). The idea behind this approach is that the rank ordering of the deterministic latent payoff component matches the rank ordering of the choice probabilities only when one correctly specifies the model parameters up to a scale. Following this idea, Cavanagh and Sherman (1998) generalize maximum rank correlation estimation and propose

a new class of rank estimator called monotonic rank (MD) estimation. Since their work, rank correlation estimation has been widely used in the literature, including Chen (1999a,b) (in a single agent model), Abrevaya (2000) (in a generalized fixed effect regression model), and Abrevaya, Hausman, and Khan (2010) (in a triangular simultaneous discrete choice model).

Pairwise-difference estimation is similar to differencing panel models with fixed effects, where variations within "pairwise comparisons" or "matched pairs" can be used to construct an estimator. The approach typically follows a two-step estimation procedure, first eliminating the nuisance components by differencing pairwise observations with approximately equal nuisance components, and second by recovering the other model parameters. The approach has been applied to a variety of models: truncated and censored models (Honoré (1992), Ahn and Powell (1993), and Honoré and Powell (1994)), panel models (Kyriazidou (1997), Abrevaya (1999), Hu (2002), and Honoré and Hu (2004)), as well as the sample selection model with heteroskadasticity (Chen and Khan (2003)). Recently, Honoré and Powell (1997) apply the idea to the general nonlinear model; Aradillas-Lopez, Honoré, and Powell (2007) extend a semilinear model to allow general nonparametric components depending on the conditional expectation; Hong and Shum (2010) use the pairwise-difference idea to estimate dynamic optimization problems; and Aradillas-Lopez (2012) provides a pairwise-difference estimation procedure for incomplete information games.

Overlapping with the rank correlation estimation and pairwise-difference literature, another stream of literature explores the identification power of symmetry conditions. Typically, the symmetry condition creates restrictions between pairs of observations, which can provide an additional source of identification power. Powell (1986), Honoré, Kyriazidou, and Udry (1997), and Hu (2002) and others use symmetry in the context of censored and truncated models. Lee (1996) uses symmetry in a model with a discrete endogenous regressor. Chen (1999a,b) uses symmetry in a

discrete choice model; Chen and Zhou (2010) in a sample selection model. Other studies examining the identification power of symmetry include Chen and Zhou (2010), Newey (1991), Cosslett (1997), Bai and Ng (2001), and Chen, Khan, and Tang (2013).

The present paper contributes to this general literature by combining rank correlation estimation and pairwise-difference estimation and providing a new pairwise-difference rank estimation procedure under the symmetry condition. A key observation is that in the discrete game context, it is hard to construct an estimator by directly using the rank-ordering property or a pairwise difference. To resolve this challenge, we do both; that is, we take differences on observations and then construct a rank estimation procedure on the differences. While the idea of combining these is not new (Abrevaya (1999, 2003)), to the best of our knowledge, we are the first to do so for discrete games.[8]

The remainder of the paper is organized as follows. Section 1.2 introduces a new strategy for identifying competition effects in entry games. Section 1.3 describes an estimation procedure constructed from this novel identification approach. Section 1.4 concludes. Appendix A.1 extends the analysis to a richer model with multivariate observables and heteroskedasticity. Appendices A.2 and A.3 collect the proofs for the theorems for identification and estimation, respectively. An online supplementary appendix (Appendix S) gives proofs for Lemmas in Appendices A.1, A.2 and A.3.

1.2 Identification

This section illustrates our identification strategy in the simple two-player entry game discussed in Section 1.1. In the two-player entry game, the entry decisions of

[8]The discussion above is to some extent related to a broader literature on the endogeneity problem in the nonlinear model. Blundell and Powell (2003) provide a excellent survey of the nonlinear endogeneity problem with the continuous regressor. Studies of the discrete endogeneous regressor in the triangular discrete choice models include Newey and Powell (2003), Chesher (2005), Chesher (2010), Vytlacil and Yildiz (2007), Bhattacharya, Shaikh, and Vytlacil (2008), Imbens and Newey (2009), and Shaikh and Vytlacil (2011).

players 1 and 2, in market i, are represented by $(Y_{1i}, Y_{2i}) \in \{0,1\}^2$. Given this, we have four possible entry outcomes: $(0,0)$, $(0,1)$, $(1,0)$ and $(1,1)$. In addition, we restrict attention to a scalar observed characteristic, Z_{pi}, for each player p and each market i.[9] We use the distance from a store to its headquarters as the scalar observable in the discount retailers' entry game context. We further normalize the coefficient $(\beta_1, \beta_2) = (-1, -1)$, as this model is only identified up to scale. Throughout the paper, we use uppercase letters to denote random variables and lowercase letters to denote their realizations. Furthermore, we use boldface to denote vectors. Let $\boldsymbol{Y}_i$, $\boldsymbol{Z}_i$ and $\boldsymbol{\varepsilon}_i$ denote (Y_{1i}, Y_{2i}), (Z_{1i}, Z_{2i}), $(\varepsilon_{1i}, \varepsilon_{2i})$, respectively. The random vectors $\boldsymbol{Y}_i$, $\boldsymbol{Z}_i$ and $\boldsymbol{\varepsilon}_i$ take values in the sets $\mathcal{S}_{\boldsymbol{Y}}$, $\mathcal{S}_{\boldsymbol{Z}}$ and $\mathcal{S}_{\boldsymbol{\varepsilon}}$, where $\mathcal{S}_{\boldsymbol{Y}} = \{0,1\}^2$, $\mathcal{S}_{\boldsymbol{Z}} \subseteq \mathbb{R}^2$, $\mathcal{S}_{\boldsymbol{\varepsilon}} = \mathbb{R}^2$. In particular, we assume that the following regularity conditions hold.

Assumption R (Random Sampling): *An independent and identically distributed (i.i.d.) sample* $\{\boldsymbol{Y}_i, \boldsymbol{Z}_i, \boldsymbol{\varepsilon}_i\}_{i=1}^n$ *is drawn from the population.*

Assumption R restricts our analysis to an *i.i.d.* sample and assumes that firms make independent decisions across markets. This assumption is crucial to establishing our identification method.[10]

Assumption S (Sign): $\Delta_1 < 0, \Delta_2 < 0$.

Assumption S requires prior knowledge of the sign of the competition effects. Under Assumption S, entry outcomes $(0,0)$ and $(1,1)$ are uniquely predicted by the model. We will use the choice probabilities of these two unique equilibria to identify the parameters of interest. Though the model tends to generate the multiplicity of the equilibrium for $(1,0)$ and $(0,1)$, we will show later that it does not affect our identification strategy. As a final remark, the identification presented below can also

[9] In Appendix A.1, we extend the analysis to multivariate observed characteristics.

[10] This assumption may be not realistic for some applications, as discussed in Ellickson and Misra (2011). We will leave possible extensions to future work.

be applied to the case with $\Delta_1 > 0, \Delta_2 > 0$ (for more details on the unique equilibria in this case, see Tamer (2003)).

Assumption ER (Exclusion Restriction): *Suppose that*

(i) (Z_{1i}, Z_{2i}) *is independent of* $(\varepsilon_{1i}, \varepsilon_{2i})$*;*

(ii) the scalar covariate Z_{pi} *enters only the payoff function of player* p*, but not the payoff function of the other player.*

Assumption ER requires that each firm has an exogenous observed characteristic affecting its own profitability which does not directly affect the profitability of its rival. Assumption ER is commonly used in the existing literature since without it, identification is extremely difficult to obtain.[11] In the entry game example, variables that shift the fixed cost of one player but not the other will satisfy Assumption ER. In the discount retailing industry context, Jia (2008) assumes that the distance from a store to its headquarters is such a fixed cost shifter. We will also use this fixed cost shifter in our empirical application.

Assumption RS (Radial Symmetry): T*he distribution of the unobserved characteristics* $(\varepsilon_1, \varepsilon_2)$ *is continuous over the support* $\mathcal{S}_{\boldsymbol{\varepsilon}}$ and *radially symmetric around* (α_1, α_2)*; that is,* $f_{\boldsymbol{\varepsilon}}(\varepsilon_1, \varepsilon_2; \alpha_1, \alpha_2) = f_{\boldsymbol{\varepsilon}}(2\alpha_1 - \varepsilon_1, 2\alpha_2 - \varepsilon_2; \alpha_1, \alpha_2)$.[12]

Assumption RS means that any two realizations of $(\varepsilon_{1i}, \varepsilon_{2i})$, radiating equal distances in opposite directions from the symmetric point, have the same density. The symmetry point does not need to be known, and we treat the symmetry point as an additional set of parameters that we identify along with competition effects. Note that symmetry implies $\mathbb{E}(\varepsilon_1) = \alpha_1$ and $\mathbb{E}(\varepsilon_2) = \alpha_2$, where α_1 and α_2 are the re-

[11] The literature that uses the exclusion restriction includes Berry and Tamer (2006), Ciliberto and Tamer (2009), Bajari, Hong, and Ryan (2010), Khan and Nekipelov (2012), and Fox and Lazzati (2013).

[12] The terminology used here follows Nelsen (1993). Alternatively, this type of symmetry is also called the central symmetry in Serfling (2006).

spective means of the respective marginal distributions. A large class of distributions commonly used in empirical applications satisfy this condition, including the bivariate normal and the bivariate Laplace allowing arbitrary correlations.[13] In future work, we plan to develop a specification test for this condition.

1.2.1 Identification Strategy

This subsection provides restrictions necessary to identify the parameters of interest. Given that $\Delta_1 < 0, \Delta_2 < 0$, entry decisions $(0,0)$ and $(1,1)$ are uniquely predicted by the model. At any point (z_1, z_2), we can define a set of realizations of unobservables, $A_0(z_1, z_2)$, such that if $(\varepsilon_1, \varepsilon_2)$ is an element of that set, neither firm would choose to enter the market. Analogously, at any point (z_1, z_2), we can define a set of realizations of unobservables $A_1(z_1, z_2; \Delta_1, \Delta_2)$, such that if $(\varepsilon_1, \varepsilon_2)$ is an element of this set, both firms would choose to enter the market. Formally, A_0 is defined as follows

$$A_0(z_1, z_2) = \{(\varepsilon_1, \varepsilon_2) : \varepsilon_1 < z_1, \varepsilon_2 < z_2\};$$

and similarly, $A_1(z_1, z_2; \Delta_1, \Delta_2)$ is defined as follows

$$A_1(z_1, z_2; \Delta_1, \Delta_2) = \{(\varepsilon_1, \varepsilon_2) : \varepsilon_1 \geq z_1 - \Delta_1, \varepsilon_2 \geq z_2 - \Delta_2\}.$$

We can illustrate these two regions in Figure 1.1, similar to Bresnahan and Reiss (1991a), Tamer (2003), and Ciliberto and Tamer (2009).

Now, integrating $f_\varepsilon(\varepsilon_1, \varepsilon_2; \alpha_1, \alpha_2)$ over either $A_0(z_1, z_2)$ or $A_1(z_1, z_2; \Delta_1, \Delta_2)$ yields the probability that neither firm will enter the market or both firms will enter the market, respectively, given the value of observed variables. We refer to these probabilities as the conditional choice probabilities (CCP), which can be formally defined

[13] Also, all elliptically-contoured distributions satisfy radial symmetry.

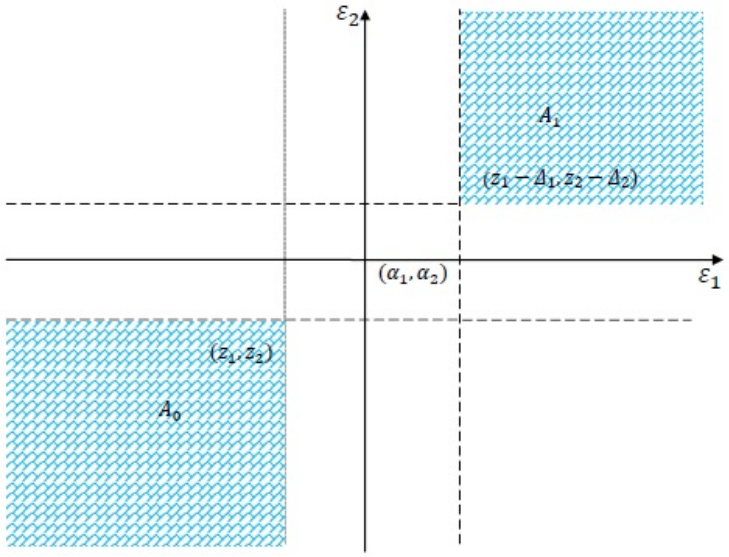

Figure 1.1: The integral region.

using the standard threshold crossing structure as follows:

$$\begin{aligned}
\Pr\left[(0,0)\,|\,(Z_{1i},Z_{2i})=(z_1,z_2)\right] &= \Pr\left(\varepsilon_{1i}<z_1,\varepsilon_{2i}<z_2\right)\\
&= \int\limits_{A_0(z_1,z_2)} f_\varepsilon(\varepsilon_1,\varepsilon_2;\alpha_1,\alpha_2)d\left(\varepsilon_1,\varepsilon_2\right);\\
\Pr\left[(1,1)\,|\,(Z_{1i},Z_{2i})=(z_1,z_2)\right] &= \Pr\left(\varepsilon_{1i}\geq z_1-\Delta_1,\varepsilon_{2i}\geq z_2-\Delta_2\right)\\
&= \int\limits_{A_1(z_1,z_2;\Delta_1,\Delta_2)} f_\varepsilon(\varepsilon_1,\varepsilon_2;\alpha_1,\alpha_2)d\left(\varepsilon_1,\varepsilon_2\right).
\end{aligned}$$

Note that $\Pr\left[(0,0)\,|\,(Z_{1i},Z_{2i})=(z_1,z_2)\right]$ depends on the symmetric point (α_1,α_2), and $\Pr\left[(1,1)\,|\,(Z_{1i},Z_{2i}]=(z_1,z_2)\right)$ depends on both (α_1,α_2) and the competition effects (Δ_1,Δ_2). Our identification strategy proceeds in two steps: (i) identifying the symmetric point (α_1,α_2) from $\Pr\left[(0,0)\,|\,(z_1,z_2)\right]$, and (ii) given (α_1,α_2), identifying the competition effects (Δ_1,Δ_2) from $\Pr\left[(1,1)\,|\,(z_1,z_2)\right]$.[14] To develop the intuition behind our identification strategy, we present a graphical illustration before proceeding to the formal derivation.

[14]Note that because α and Δ enter $\Pr\left[(1,1)\,|\,(Z_{1i},Z_{2i})=(z_1,z_2)\right]$ additively, one cannot recover (Δ_1,Δ_2) directly from this probability.

1.2.1.1 Identifying Restriction for the Symmetric Point

To identify the symmetric point (α_1, α_2), we use the choice probability of no entry, that is, $\Pr[(0,0)|z_1, z_2]$ $(=\Pr[(0,0) \mid (Z_{1i}, Z_{2i}) = (z_1, z_2)])$. The goal of our analysis is to find a restriction on choice probabilities for the same unique equilibrium $(0,0)$ across different locations, such that this restriction holds only at the true (α_1, α_2). In particular, the identifying restriction can be constructed in the following three steps.

First, consider two values for the observed variables (z_1, z_2) and $(\tilde{z}_1, \tilde{z}_2)$, where $z_1 < \tilde{z}_1$ and $z_2 < \tilde{z}_2$. Then, consider values of the observables that combine one element from each of those values, that is, $(z_1, \tilde{z}_2)$ and $(\tilde{z}_1, z_2)$. Define

$$R_0(\boldsymbol{z}, \tilde{\boldsymbol{z}}) = R_0(z_1, z_2, \tilde{z}_1, \tilde{z}_2) \equiv \{(\varepsilon_1, \varepsilon_2) : z_1 < \varepsilon_1 < \tilde{z}_1, z_2 < \varepsilon_2 < \tilde{z}_2\}$$

where $\boldsymbol{z} = (z_1, z_2)$ and $\tilde{\boldsymbol{z}} = (\tilde{z}_1, \tilde{z}_2)$. We will now construct the probability of observing $(\varepsilon_1, \varepsilon_2)$ in the region $R_0(z_1, z_2, \tilde{z}_1, \tilde{z}_2)$,

$$\begin{aligned}
& \Pr((\varepsilon_1, \varepsilon_2) \in R_0(z_1, z_2, \tilde{z}_1, \tilde{z}_2)) \\
= \; & \Pr((\varepsilon_1, \varepsilon_2) \in A_0(z_1, z_2)) + \Pr((\varepsilon_1, \varepsilon_2) \in A_0(\tilde{z}_1, \tilde{z}_2)) \\
& - \Pr((\varepsilon_1, \varepsilon_2) \in A_0(z_1, \tilde{z}_2)) - \Pr((\varepsilon_1, \varepsilon_2) \in A_0(\tilde{z}_1, z_2)) \\
= \; & \Pr[(0,0) \mid (z_1, z_2)] + \Pr[(0,0) \mid (\tilde{z}_1, \tilde{z}_2)] - \Pr[(0,0) \mid (z_1, \tilde{z}_2)] - \Pr[(0,0) \mid (\tilde{z}_1, z_2)] \\
\equiv \; & B_0(\boldsymbol{z}, \tilde{\boldsymbol{z}}; \boldsymbol{\alpha}).
\end{aligned}$$

The intuition here is that the set $R_0(z_1, z_2, \tilde{z}_1, \tilde{z}_2)$ can be decomposed as a combination of the sets $A_0(z_1, z_2)$, $A_0(\tilde{z}_1, \tilde{z}_2)$, $A_0(z_1, \tilde{z}_2)$ and $A_0(\tilde{z}_1, z_2)$, which is shown in Figure 1.5, with the assumption that $(\alpha_1, \alpha_2) = (0,0)$. This probability is equal to the linear combination of the choice probabilities given four values of observables.

Importantly, each of the four choice probabilities that are used in constructing $B_0(\boldsymbol{z}, \tilde{\boldsymbol{z}}; \boldsymbol{\alpha})$ can be obtained directly from the data, that is, we can recover the prob-

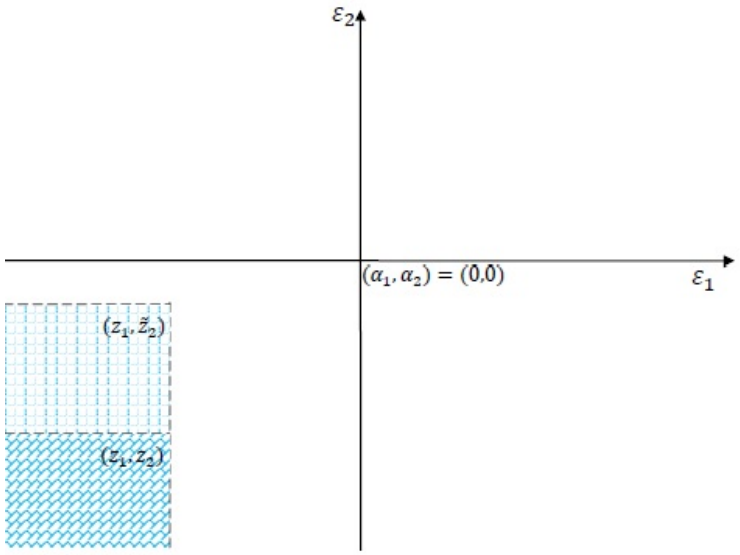

Figure 1.2: $A_0(\tilde{z}_1, \tilde{z}_2) \backslash A_0(z_1, z_2)$

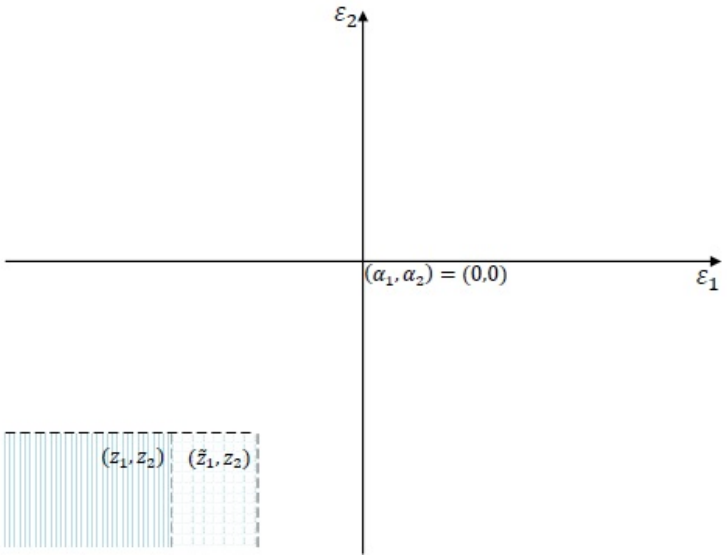

Figure 1.3: $A_0(z_1, \tilde{z}_2) \backslash A_0(z_1, z_2)$

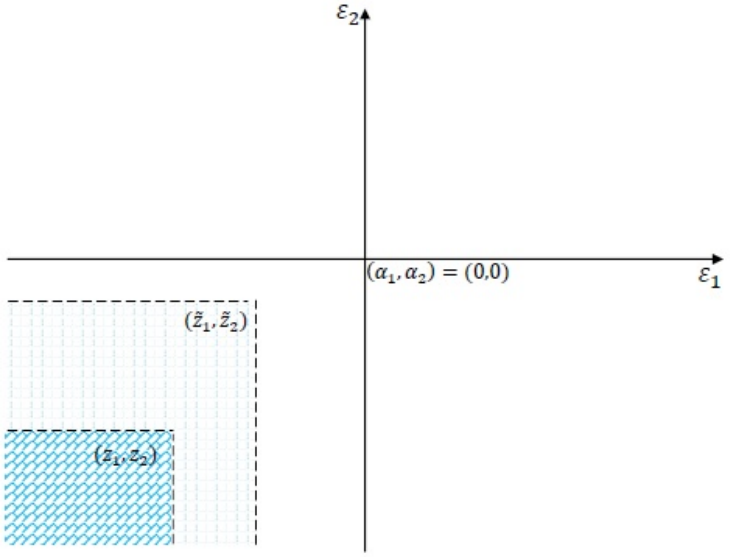

Figure 1.4: $A_0(\tilde{z}_1, z_2) \backslash A_0(z_1, z_2)$

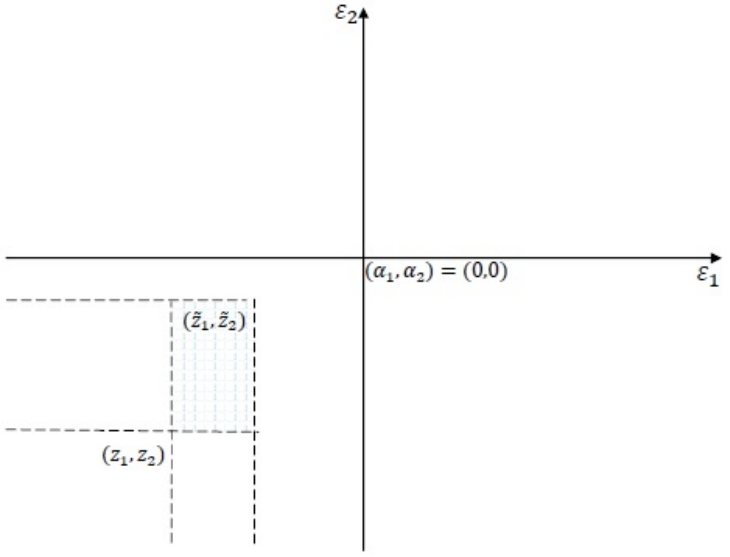

Figure 1.5: $R_0(z_1, z_2, \tilde{z}_1, \tilde{z}_2)$

ability of $(\varepsilon_1, \varepsilon_2)$ lying in the rectangular area $R_0(z_1, z_2, \tilde{z}_1, \tilde{z}_2)$ by adding and subtracting choice probabilities from the data.[15] $\Pr[(0,0) \mid (z_1, z_2)]$, for example, can be recovered by considering the proportion of locations in the data, with the observed variable value (z_1, z_2) for which the outcome $(0,0)$ is observed.

Next, we utilize our symmetry assumption (Assumption RS). For the true symmetric point given by (α_1, α_2), consider the values of observed variables that are reflected through the symmetry point, as follows:

$$(2\alpha_1 - z_1, 2\alpha_2 - z_2), (2\alpha_1 - \tilde{z}_2, 2\alpha_2 - \tilde{z}_2), (2\alpha_1 - z_1, 2\alpha_2 - \tilde{z}_2), (2\alpha_1 - \tilde{z}_1, 2\alpha_2 - z_2).$$

These are the original four locations reflected through the symmetry point. Following from the above, we can construct

$$R_0(2\boldsymbol{\alpha} - \boldsymbol{z}, 2\boldsymbol{\alpha} - \tilde{\boldsymbol{z}}) \quad \text{and} \quad B_0(2\boldsymbol{\alpha} - \boldsymbol{z}, 2\boldsymbol{\alpha} - \tilde{\boldsymbol{z}}; \boldsymbol{\alpha});$$

with the four new locations. The rectangle $R_0(2\boldsymbol{\alpha} - \boldsymbol{z}, 2\boldsymbol{\alpha} - \tilde{\boldsymbol{z}})$ is the reflection of $R_0(\boldsymbol{z}, \tilde{\boldsymbol{z}})$ through the symmetry point. We illustrate $R_0(\boldsymbol{z}, \tilde{\boldsymbol{z}})$ and $R_0(2\boldsymbol{\alpha} - \boldsymbol{z}, 2\boldsymbol{\alpha} - \tilde{\boldsymbol{z}})$ in Figure 1.6 with the assumption that $(\alpha_1, \alpha_2) = (0,0)$.

Finally, Assumption RS then implies that the density of each point $(\varepsilon_1, \varepsilon_2)$ in $R_0(\boldsymbol{z}, \tilde{\boldsymbol{z}})$ is equal to the density of its reflected point $(2\alpha_1 - \varepsilon_1, 2\alpha_2 - \varepsilon_2)$ in $R_0(2\boldsymbol{\alpha} - \boldsymbol{z}, 2\boldsymbol{\alpha} - \tilde{\boldsymbol{z}})$. Hence,

$$B_0(\boldsymbol{z}, \tilde{\boldsymbol{z}}; \boldsymbol{\alpha}) = B_0(2\boldsymbol{\alpha} - \boldsymbol{z}, 2\boldsymbol{\alpha} - \tilde{\boldsymbol{z}}; \boldsymbol{\alpha}).$$

Now, we formalize our discussion in Lemma I.1.

[15] $R_0(z_1, z_2, \tilde{z}_1, \tilde{z}_2) = [A_0(\tilde{z}_1, \tilde{z}_2) \setminus A_0(z_1, z_2)] \setminus [(A_0(\tilde{z}_1, z_2) \setminus A_0(z_1, z_2)) \cup (A_0(z_1, \tilde{z}_2) \setminus A_0(z_1, z_2))]$.

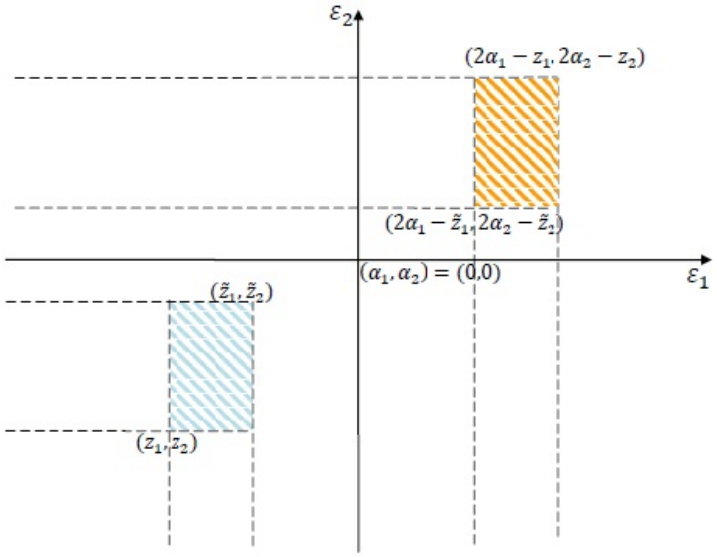

Figure 1.6: Identifying restriction for (α_1, α_2)

Lemma I.1. *For any two vectors* $\boldsymbol{z} = (z_1, z_2)$, $\tilde{\boldsymbol{z}} = (\tilde{z}_1, \tilde{z}_2) \in \mathcal{S}_{\boldsymbol{z}}$, *consider*

$$\begin{aligned}
\boldsymbol{Z}_1 &= (z_1, z_2)\,; \qquad \boldsymbol{Z}_5 = (2\alpha_1 - z_1, 2\alpha_2 - z_2)\,; \\
\boldsymbol{Z}_2 &= (\tilde{z}_1, \tilde{z}_2)\,; \qquad \boldsymbol{Z}_6 = (2\alpha_1 - \tilde{z}_1, 2\alpha_2 - \tilde{z}_2)\,; \\
\boldsymbol{Z}_3 &= (z_1, \tilde{z}_2)\,; \qquad \boldsymbol{Z}_7 = (2\alpha_1 - z_1, 2\alpha_2 - \tilde{z}_2)\,; \\
\boldsymbol{Z}_4 &= (\tilde{z}_1, z_2)\,; \qquad \boldsymbol{Z}_8 = (2\alpha_1 - \tilde{z}_1, 2\alpha_2 - z_2)\,.
\end{aligned}$$

Given that Assumptions R, S and ER hold, define

$$\begin{aligned}
B_0(\boldsymbol{z}, \tilde{\boldsymbol{z}}; \boldsymbol{\alpha}) &= \Pr[(0,0)\,|\boldsymbol{Z}_1] + \Pr[(0,0)\,|\boldsymbol{Z}_2] - \Pr[(0,0)\,|\boldsymbol{Z}_3] - \Pr[(0,0)\,|\boldsymbol{Z}_4]; \\
B_0(2\boldsymbol{\alpha} - \boldsymbol{z}, 2\boldsymbol{\alpha} - \tilde{\boldsymbol{z}}; \boldsymbol{\alpha}) &= \Pr[(0,0)\,|\boldsymbol{Z}_5] + \Pr[(0,0)\,|\boldsymbol{Z}_6] - \Pr[(0,0)\,|\boldsymbol{Z}_7] - \Pr[(0,0)\,|\boldsymbol{Z}_8].
\end{aligned}$$

By Assumption RS, we have

$$B_0(\boldsymbol{z}, \tilde{\boldsymbol{z}}; \boldsymbol{\alpha}) - B_0(2\boldsymbol{\alpha} - \boldsymbol{z}, 2\boldsymbol{\alpha} - \tilde{\boldsymbol{z}}; \boldsymbol{\alpha}) = 0. \tag{1.2}$$

Lemma I.1 gives us our fundamental identifying restriction on (α_1, α_2), which we

will use to define our identification definition later.

Next, we will apply the same procedure to derive the identifying restriction for the competition effects.

1.2.1.2 Identifying Restriction for the Competition Effects

The strategy for identifying the competition effects (Δ_1, Δ_2) is analogous to the one above but focuses on the equilibrium $(1,1)$ rather than $(0,0)$ that is used in identifying the symmetry point. More specifically, we use the choice probability of entry to identifying $(\alpha_1 + \Delta_1, \alpha_2 + \Delta_2)$, that is, $\Pr[(1,1)|z_1, z_2]$ $(= \Pr[(1,1) \mid (Z_{1i}, Z_{2i}) = (z_1, z_2)])$, and since we have already identified (α_1, α_2), so that we can recover (Δ_1, Δ_2).

First, consider four values as above, $(z_1, z_2), (\tilde{z}_1, \tilde{z}_2), (z_1, \tilde{z}_2)$ and $(\tilde{z}_1, z_2)$, where $z_1 < \tilde{z}_1$ and $z_2 < \tilde{z}_2$. Define

$$\begin{aligned} R_1(\boldsymbol{z}, \tilde{\boldsymbol{z}}) &= R_1(z_1, z_2, \tilde{z}_1, \tilde{z}_2) \\ &\equiv \{(\varepsilon_1, \varepsilon_2) : z_1 - \Delta_1 < \varepsilon_1 < \tilde{z}_1 - \Delta_1, z_2 - \Delta_2 < \varepsilon_2 < \tilde{z}_2 - \Delta_2\}, \end{aligned}$$

where $\boldsymbol{z} = (z_1, z_2)$ and $\tilde{\boldsymbol{z}} = (\tilde{z}_1, \tilde{z}_2)$. The probability of observing $(\varepsilon_1, \varepsilon_2)$ in the region $R_1(\boldsymbol{z}, \tilde{\boldsymbol{z}})$ is given by

$$\begin{aligned} & \Pr((\varepsilon_1, \varepsilon_2) \in R_1(\boldsymbol{z}, \tilde{\boldsymbol{z}})) \\ = \; & \Pr((\varepsilon_1, \varepsilon_2) \in A_1(z_1, z_2; \Delta_1, \Delta_2)) + \Pr((\varepsilon_1, \varepsilon_2) \in A_1(\tilde{z}_1, \tilde{z}_2; \Delta_1, \Delta_2)) \\ & - \Pr((\varepsilon_1, \varepsilon_2) \in A_1(z_1, \tilde{z}_2; \Delta_1, \Delta_2)) - \Pr((\varepsilon_1, \varepsilon_2) \in A_1(\tilde{z}_1, z_2; \Delta_1, \Delta_2)) \\ \equiv \; & \Pr[(1,1) \mid (z_1, z_2)] + \Pr[(1,1) \mid (\tilde{z}_1, \tilde{z}_2)] - \Pr[(1,1) \mid (z_1, \tilde{z}_2)] - \Pr[(1,1) \mid (\tilde{z}_1, z_2)] \\ \equiv \; & B_1(\boldsymbol{z}, \tilde{\boldsymbol{z}}; \boldsymbol{\alpha}, \boldsymbol{\Delta}). \end{aligned}$$

Next, we use the symmetry condition (Assumption RS) again, for the symmetric points plus the competition effects given by $(\alpha_1 + \Delta_1, \alpha_2 + \Delta_2)$. Now, we also consider

four combinations of observables, which are given by

$$\left(2\left(\alpha_1+\Delta_1\right)-z_1, 2\left(\alpha_2+\Delta_2\right)-z_2\right), \left(2\left(\alpha_1+\Delta_1\right)-\tilde{z}_1, 2\left(\alpha_2+\Delta_2\right)-\tilde{z}_2\right),$$
$$\left(2\left(\alpha_1+\Delta_1\right)-z_1, 2\left(\alpha_2+\Delta_2\right)-\tilde{z}_2\right), \left(2\left(\alpha_1+\Delta_1\right)-\tilde{z}_1, 2\left(\alpha_2+\Delta_2\right)-z_2\right).$$

These four points are then reflected through the symmetry point. For $z_1 < \tilde{z}_1$ and $z_2 < \tilde{z}_2$, define

$$\begin{aligned} & R_1\left(2\left(\boldsymbol{\alpha}+\boldsymbol{\Delta}\right)-\boldsymbol{z}, 2\left(\boldsymbol{\alpha}+\boldsymbol{\Delta}\right)-\tilde{\boldsymbol{z}}\right) \\ = & \left\{ \begin{array}{ll} (\varepsilon_1,\varepsilon_2): & 2\left(\alpha_1+\Delta_1\right)-\tilde{z}_1-\Delta_1 < \varepsilon_1 < 2\left(\alpha_1+\Delta_1\right)-z_1-\Delta_1; \\ & 2\left(\alpha_2+\Delta_2\right)-\tilde{z}_2-\Delta_2 < \varepsilon_2 < 2\left(\alpha_2+\Delta_2\right)-z_2-\Delta_2. \end{array} \right\} \\ = & \left\{ \begin{array}{ll} (\varepsilon_1,\varepsilon_2): & 2\alpha_1-\left(\tilde{z}_1-\Delta_1\right) < \varepsilon_1 < 2\alpha_1-\left(z_1-\Delta_1\right); \\ & 2\alpha_2-\left(\tilde{z}_2-\Delta_2\right) < \varepsilon_2 < 2\alpha_2-\left(z_2-\Delta_2\right). \end{array} \right\} \end{aligned}$$

We illustrate $R_1\left(\boldsymbol{z}, \tilde{\boldsymbol{z}}\right)$ and $R_1\left(2\left(\boldsymbol{\alpha}+\boldsymbol{\Delta}\right)-\boldsymbol{z}, 2\left(\boldsymbol{\alpha}+\boldsymbol{\Delta}\right)-\tilde{\boldsymbol{z}}\right)$ under the assumption $(\alpha_1, \alpha_2) = (0, 0)$, in Figure 1.7. Then we can also define $B_1\left(2\left(\boldsymbol{\alpha}+\boldsymbol{\Delta}\right)-\boldsymbol{z}, 2\left(\boldsymbol{\alpha}+\boldsymbol{\Delta}\right)-\tilde{\boldsymbol{z}}; \boldsymbol{\alpha}, \boldsymbol{\Delta}\right)$ accordingly.

Finally, by the same argument for the identification of the symmetry points using Assumption RS, we can show

$$B_1\left(\boldsymbol{z}, \tilde{\boldsymbol{z}}; \boldsymbol{\alpha}, \boldsymbol{\Delta}\right) = B_1\left(2\left(\boldsymbol{\alpha}+\boldsymbol{\Delta}\right)-\boldsymbol{z}, 2\left(\boldsymbol{\alpha}+\boldsymbol{\Delta}\right)-\tilde{\boldsymbol{z}}; \boldsymbol{\alpha}, \boldsymbol{\Delta}\right)$$

which is similar to the identifying restriction of (α_1, α_2).

Now, we formalize our argument to define the identification of (Δ_1, Δ_2) in this paper.

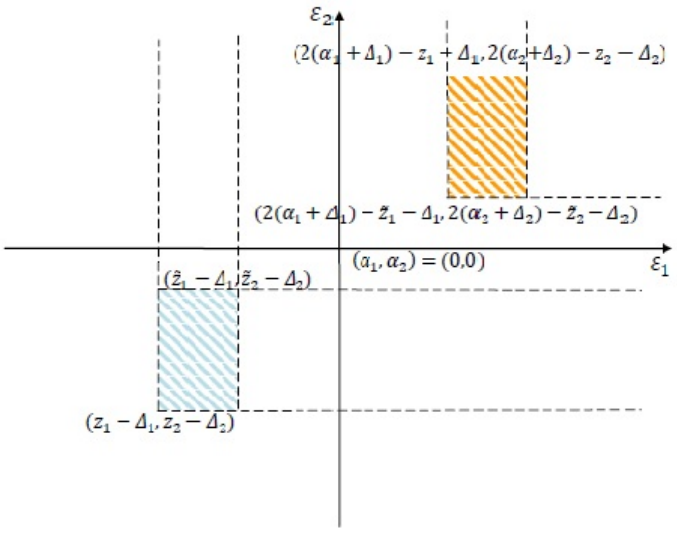

Figure 1.7: Identifying restriction for $(\alpha_1 + \Delta_1, \alpha_2 + \Delta_2)$

Lemma I.2. *For any two vectors* $\boldsymbol{z} = (z_1, z_2)$, $\tilde{\boldsymbol{z}} = (\tilde{z}_1, \tilde{z}_2) \in \mathcal{S}_{\boldsymbol{z}}$, *consider*

$$\begin{aligned}
\boldsymbol{Z}_1 &= (z_1, z_2)\,; \qquad \boldsymbol{Z}_5 = (2(\alpha_1+\Delta_1) - z_1, 2(\alpha_2+\Delta_2) - z_2)\,; \\
\boldsymbol{Z}_2 &= (\tilde{z}_1, \tilde{z}_2)\,; \qquad \boldsymbol{Z}_6 = (2(\alpha_1+\Delta_1) - \tilde{z}_1, 2(\alpha_2+\Delta_2) - \tilde{z}_2)\,; \\
\boldsymbol{Z}_3 &= (z_1, \tilde{z}_2)\,; \qquad \boldsymbol{Z}_7 = (2(\alpha_1+\Delta_1) - z_1, 2(\alpha_2+\Delta_2) - \tilde{z}_2)\,; \\
\boldsymbol{Z}_4 &= (\tilde{z}_1, z_2)\,; \qquad \boldsymbol{Z}_8 = (2(\alpha_1+\Delta_1) - \tilde{z}_1, 2(\alpha_2+\Delta_2) - z_2)\,.
\end{aligned}$$

Given that Assumptions R, S and ER hold, define

$$\begin{aligned}
& B_1(\boldsymbol{z}, \tilde{\boldsymbol{z}}; \boldsymbol{\alpha}, \boldsymbol{\Delta}) \\
= \;& \Pr[(1,1)\,|\boldsymbol{Z}_1] + \Pr[(1,1)\,|\boldsymbol{Z}_2] - \Pr[(1,1)\,|\boldsymbol{Z}_3] - \Pr([1,1)\,|\boldsymbol{Z}_4]; \\
& B_1(2(\boldsymbol{\alpha}+\boldsymbol{\Delta}) - \boldsymbol{z}, 2(\boldsymbol{\alpha}+\boldsymbol{\Delta}) - \tilde{\boldsymbol{z}}; \boldsymbol{\alpha}, \boldsymbol{\Delta}) \\
= \;& \Pr[(1,1)\,|\boldsymbol{Z}_5] + \Pr[(1,1)\,|\boldsymbol{Z}_6] - \Pr[(1,1)\,|\boldsymbol{Z}_7] - \Pr[(1,1)\,|\boldsymbol{Z}_8].
\end{aligned}$$

By Assumption RS, we have

$$B_1(\boldsymbol{z}, \tilde{\boldsymbol{z}}; \boldsymbol{\alpha}, \boldsymbol{\Delta}) - B_1(2(\boldsymbol{\alpha}+\boldsymbol{\Delta}) - \boldsymbol{z}, 2(\boldsymbol{\alpha}+\boldsymbol{\Delta}) - \tilde{\boldsymbol{z}}; \boldsymbol{\alpha}, \boldsymbol{\Delta}) = 0. \tag{1.3}$$

Equalities (1.2) and (1.3) provide the fundamental identifying restrictions on $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$ and $\boldsymbol{\alpha}+\boldsymbol{\Delta} = (\alpha_1 + \Delta_1, \alpha_2 + \Delta_2)$. Again, $B_0(\boldsymbol{z}, \tilde{\boldsymbol{z}}; \boldsymbol{\alpha})$ and $B_0(2\boldsymbol{\alpha} - \boldsymbol{z}, 2\boldsymbol{\alpha} - \tilde{\boldsymbol{z}}; \boldsymbol{\alpha})$ can be constructed directly from the data. Note that at the true (α_1, α_2), equality (1.2) will hold for all $\boldsymbol{z}, \tilde{\boldsymbol{z}}$. The question of identification is then whether (1.2) might also hold for $\boldsymbol{a} \neq \boldsymbol{\alpha}$, $\boldsymbol{a} \in \Theta_\alpha \subset \mathbb{R}^2$ for all $\boldsymbol{z}, \tilde{\boldsymbol{z}}$, where Θ_α is the (bounded) parameter space.

1.2.2 Definitions and Sufficient Conditions

In this section, we formalize our identification definitions and provide a set of sufficient conditions for point identification.

Definition I.3. (Radial Symmetry-Discrete Response Identification) Let $\boldsymbol{a} = (a_1, a_2) \in \Theta_\alpha \subset \mathbb{R}^2$. Let

$$T(\boldsymbol{a}) = \left\{ (\boldsymbol{z}, \tilde{\boldsymbol{z}}) \;\middle|\; \begin{array}{l} B_0(\boldsymbol{z}, \tilde{\boldsymbol{z}}; \boldsymbol{\alpha}) - B_0(2\boldsymbol{a} - \boldsymbol{z}, 2\boldsymbol{a} - \tilde{\boldsymbol{z}}; \boldsymbol{\alpha}) \neq 0; \\ \boldsymbol{z}, \tilde{\boldsymbol{z}} \in \mathcal{S}_{\boldsymbol{z}}, 2\boldsymbol{a} - \boldsymbol{z}, 2\boldsymbol{a} - \tilde{\boldsymbol{z}} \in \mathcal{S}_{\boldsymbol{z}} \end{array} \right\}$$

(i) We say that $\boldsymbol{\alpha}$ is RSDR identified relative to $\boldsymbol{a}$ if

$$\Pr\left[\left(\boldsymbol{Z}, \tilde{\boldsymbol{Z}}\right) \in T(\boldsymbol{a})\right] > 0.$$

(ii) In addition, we say that $\boldsymbol{\alpha}$ is RSDR point identified if for all $\boldsymbol{a} \neq \boldsymbol{\alpha}$,

$$\Pr\left[\left(\boldsymbol{Z}, \tilde{\boldsymbol{Z}}\right) \in T(\boldsymbol{a})\right] > 0.$$

Definition I.4. (Radial Symmetry-Discrete Response Identification) Let $\boldsymbol{a} = (a_1, a_2) \in \Theta_\alpha$, $\boldsymbol{\delta} = (\delta_1, \delta_2) \in \Theta_\Delta$, where $\Theta_\alpha, \Theta_\Delta \subset \mathbb{R}^2$. Let

$$T(\boldsymbol{a} + \boldsymbol{\delta}) = \left\{ (\boldsymbol{z}, \tilde{\boldsymbol{z}}) \;\middle|\; \begin{array}{l} B_1(\boldsymbol{z}, \tilde{\boldsymbol{z}}; \boldsymbol{\alpha}, \boldsymbol{\Delta}) - B_0(2(\boldsymbol{a} + \boldsymbol{\delta}) - \boldsymbol{z}, 2(\boldsymbol{a} + \boldsymbol{\delta}) - \tilde{\boldsymbol{z}}; \boldsymbol{\alpha}, \boldsymbol{\Delta}) \neq 0; \\ \boldsymbol{z}, \tilde{\boldsymbol{z}} \in \mathcal{S}_{\boldsymbol{z}}, 2(\boldsymbol{a} + \boldsymbol{\delta}) - \boldsymbol{z}, 2(\boldsymbol{a} + \boldsymbol{\delta}) - \tilde{\boldsymbol{z}} \in \mathcal{S}_{\boldsymbol{z}}. \end{array} \right\}$$

(i)We say that $\boldsymbol{\alpha}+\boldsymbol{\Delta}$ is RSDR identified relative to $\boldsymbol{a}+\boldsymbol{\delta}$ if

$$\Pr\left[\left(\boldsymbol{Z},\tilde{\boldsymbol{Z}}\right)\in T\left(\boldsymbol{a}+\boldsymbol{\delta}\right)\right]>0.$$

(ii) In addition, we say that $\boldsymbol{\alpha}+\boldsymbol{\Delta}$ is RSDR point identified if for all $\boldsymbol{a}+\boldsymbol{\delta}\neq\boldsymbol{\alpha}+\boldsymbol{\Delta}$,

$$\Pr\left[\left(\boldsymbol{Z},\tilde{\boldsymbol{Z}}\right)\in T\left(\boldsymbol{a}+\boldsymbol{\delta}\right)\right]>0.$$

Part(i) of Definition I.3 suggests that we can identify $\boldsymbol{\alpha}$ relative to a particular alternative $\boldsymbol{a}\neq\boldsymbol{\alpha}$, if there exists a set of $(\boldsymbol{z},\tilde{\boldsymbol{z}})$ with positive probability such that (1.2) is violated at $\boldsymbol{a}$. Part (ii) suggests that we can point identify $\boldsymbol{\alpha}$ if, for any arbitrary $\boldsymbol{a}\neq\boldsymbol{\alpha}$, $\boldsymbol{a}\in\Theta_{\alpha}$, we can find such a set. Similar arguments can be applied to identify $(\boldsymbol{\alpha}+\boldsymbol{\Delta})$ in Definition I.4. These definitions will be used to show that model parameters are identified. We will formally describe sufficient conditions below.

Now, we provide a set of sufficient conditions for point identification. Given the construction of the identification strategy, the sufficient conditions ensure that the support of the excluded observables covers the symmetry point (α_1,α_2) and the symmetry point plus the competition effects $(\alpha_1+\Delta_1,\alpha_2+\Delta_2)$. In addition, the support of observables must be sufficiently large to rule out alternative points which might appear to be symmetric over the support of the observables. Now, we introduce sufficient conditions for point identification in our paper.

Assumption SV (Sufficient Variation) Given any set $S\subset\mathcal{S}_{\boldsymbol{Z}}$ and a vector $\boldsymbol{a}=(a_1,a_2)$, $\boldsymbol{\delta}=(\delta_1,\delta_2)$, define the symmetrically reflected sets

$$S^{'}\left(S,\boldsymbol{a}\right)=\{\ \ (2a_1-z_1,2a_2-z_2)|\ \text{for all}\ \ (z_1,z_2)\in S\};$$

and

$$S^{'}\left(S,\boldsymbol{a}+\boldsymbol{\delta}\right)=\{\ \ \left(2\left(a_1+\delta_1\right)-z_1,2\left(a_2+\delta_2\right)-z_2\right)|\ \text{for all}\ \ \left(z_1,z_2\right)\in S\}\,.$$

(i) The points (α_1,α_2) and $(\alpha_1+\Delta_1,\alpha_2+\Delta_2)$ are in the interior of the support $\mathcal{S}_{\boldsymbol{Z}}$;

(ii) The random vector $\boldsymbol{Z}=(Z_1,Z_2)$ is absolutely continuously distributed with the positive density $f_{(Z_1,Z_2)}(\cdot,\cdot)$ over the support of $\mathcal{S}_{\boldsymbol{Z}}$, with respect to the Lebesgue measure;

(iii) For all $\boldsymbol{a}\in\mathcal{S}_{\boldsymbol{Z}}$ such that $\boldsymbol{a}\neq\boldsymbol{\alpha}$, there exists a Lebesgue measurable set $S\subset\mathcal{S}_{\boldsymbol{Z}}$ with positive measure such that $S'(S,\boldsymbol{a})\subset\mathcal{S}_{\boldsymbol{Z}}$ and

$$f_\varepsilon\left(z_1,z_2\right)\neq f_\varepsilon\left(2a_1-z_1,2a_2-z_2\right)\ \text{a.e. for all}\ \ \left(z_1,z_2\right)\in S.$$

Moreover, for all $\boldsymbol{a}+\boldsymbol{\delta}\in\mathcal{S}_{\boldsymbol{Z}}$ such that $\boldsymbol{a}+\boldsymbol{\delta}\neq\boldsymbol{\alpha}+\boldsymbol{\Delta}$, there exists a Lebesgue measurable set with positive measure $S\subset\mathcal{S}_{\boldsymbol{Z}}$ such that $S^{'}\left(S,\boldsymbol{a}+\boldsymbol{\delta}\right)\subset\mathcal{S}_{\boldsymbol{Z}}$ and

$$f_\varepsilon\left(z_1,z_2\right)\neq f_\varepsilon\left(2\left(a_1+\delta_1\right)-z_1,2\left(a_2+\delta_2\right)-z_2\right)\ \text{a.e. for all}\ \ \left(z_1,z_2\right)\in S.$$

Assumption SV-(i) assumes that support of observables depends on the parameter value. Similar parameter-dependent support assumptions are made in Vytlacil and Yildiz (2007, pp.764), where the size of the support depends on the strength of the exogenous regressor relative to the effect of the endogenous regressor. Assumption SV-(iii) essentially rules out alternative parameter values in the support of the data that "look like" symmetry points. A joint distribution can have only one symmetry point, but if one observes that distribution over only a portion of its support, there may be multiple points that "appear" symmetric. For example, suppose one observes a distribution only over a portion of its support. Then, over this portion of the

support, the distribution is periodic, and it integrates to 0.1. In this case, any of the local symmetric points in the observed support could be the symmetric point of the distribution. Assumption SV-(iii) rules out such points; that is, it rules out the distribution function being periodic over the support of the data.[16]

Under Assumption SV, we provide our key identification result in Theorem I.5.

Theorem I.5. *Suppose that Assumptions R, S, ER, RS and SV hold. Then* $\boldsymbol{\alpha}$ *and* $\boldsymbol{\alpha} + \boldsymbol{\Delta}$ *are point identified as defined in Definitions I.3 and I.4.*

In addition, if we are willing to assume that the distribution of unobservables is unimodal, then Assumption SV-(iii) is implied by a unimodal distribution and we have Theorem I.6.

Theorem I.6. *Suppose that Assumptions R, S, ER, RS and SV-(i)(ii) hold, and the joint distribution of unobservables* $(\varepsilon_1, \varepsilon_2)$ *is unimodal. Then* $\boldsymbol{\alpha}$ *and* $\boldsymbol{\alpha} + \boldsymbol{\Delta}$ *are point identified.*

As our point identification result relies highly on Assumption SV, further discussion of this relationship is warranted. Conditional on the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\alpha} + \boldsymbol{\Delta}$ and $\boldsymbol{a}$, $\boldsymbol{a} + \boldsymbol{\delta}$ as well as $f_{\boldsymbol{\varepsilon}}(\varepsilon_1, \varepsilon_2)$, Assumption SV only requires bounded support of observables. That said, our support condition is weaker than that used in the prior literature for this model (Tamer (2003) and Fox and Lazzati (2013)). But since one obviously does not know the parameters ahead of time, it would be inappropriate to describe our identification result as the one that depends on bounded support. Rather, we would describe our identification result as the one that depends on the bounded support given a compact set of parameters. More specifically, for their point identification results, one would need unbounded support of the observables even if the parameters were known to lie in a compact set. For our point identification result,

[16]Note that one may think that uniform distribution could always violate our identification strategy. However, since uniform distribution has finite support, it is automatically excluded from our discussion as we focus on the infinite support of unobservables.

if the parameters are known to lie in bounded support, we need only bounded support for the observables.

To conclude this section, we will give several remarks related to the identification strategy presented above.

Remark I.7. The identification of model 1.1 is confronted with two main difficulties: (i) if we use only the unique equilibrium to identify the parameter, we have the limited information that can be used in identification, that is, only the choice probability of $(0,0)$ and $(1,1)$ can be used for identification; and (ii) the binary feature of the endogenous regressor. To tackle the first difficulty, we identify the parameters by observing how the choice probabilities change as the observables change across the locations. In addition, to address the binary feature of the endogenous regressor, we transform the problem into a pairwise-difference comparison rather than a pairwise comparison problem.[17,18]

Remark I.8. Though we illustrate the method in the two-player entry game case, the proposed method can be directly extended to the case with more than two-player, if we assume that one firm's negative effects on its rivals are the same. We note that as the number of players increases, the ratio of the uniquely predicted entry outcome is decreasing relative to the total possible entry outcomes in the model. This fact suggests that our method will likely have decreased identification power as

[17]Because of these two difficulties, we note that the standard pairwise-difference identification and rank-order identification arguments cannot be directly applied to this model, since the location parameters will be differenced out by using the standard pairwise difference or rank estimation approach. As a motivation for our identification procedure, we observe that the location parameters do not affect the relative magnitude rather than the absolute magnitude of the choice probability. Given this observation, we first use the pairwise-difference to obtain a certain form of the absolute magnitude of the choice probabilities. Second, under the symmetry condition, we can assign the relation on these forms of the magnitude of the choice probabilities.

[18]In addition, we also find that though Chen (2000) provides a novel approach for identifying the location parameters in the single-agent model, his approach cannot be directly extended to the two-agent model unless we are willing to assume a stronger symmetry condition. We have shown that the location parameters can be identified under spherical symmetry (or called the joint symmetry) following Chen (2000). However, since spherical symmetry requires that the two unobservables be uncorrelated, which cannot be satisfied in most empirical applications, we do not present the results here.

we increase the number of players. Note that when we allow for competition effects indexed by the rival's identity, the number of competition effects increases as the number of players increases as discussed in Fox and Lazzati (2013). In this case, our identification strategy can only identify the sum of competition effects up to the constant. Furthermore, when the competition effects (Δ_1, Δ_2) are positive, the argument above cannot be directly applied, since with three or more players there does not necessarily exist a unique equilibrium, as emphasized in Fox and Lazzati (2013). They suggest that one possible way to solve this issue is to impose an equilibrium selection mechanism.

Remark I.9. We construct the identifying restriction using a certain combination of observables. In fact, a similar identifying restriction can be constructed by using other possible combination of observables. In other words, the parameters will be overidentified. However, the more combinations we use, the larger the computational burden is. Thus, in this paper, we only focus on the one proposed here.

Remark I.10. Finally, note that our identification strategy does not use the choice probabilities of $(0, 1)$ and $(1, 0)$ in order to avoid issues associated with multiple equilibria. A caveat of this is that our approach may lose efficiency relative to a procedure that does use these choice probabilities..

1.3 Estimation

In this section, we propose an estimation procedure based on the identifying restriction discussed in Section 1.2. When identifying the symmetric points (α_1, α_2) and the symmetric point plus the competition effects $(\alpha_1 + \Delta_1, \alpha_2 + \Delta_2)$, we use the same identification strategy but apply to different unique equilibrium outcomes ($(0, 0)$ and $(1, 1)$ respectively). As such we do the same with our estimator. When estimating

(α_1, α_2), we define

$$d = d^{00} = \mathbb{I}\left((Y_{1i}, Y_{2i}) = (0,0)\right);$$

and when estimating $(\alpha_1 + \Delta_1, \alpha_2 + \Delta_2)$, we define

$$d = d^{11} = \mathbb{I}\left((Y_{1i}, Y_{2i}) = (1,1)\right).$$

Hence, d generically represents the outcome variable depending on the parameter under consideration. Following this generic representation, denote the true parameter as $\boldsymbol{\theta}^0$ (equal to (α_1, α_2) or $(\alpha_1 + \Delta_1, \alpha_2 + \Delta_2)$) and $\hat{\boldsymbol{\theta}}_n$ as an estimator.

Now, for any two points $\boldsymbol{z} = (z_1, z_2)$ and $\tilde{\boldsymbol{z}} = (\tilde{z}_1, \tilde{z}_2)$, and for any arbitrary value of the parameter $\boldsymbol{\theta}$, define

$$\begin{aligned}
\varphi_1(\boldsymbol{z}, \tilde{\boldsymbol{z}}, \boldsymbol{\theta}) &= \varphi(z_1, z_2) = \Pr\left(d = 1 | \boldsymbol{Z}_1 = (z_1, z_2)\right); \\
\varphi_2(\boldsymbol{z}, \tilde{\boldsymbol{z}}, \boldsymbol{\theta}) &= \varphi(\tilde{z}_1, \tilde{z}_2) = \Pr\left(d = 1 | \boldsymbol{Z}_2 = (\tilde{z}_1, \tilde{z}_2)\right); \\
\varphi_3(\boldsymbol{z}, \tilde{\boldsymbol{z}}, \boldsymbol{\theta}) &= \varphi(z_1, \tilde{z}_2) = \Pr\left(d = 1 | \boldsymbol{Z}_3 = (z_1, \tilde{z}_2)\right); \\
\varphi_4(\boldsymbol{z}, \tilde{\boldsymbol{z}}, \boldsymbol{\theta}) &= \varphi(\tilde{z}_1, z_2) = \Pr\left(d = 1 | \boldsymbol{Z}_4 = (\tilde{z}_1, z_2)\right); \\
\varphi_5(\boldsymbol{z}, \tilde{\boldsymbol{z}}, \boldsymbol{\theta}) &= \varphi(2\theta_1 - z_1, 2\theta_2 - z_2) = \Pr\left(d = 1 | \boldsymbol{Z}_5 = (2\theta_1 - z_1, 2\theta_2 - z_2)\right); \\
\varphi_6(\boldsymbol{z}, \tilde{\boldsymbol{z}}, \boldsymbol{\theta}) &= \varphi(2\theta_1 - \tilde{z}_1, 2\theta_2 - \tilde{z}_2) = \Pr\left(d = 1 | \boldsymbol{Z}_6 = (2\theta_1 - \tilde{z}_1, 2\theta_2 - \tilde{z}_2)\right); \\
\varphi_7(\boldsymbol{z}, \tilde{\boldsymbol{z}}, \boldsymbol{\theta}) &= \varphi(2\theta_1 - z_1, 2\theta_2 - \tilde{z}_2) = \Pr\left(d = 1 | \boldsymbol{Z}_7 = (2\theta_1 - z_1, 2\theta_2 - \tilde{z}_2)\right); \\
\varphi_8(\boldsymbol{z}, \tilde{\boldsymbol{z}}, \boldsymbol{\theta}) &= \varphi(2\theta_1 - \tilde{z}_1, 2\theta_2 - z_2) = \Pr\left(d = 1 | \boldsymbol{Z}_8 = (2\theta_1 - \tilde{z}_1, 2\theta_2 - z_2)\right).
\end{aligned}$$

These eight φ functions correspond to the corners of the two rectangles in our identification analysis. Note that $\varphi_1(\boldsymbol{z}, \tilde{\boldsymbol{z}}, \boldsymbol{\theta})$, for example, does not depend on $\tilde{\boldsymbol{z}}$ or $\boldsymbol{\theta}$ since the corresponding corner is defined solely by $\boldsymbol{z}$. In another example, $\varphi_3(\boldsymbol{z}, \tilde{\boldsymbol{z}}, \boldsymbol{\theta})$ depends only elements in $\boldsymbol{z}, \tilde{\boldsymbol{z}}$, respectively. This is done for notational simplicity.

Next, consider

$$\sum_{\upsilon=1}^{8} \kappa_\upsilon \varphi_\upsilon(\boldsymbol{z}, \tilde{\boldsymbol{z}}, \boldsymbol{\theta}),$$

where $(\kappa_1, \ldots, \kappa_8) = (1, 1, -1, -1, -1, -1, 1, 1)$. Adding and subtracting the φ_υ terms in this way generates the difference in choice probabilities defined in two rectangles from the identification analysis. In other words,

$$\sum_{\upsilon=1}^{8} \kappa_\upsilon \varphi_\upsilon(\boldsymbol{z}, \tilde{\boldsymbol{z}}, \boldsymbol{\theta}) = B\left(\boldsymbol{z}, \tilde{\boldsymbol{z}}; \boldsymbol{\theta}^0\right) - B\left(2\boldsymbol{\theta} - \boldsymbol{z}, 2\boldsymbol{\theta} - \tilde{\boldsymbol{z}}; \boldsymbol{\theta}^0\right).$$

Hence, given our identification assumptions, $\sum_{\upsilon=1}^{8} \kappa_\upsilon \varphi_\upsilon(\boldsymbol{z}, \tilde{\boldsymbol{z}}, \boldsymbol{\theta}) = 0$ for all $\boldsymbol{z}, \tilde{\boldsymbol{z}}$ only when $\boldsymbol{\theta} = \boldsymbol{\theta}^0$ (see Lemmas I.1 and I.2).

To proceed to our asymptotic analysis, we propose a population objective function based on the above,

$$Q(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{Z}, \tilde{\boldsymbol{Z}}}\left[\tau\left(\boldsymbol{Z}, \tilde{\boldsymbol{Z}}, \boldsymbol{\theta}\right) \sum_{\upsilon=1}^{8} \kappa_\upsilon \varphi_\upsilon(\boldsymbol{Z}, \tilde{\boldsymbol{Z}}, \boldsymbol{\theta})\right]^2, \tag{1.4}$$

where the expectation is taken over all possible values $\boldsymbol{z}, \tilde{\boldsymbol{z}} \in \mathcal{S}_{\boldsymbol{Z}}^o$, where $\mathcal{S}_{\boldsymbol{Z}}^o$ denotes the interior of $\mathcal{S}_{\boldsymbol{Z}}$. Here, $Q(\boldsymbol{\theta})$ is similar to the quadratic objective function used in nonlinear least squares estimation.[19] $\tau(\boldsymbol{z}, \tilde{\boldsymbol{z}}, \boldsymbol{\theta})$ is a smooth trimming function which is positive on the interior of the compact set $\mathcal{S}_{\boldsymbol{Z}}$, $\mathcal{S}_{\boldsymbol{Z}}^o$, and zero otherwise (see more details in Assumption TR).[20] The trimming function ensures that we only evaluate the identifying restriction at the points on $\mathcal{S}_{\boldsymbol{Z}}^o$ and have symmetrically reflected points also contained in $\mathcal{S}_{\boldsymbol{Z}}^o$ for a given $\boldsymbol{\theta}$. Otherwise, $\sum_{\upsilon=1}^{8} \kappa_\upsilon \varphi_\upsilon(\boldsymbol{z}, \tilde{\boldsymbol{z}}, \boldsymbol{\theta})$ is not well defined.

Assumption TR Define $\tau_{ij}(\boldsymbol{\theta}) = \left(\Pi_{\upsilon=1}^{8} \tau(z_{\upsilon,1}, z_{\upsilon,2})\right)^{1/8}$, where for $\upsilon = 1, \ldots, 8$, write $(z_{\upsilon,1}, z_{\upsilon,2})$ as the generic points for the eight choice probabilities. The trimming

[19] Though other smooth or nonsmooth functional forms can also be used here, for analytical tractability, we will focus on this quadratic form in our analysis. We note that this quadratic function is not robust to outliers.

[20] In doing so, we prevent the estimator of the choice probabilities from the boundary bias. We will discuss more in Section 1.3.3.

function $\tau(z_{v,1}, z_{v,2}) : \mathcal{S}_{\boldsymbol{Z}}^o \to \mathbb{R}$ is bounded on the set $\mathcal{S}_{\boldsymbol{Z}}^o$ and equal to zero outside $\mathcal{S}_{\boldsymbol{Z}}^o$. In addition, the trimming function τ is at least ι times continuously differentiable and has bounded derivatives on $\mathcal{S}_{\boldsymbol{Z}}^o$.

Assumption TR specifies the properties of the smooth trimming function, which guarantees the identifying restriction is well-defined and we restrict it as the interior.[21] Assumption TR further guarantees that the corresponding kernel estimators of the choice probabilities have no boundary bias in the estimation later.

Theorem I.11. *Suppose that R, S, ER, RS and SV as well as TR hold. Then, (i) for all* $\boldsymbol{\theta} = (\theta_1, \theta_2) \in \Theta \subset \mathbb{R}^2$,$Q(\boldsymbol{\theta}) \geq 0$*; and (ii)* $Q(\boldsymbol{\theta}) = 0$ *when* $\boldsymbol{\theta} = \boldsymbol{\theta}^0$*, and* $Q(\boldsymbol{\theta}^*) > Q\left(\boldsymbol{\theta}^0\right) = 0$ *for all* $\boldsymbol{\theta}^* \neq \boldsymbol{\theta}^0$.

The proof is provided in Appendix A.3.1. Theorem I.11 shows that the population objective function is uniquely minimized at $\boldsymbol{\theta}^0$, implying that the true parameters can be identified from the population objective function.

By the analogy principle, let $Q_n(\boldsymbol{\theta})$ denote the sample analog of $Q(\boldsymbol{\theta})$. Replacing the expectation with a sample average and replacing the choice probabilities with corresponding kernel estimators, we obtain

$$Q_n(\boldsymbol{\theta}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \tau_{ij}(\boldsymbol{\theta}) \left[\sum_{v=1}^{8} \kappa_v \hat{\varphi}_v(\boldsymbol{Z}_i, \boldsymbol{Z}_j, \boldsymbol{\theta}) \right]^2.$$

In order to construct an estimator that converges at rate $n^{-1/2}$, we will need to use higher-order kernel functions.[22] Using higher-order kernel functions has the caveat that the predicted choice probabilities may be below zero or above one. Therefore, we consider the alternative sample objective function $\tilde{Q}_n(\boldsymbol{\theta})$ and define an estimator

[21] Note that the smooth trimming functions are typically assumed for analytical convenience. In practice, commonly trimming functions are specified with the combination of the smooth function and the indicator function.

[22] This is fairly common in the literature (e.g., Buchinsky and Hahn (1998)).

$\hat{\boldsymbol{\theta}}_n$ as

$$\begin{aligned}\tilde{Q}_n(\boldsymbol{\theta}) &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\tau_{ij}(\boldsymbol{\theta})\,G_{ij}(\boldsymbol{\theta})\left[\sum_{\upsilon=1}^{8}\kappa_\upsilon\hat{\varphi}_\upsilon(\boldsymbol{Z}_i,\boldsymbol{Z}_j,\boldsymbol{\theta})\right]^2; \quad (1.5)\\ \text{and } \hat{\boldsymbol{\theta}}_n &= \arg\min_{\boldsymbol{\theta}}\tilde{Q}_n(\boldsymbol{\theta})\end{aligned}$$

where $\tau_{ij}(\boldsymbol{\theta})$ restricts the observed variable taking the values in the interior of $\mathcal{S}_{\boldsymbol{Z}}$, $\mathcal{S}^o_{\boldsymbol{Z}}$, to protect against the boundary bias in the estimated choice probabilities; and G_{ij} is a trimming function that ensures that the predicted choice probabilities are well-defined (the choice probability is not below zero or above one). Additional details for the trimming function can be found in Appendix A.3.

In the objective function $\tilde{Q}_n(\boldsymbol{\theta})$, we follow standard procedure for nonparametrically estimating the choice probabilities $\hat{\varphi}_\upsilon$ for $\upsilon = 1,\ldots,8$. For example, for $\upsilon = 3$, our kernel estimator is

$$\begin{aligned}\hat{\varphi}_{n,3}\left(\boldsymbol{z_i},\boldsymbol{z_j},\boldsymbol{\theta}\right) &= \hat{\varphi}_n(z_{1i},z_{2j}) = \frac{\hat{g}_n(z_{1i},z_{2j})}{\hat{f}_n(z_{1i},z_{2j})},\\ \hat{g}_n(z_{1i},z_{2j}) &= \frac{1}{n-2}\sum_{k=1,k\neq i\neq j}^{n} d_k K_n\left(\frac{Z_{1k}-z_{1i}}{h},\frac{Z_{2k}-z_{2j}}{h}\right),\\ \hat{f}_n(z_{1i},z_{2j}) &= \frac{1}{n-2}\sum_{k=1,k\neq i\neq j}^{n} K_n\left(\frac{Z_{1k}-z_{1i}}{h},\frac{Z_{2k}-z_{2j}}{h}\right);\end{aligned}$$

where $K_n(\boldsymbol{u}) = \frac{1}{h^2}K\left(\Omega_t^{-1}(\boldsymbol{u})\right)$ is a kernel function depending on the covariance matrix Ω_t and the bandwidth $h = h_n$, which is defined as a decreasing function of n. Other terms $\hat{\varphi}_\upsilon$ are constructed in the same way. In summary, our estimation strategy follows a plug-in two-step procedure, i.e., we first nonparametrically estimate the choice probabilities, and then use these choice probabilities to evaluate the objective function given the parameters. We search for the parameter values to minimize this objective function.

The rest of this section proceeds as follows: we first show the properties of the

proposed estimator in Section 1.3.1. Then, we derive an optimal bandwidth and a feasible estimator for standard error of $\hat{\boldsymbol{\theta}}_n$ in Section 1.3.2. Finally in Section 1.3.3, we discuss the practical issues one might encounter when choosing the kernel function, bandwidth selection and trimming.

1.3.1 Properties of Estimator

In this section, we will derive the following asymptotic and finite-sample properties of the proposed estimator: consistency, rate of convergence and asymptotic normality as well as higher-order mean squared error approximation. Throughout this section, we assume that the following additional regularity conditions hold.

Assumption 1 Θ is a compact subset of $\mathbb{R}^2$; $\boldsymbol{\theta}^0 \in \text{int}(\Theta)$.

Assumption 1 is standard in the literature. The compactness condition is always required for consistency, while the restriction to the interior of the parameter space is required only for asymptotic normality. The compactness condition is commonly used in the literature of discrete choice models (e.g., Manski (1985), Horowitz (1992), and Ichimura (1993)).

Assumption 2 Assume that

(i) The random vector (Z_1, Z_2) is continuously distributed on a compact support $\mathcal{S}_{\boldsymbol{Z}}$, with the joint density $f_{(Z_1,Z_2)}(\cdot,\cdot)$. Further, $f_{(Z_1,Z_2)}(\cdot,\cdot)$ is bounded away from zero by some positive constant over its support and $C_f = \sup_{z_1,z_2} f(z_1,z_2) < \infty$.

(ii) The marginal densities $f_{Z_1}(\cdot)$ and $f_{Z_2}(\cdot)$ and the joint density $f_{(Z_1,Z_2)}(\cdot,\cdot)$, as well as the product $\varphi(\cdot,\cdot)f_{Z_1}(\cdot), \varphi(\cdot,\cdot)f_{Z_2}(\cdot)$ and $\varphi(\cdot,\cdot)f_{(Z_1,Z_2)}(\cdot,\cdot)$ are at least ι times continuously differentiable and have bounded derivatives on the sets $\mathcal{S}^o_{Z_1}$, $\mathcal{S}^o_{Z_2}$ and $\mathcal{S}^o_{\boldsymbol{Z}}$, where $\mathcal{S}^o_{Z_1}$, $\mathcal{S}^o_{Z_2}$ denotes as the interior of $\mathcal{S}_{Z_1}$, $\mathcal{S}_{Z_2}$.

(iii) $\mathbb{E}_{[i,j]}\left[\nabla_\theta \zeta_{ij}\left(\boldsymbol{\theta}^0\right)\varphi(\cdot,\cdot)^{(\iota_1)} f_{(Z_1,Z_2)}(\cdot,\cdot)^{(\iota_2)}\right]$ exists for $\iota_1 + \iota_2 = \iota, 0 < \iota_1, \iota_2 \leq \iota$.

In addition, $\mathbb{E}_{[i,j]}\left[\nabla_\theta f_{\upsilon,ij}\left(\cdot,\cdot\right)/f_{\upsilon,ij}^{-2}\left(\cdot,\cdot\right)\right]$ exists for all $\upsilon = 1,\ldots,8$.

Assumption 2-(i) avoids zero denominator problems. Assumption 2-(ii) imposes smoothness conditions on the unknown densities and the conditional choice probabilities, which are standard in kernel regression estimation. Assumption 2-(ii) also requires that the bounded derivatives exist in the interior of the support, in order to prevent our estimator from suffering from a boundary bias. Note that the higher the differentiability ι is, the higher the kernel function order one can use, and the smaller the bias is. We will save the detailed discussion on this point for Section 1.3.3. Assumption 2-(iii) assumes that the first-order and higher-order means exist for the Hoeffiding decomposition. In addition, it guarantees that the first highest-order terms that involve the bandwidth exist. Alternatively, if we follow the derivation of density-weighted average derivative in Powell, Stock, and Stoker (1989) and Powell and Stoker (1996), by adding the product of corresponding densities as a weight in the objective function, then Assumption 2-(iii) is not necessary. However, it will substantially increase complicity in our derivation. To keep the derivation simple, we impose Assumption 2-(iii) here.

Assumption 3 The kernel function K and the bandwidth h satisfy the following conditions:

(i) the bivariate kernel function K is a function of bounded variation that satisfies

(a) $K\left(\mathbf{u}\right) = K\left(-\mathbf{u}\right)$;

(b) $\left|K\left(\mathbf{u}\right)\right| \leq \bar{K} < \infty$ and $\int_{\mathbb{R}^2}\left|K\left(\mathbf{u}\right)\right| d\mathbf{u} \leq c < \infty$;

(c) For some $\iota \geq 2$,

$$\int_{\mathbb{R}^2} u_1^{\iota_1} u_2^{\iota_2} K\left(\mathbf{u}\right) d\mathbf{u} \begin{cases} = 1 & \text{if} \quad \iota_1 + \iota_2 = 0, \\ = 0 & \text{if} \quad 0 < \iota_1 + \iota_2 < \iota, \\ < \infty & \text{if} \quad \iota_1 + \iota_2 = \iota; \end{cases}$$

(d) For some $C < \infty$, $K(\mathbf{u}) = 0$ for $\|\mathbf{u}\| > C$ and for all $\mathbf{u}, \mathbf{u}' \in \mathbb{R}^2$,

$$|K(\mathbf{u}') - K(\mathbf{u})| \leq K^*(\mathbf{u}) \|\mathbf{u}' - \mathbf{u}\|;$$

and

$$C_K = \sup_{\mathbf{u} \in \mathbb{R}^2} (\mathbf{u}) + \int_{\mathbb{R}^2} K(\mathbf{u}) \, d\mathbf{u} \text{ and } C_{K^*} = \sup_{\mathbf{u} \in \mathbb{R}^2} (\mathbf{u}) + \int_{\mathbb{R}^2} K^*(\mathbf{u}) \, d\mathbf{u}.$$

(ii) h is a sequence of positive numbers that satisfies $h \to 0$ as $n \to \infty$.

In Assumption 3-(i), Conditions (a)-(c) are standard in the literature. Condition-(d) corresponds to Assumption 3 in Hansen (2008) and to Assumption 2-(d) in Cattaneo, Crump, and Jansson (2013). Since we allow for a relaxed support condition on the observables, we need our kernel function to both have truncated support and satisfy the Lipschitz condition in order to show uniform convergence. More details on the kernel functions that satisfy this condition are given in Section 1.3.3.

Assumption 3-(ii) does not impose any specific restriction on the rate at which h will decrease as the sample size n increases. For analytical simplicity, we assume that the bandwidths are the same across different dimensions of each kernel regression estimator and are the same for different kernel regression estimators.

Assumption 4 The trimming function $G(\cdot)$ is $(L+1)$th order differentiable for some $L > 4$.

Assumption 4 is required when we use a higher-order kernel function. This condition ensures that the kernel regression estimator of the choice probability is well-defined, that is, the estimated choice probability is not below zero or above one. In Appendix A.3, we follow Linton and Xiao (2001) when specifying the form of the trimming function. The smooth trimming of $G(\cdot)$, in particular, guarantees that the trimming does not affect the higher-order mean squared error (MSE) approximation,

in comparison to an indicator trimming function. We will discuss in greater detail on the difference between the smooth trimming function and the indicator trimming function on the higher-order MSE approximation in Section 1.3.3.

1.3.1.1 Consistency

To prove that the estimator is consistent, we will first show the uniform convergence of the densities and the conditional choice probabilities.

Lemma I.12. *Suppose that Assumptions 2-3 hold. Then, for* $\upsilon = 1, \cdots, 8$*, with* $\boldsymbol{z}_\upsilon = (z_{\upsilon,1}, z_{\upsilon,2})$*,*

$$
\begin{array}{ll}
(i) & \sup\limits_{\boldsymbol{z}_\upsilon \in \mathcal{S}^o_{\boldsymbol{Z}}} \left| \hat{f}_n \left(z_{\upsilon,1}, z_{\upsilon,2} \right) - \mathbb{E} \left[\hat{f}_n \left(z_{\upsilon,1}, z_{\upsilon,2} \right) \right] \right| = O_p \left(\sqrt{\frac{\log n}{nh^2}} \right) \\
 & \sup\limits_{\boldsymbol{z}_\upsilon \in \mathcal{S}^o_{\boldsymbol{Z}}} \left| \mathbb{E} \left[\hat{f}_n \left(z_{\upsilon,1}, \mu_{\upsilon,2} \right) \right] - f_n \left(z_{\upsilon,1}, z_{\upsilon,2} \right) \right| = O \left(h^\iota \right); \\
(ii) & \sup\limits_{\boldsymbol{z}_\upsilon \in \mathcal{S}^o_{\boldsymbol{Z}}} \left| \hat{g}_n \left(z_{\upsilon,1}, z_{\upsilon,2} \right) - \mathbb{E} \left[\hat{g}_n \left(z_{\upsilon,1}, z_{\upsilon,2} \right) \right] \right| = O_p \left(\sqrt{\frac{\log n}{nh^2}} \right) \\
 & \sup\limits_{\boldsymbol{z}_\upsilon \in \mathcal{S}^o_{\boldsymbol{Z}}} \left| \mathbb{E} \left[\hat{g}_n \left(z_{\upsilon,1}, z_{\upsilon,2} \right) \right] - g_n \left(z_{\upsilon,1}, z_{\upsilon,2} \right) \right| = O \left(h^\iota \right).
\end{array}
$$

Our proof of Lemma I.12, found in the Appendix S.A, follows Newey (1994), Hansen (2008) and Cattaneo, Crump, and Jansson (2013). Using Lemma I.12, we can now show the uniform convergence of the choice probability of $\hat{\varphi}_\upsilon$, for $\upsilon = 1, \cdots, 8$, as follows.

Lemma I.13. *Suppose that Assumptions 2-3 hold. Then, for* $\upsilon = 1, \cdots, 8$*, with* $\boldsymbol{z}_\upsilon = (z_{\upsilon,1}, z_{\upsilon,2})$*,*

$$
\begin{array}{rcl}
\sup\limits_{\boldsymbol{z}_\upsilon \in \mathcal{S}^o_{\boldsymbol{Z}}} \left| \hat{\varphi}_n \left(z_{\upsilon,1}, z_{\upsilon,2} \right) - \mathbb{E} \left[\hat{\varphi}_n \left(z_{\upsilon,1}, z_{\upsilon,2} \right) \right] \right| & = & O_p \left(\sqrt{\frac{\log n}{nh^2}} \right) \\
\sup\limits_{\boldsymbol{z}_\upsilon \in \mathcal{S}^o_{\boldsymbol{Z}}} \left| \mathbb{E} \left[\hat{\varphi}_n \left(z_{\upsilon,1}, z_{\upsilon,2} \right) \right] - \varphi_n \left(z_{\upsilon,1}, z_{\upsilon,2} \right) \right| & = & O \left(h^\iota \right)).
\end{array}
$$

Lemma I.13 shows that the optimal uniform rate of convergence of the estimated choice probability is $(n/\log n)^{\frac{\iota}{2\iota+2}}$, and the corresponding bandwidth is of

order $(n/\log n)^{\frac{\iota}{2\iota+2}}$. In addition, this lemma will be used to show the uniform convergence of the sample objective function, which is the key to showing the consistency of the estimator.

Theorem I.14. *Suppose that Assumptions R, S, ER, RS, SV, TR and Assumptions 1-4 hold. Then, provided that* $nh^2/\log n \to \infty$, *we have* $\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0 = o_p(1)$.

Theorem I.14 gives the consistency of the estimator. We note that the optimal bandwidth of the choice probability satisfies the requirement $nh^2/\log n \to \infty$. As such, an estimator using the optimal bandwidth will be consistent. Theorem I.14 allows us to construct a consistent estimator for the plug-in bandwidth selector.

1.3.1.2 Root–n Consistency and Asymptotic Normality

In this subsection, we show that the proposed estimator is $\sqrt{n}$-consistent and asymptotically normal following Sherman (1994). First, we will apply Theorem 1 in Sherman (1994) to show the $\sqrt{n}$-consistency.

Theorem I.15. *Suppose that Assumptions R, S, ER, RS, SV, TR and Assumptions 1-4 hold. Then, provided that* $nh^{2\iota} \to 0$ *and* $nh^4 \to \infty$, *we have* $\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0 = O_p\left(n^{-1/2}\right)$.

Theorem I.15 gives the $\sqrt{n}$-consistency result. Central to this result is that the leading term of the estimator takes the form of a U-statistic, similar to the full mean of Newey (1994) and Imbens and Ridder (2009). And, in a broader sense, it is also similar to the kernel-based average derivatives of Powell, Stock, and Stoker (1989). They find that such U-statistic can be written as an average over the plug-in nonparametric estimator. Similarly, our estimator is a U-statistic, as it can be written as an average over the plug-in kernel regression estimator of the choice probabilities.

Given this form, Newey (1994) and Imbens and Ridder (2009) further show that the rate of convergence is determined by the dimensions of components of covariates

that are averaged over: the more components are averaged out, the faster the convergence rate is. Following their discussion, we find that in our context, the leading term in our U-statistic averages out entire components of covariates. In this way, we can achieve the $n^{-1/2}$ rate of convergence.

In addition, in this theorem, we impose conditions on the bandwidth sequence. Two bandwidth conditions are used to control both the first-order and higher-order biases, respectively, in order to guarantee $\sqrt{n}$-consistency of the estimator, where we will show the form of the bias terms in Section 1.3.1.3. Specifically, we require that the order of the kernel function be greater than two, in order to satisfy $nh^{2\iota} \to 0$ and $nh^4 \to \infty$. In other words, we need to use higher-order kernel functions in estimation. Admittedly, the need to use a higher-order kernel function is a limitation of this estimation procedure, in the sense that it requires that the underlying distribution function have additional smoothness. Moreover, the higher-order kernel has negative components, which can lead to kernel regression estimates of the choice probabilities that are below zero or above one. To avoid this, we further impose an additional trimming function when estimating the choice probabilities, as stated in Assumption 4.

Despite the cost of using the higher-order kernel function, the bandwidth choice allows the estimation error of the nonparametric plug-in kernel regression estimator to have order $o_p\left(n^{-1/6}\right)$ in a suitable norm, similar to Cattaneo, Crump, and Jansson (2013). This result is weaker than the commonly used requirement in the literature while not invalidating the asymptotic linearity and asymptotic normality of our estimator, as shown below.

Next, we follow Theorem 2 in Sherman (1994) to show asymptotic linearity and asymptotic normality as follows.

Theorem I.16. *Suppose that Assumptions R, S, ER, RS, SV, TR and Assumptions 1-4 hold. Then, provided that* $nh^{2\iota} \to 0$ *and* $nh^4 \to \infty$*, (i) (Asymptotic Linearity)*

The estimator is asymptotically linear with

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0\right) = n^{-1/2}\sum_{k=1}^{n}\Gamma^{-1}\psi_k + o_p\left(1\right);$$

(ii) (Asymptotic Normality) The estimator is asymptotically normal

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0\right) \xrightarrow{d} N\left(0, \Gamma^{-1}\Sigma\Gamma^{-1\prime}\right)$$

where $\Gamma = \mathbb{E}\left[\nabla_\theta\zeta\left(\boldsymbol{z}_i, \boldsymbol{z}_j, \boldsymbol{\theta}^0\right)\nabla_\theta\zeta\left(\boldsymbol{z}_i, \boldsymbol{z}_j, \boldsymbol{\theta}^0\right)\right]$ *with*

$$\nabla_\theta\zeta\left(\boldsymbol{z}_i, \boldsymbol{z}_j, \boldsymbol{\theta}^0\right) = \tau_{ij}\left(\boldsymbol{\theta}^0\right)\left(\sum_{\upsilon=5}^{8}\kappa_\upsilon\nabla_\theta\varphi_{\upsilon,ij}\left(\boldsymbol{\theta}^0\right)\right);$$

and $\psi_k = 2\sum_{\upsilon=1}^{8}\kappa_\upsilon\left(\nabla_\theta\xi_{n,\upsilon} - \mathbb{E}\nabla_\theta\xi_{n,\upsilon}\right)$ *with*[23]

$$\nabla_\theta\xi_{n,\upsilon k} = \begin{cases} \left(d_k - \varphi\left(z_{1k}, z_{2k}\right)\right)\int\int\nabla_\theta\zeta\left(\boldsymbol{z}_k, \boldsymbol{s}, \boldsymbol{\theta}^0\right)f\left(s_1, s_2\right)d\left(s_1, s_2\right), & \upsilon = 1; \\ \left(d_k - \varphi\left(z_{1k}, z_{2k}\right)\right)\int\int\nabla_\theta\zeta\left(z_{1k}, r_2, s_1, z_{2k}, \boldsymbol{\theta}^0\right)\frac{f(z_{1k}, r_2)f(s_1, z_{2k})}{f(z_{1k}, z_{2k})}d\left(r_2, s_1\right), & \upsilon = 3; \\ \left(d_k - \varphi\left(z_{1k}, z_{2k}\right)\right)\int\int\nabla_\theta\zeta\left(2\boldsymbol{\theta}^0 - \boldsymbol{z}_k, \boldsymbol{s}, \boldsymbol{\theta}^0\right)f\left(s_1, s_2\right)d\left(s_1, s_2\right) & \upsilon = 5; \\ \left(d_k - \varphi\left(z_{1k}, z_{2k}\right)\right)\int\int\nabla_\theta\zeta\left(\left(2\theta_1^0 - z_{1k}\right), r_2, s_1, \left(2\theta_2^0 - z_{2k}\right), \boldsymbol{\theta}^0\right) & \upsilon = 7; \\ \times\frac{f\left(\left(2\theta_1^0 - z_{1k}\right), r_2\right)f\left(s_1, \left(2\theta_2^0 - z_{2k}\right)\right)}{f(z_{1k}, z_{2k})}d\left(r_2, s_1\right), \end{cases}$$

and $n^{-1/2}\sum_{k=1}^{n}\psi_k \to^d N\left(0, \Sigma\right)$, *where* $\Sigma = \mathbb{E}\left[\psi_k\psi_k'\right]$.

Theorem I.16 provides asymptotic linearity and asymptotic normality results, under the bandwidth conditions the same as those in Theorem I.15. This theorem shows that ψ_k is a linear combination over the different values of $\nabla_\theta\xi_{n,\upsilon k}$ for $\upsilon = 1, \ldots, 8$. Recall that our estimation procedure begins with a pair of markets 1 and 2. From these markets, we are able to construct two artificial markets, 3 and 4, by taking the characteristic for player 1 from a market and combining it with the characteristic

[23]Due to the symmetry of the indices i and j, we only represent the cases with $\upsilon = 1, 3, 5, 7$. The other remaining cases take a similar form. To save space, we do not explicitly provide the expressions of these four remaining cases.

for player 2 from the other market. We form the market $\upsilon = 5, 6, 7, 8$ by taking the values for each player reflected through the symmetric points.

When $\upsilon = 1, 2, 5, 6$, the individual U-statistics are similar to the standard full mean case of Newey (1994) and Imbens and Ridder (2009), which can be treated as the sample average of the kernel regression estimates over the all components of observables. In addition, when $\upsilon = 3, 4, 7, 8$, the individual U-statistics, instead, depart slightly from the standard full mean form, which can be treated as the double average over each dimension separately of the kernel regression estimator but across all dimensions of the observables as well, which can also lead to the $n^{-1/2}$ rate. Therefore, these two types of U-statistics with the rate $n^{-1/2}$ ensure that the final parameter estimator has the rate $n^{-1/2}$, as well.

Theorem I.16 also suggests an analytical expression for the standard error. Note that the variance of the estimator contains the matrix Σ, which can be drawn from joint distribution of the vector $(\nabla_\theta \xi_{n,1}, \ldots, \nabla_\theta \xi_{n,8})'$; that is,

$$\begin{pmatrix} \nabla_\theta \xi_{n,1} \\ \vdots \\ \nabla_\theta \xi_{n,8} \end{pmatrix} \sim N \left(\begin{pmatrix} \mathbb{E} \nabla_\theta \xi_{n,1} \\ \vdots \\ \mathbb{E} \nabla_\theta \xi_{n,8} \end{pmatrix}, \begin{pmatrix} \mathcal{V}_{11,k} & \cdots & \mathcal{V}_{18,k} \\ \vdots & \ddots & \vdots \\ \mathcal{V}_{81,k} & \cdots & \mathcal{V}_{88,k} \end{pmatrix} \right).$$

As we show in Appendix A.3.3, for all $\upsilon \neq \upsilon'$, $\nabla_\theta \xi_{n,\upsilon}$ and $\nabla_\theta \xi_{n,\upsilon'}$ are correlated, and all the off-diagonal covariance terms are of order $O\left(n^{-1}\right)$. This structure immediately suggests that the variance of the estimator contains both variance and covariance components of this vector, and to calculate the standard error of our estimator, we need to calculate each element in this variance-covariance matrix. The kernel estimator for each element can be written as

$$\mathcal{V}_{\upsilon\upsilon} = \sigma^2 \mathbb{E}\left[\chi_{\upsilon,k} \chi'_{\upsilon,k}\right] \text{ and } \mathcal{V}_{\upsilon\upsilon'} = \sigma^2 \mathbb{E}\left[\chi_{\upsilon,k} \chi'_{\upsilon',k}\right],$$

where $\chi_{\upsilon,k}$ denotes the integral part of $\nabla_\theta \xi_{n,\upsilon k}$. For example, when $\upsilon = 1$, we have $\chi_{1,k} = \int\int \nabla_\theta \zeta\left(\boldsymbol{z}_k, \boldsymbol{s}, \boldsymbol{\theta}^0\right) dF\left(s_1, s_2\right)$. In addition, we complete the analysis by emphasizing that Γ is a linear combination of the first derivative of the choice probability, which we estimate using the kernel estimator of the derivatives of the choice probability. We will use these formulas to construct a consistent estimator of the standard error.

1.3.1.3 Higher-order MSE Approximation

In this subsection, we provide a mean squared error (MSE) expansion of the estimator $\hat{\boldsymbol{\theta}}_n$ in order to derive the plug-in "optimal" bandwidth selector. The expansion procedure is consistent with the asymptotic results shown in Appendix A.3.

Theorem I.17. *Suppose that Assumptions R, S, ER, RS and SV and Assumptions 1-4 hold. Then, the approximate MSE of* $\bar{\Gamma}_n\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0\right)$ *is given by*

$$\frac{1}{n}\Sigma + h^{2\iota}\mathcal{B}\mathcal{B}' + \frac{1}{n^2 h^4}\mathcal{B}^h\mathcal{B}^{h\prime}, \tag{1.6}$$

where

$$\begin{aligned}\bar{\Gamma}_n &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\left[\tau_{ij}\left(\boldsymbol{\theta}\right)\sum_{\upsilon=1}^{8}\kappa_\upsilon \nabla_\theta \varphi_\upsilon\left(\boldsymbol{Z}_i, \boldsymbol{Z}_j, \boldsymbol{\theta}\right)\right] \\ &\quad \times\left[\tau_{ij}\left(\boldsymbol{\theta}\right)\sum_{\upsilon=1}^{8}\kappa_\upsilon \nabla_\theta \varphi_\upsilon\left(\boldsymbol{Z}_i, \boldsymbol{Z}_j, \boldsymbol{\theta}\right)\right]';\end{aligned}$$

with $\mathcal{B} = \sum_{\upsilon=1}^{8}\kappa_\upsilon \mathcal{B}_\upsilon$ *and* $\mathcal{B}^h = \sum_{\upsilon=1}^{8}\kappa_\upsilon \mathcal{B}_\upsilon^h$, *where*

$$\begin{aligned}\mathcal{B}_\upsilon &= \mathbb{E}_{[i,j]}\left[\nabla_\theta \zeta_{ij}\left(\boldsymbol{\theta}^0\right)\sum_{\iota_1+\iota_2=\iota, 0<\iota_1,\iota_2\leq\iota}\left[\int\int u_1^{\iota_1}u_2^{\iota_2}K\left(\mathbf{u}\right)d\mathbf{u}\right]\vartheta_{\upsilon,\iota_1,\iota_2}\left(\cdot,\cdot\right)\right], \\ \mathcal{B}_\upsilon^h &= \sigma_\upsilon^2\mathbb{E}_{[i,j]}\left[-\frac{\nabla_\theta f_{\upsilon,ij}\left(\boldsymbol{\theta}^0\right)}{f_{\upsilon,ij}^{-2}\left(\boldsymbol{\theta}^0\right)}\right]\int\int K^2\left(\mathbf{u}\right)d\mathbf{u};\end{aligned}$$

and $\mathbb{E}_{[i,j]}$ *is the expectation taken over* i *and* j *and* $\vartheta_{\upsilon,\iota_1,\iota_2}$ *are the corresponding bias components.*

Whereas Theorem I.16 discusses the first-order asymptotics, Theorem I.17 provides the higher-order MSE approximation. The first term in equation (1.6) corresponds to the variance matrix of the estimators; the second term $\mathcal{B}$ corresponds to the first-order bias; and the third term corresponds to the higher-order bias $\mathcal{B}^h$ where its element $\mathcal{B}^h_\upsilon$ takes a form similar to the variance of the kernel regression estimator for the choice probability.

Theorem I.17 not only verifies the asymptotic linear representation of the estimator, but also suggests a way of selecting an optimal bandwidth h^*. Specifically, we define the optimal bandwidth selector h^* as the one that minimizes the second and third terms in equation (1.6). To describe this bandwidth selector, let $c = (1,1)' \in \mathbb{R}^2$. Then, the optimal bandwidth selector can be defined as,

$$h^* = \left(\frac{4\left(c'\mathcal{B}^h\right)^2}{2\iota\left(c'\mathcal{B}\right)^2 n^2} \right)^{1/2\iota+4} = C_h n^{-1/\iota+2}; \tag{1.7}$$

where C_h is a constant. The last expression implies that the bandwidth selector is proportional to $n^{-1/\iota+2}$, where ι is the order of a kernel function and also the smoothness of the underlying distribution. For example, if we use the fourth-order kernel function, that is, $\iota = 4$, the bandwidth selector is proportional to $n^{-1/6}$. This expression also suggests that we might be able to construct a consistent estimator of h^* if consistent estimators of $\mathcal{B}$ and $\mathcal{B}^h$ are available. Consistent estimators $\hat{\mathcal{B}}$ and $\hat{\mathcal{B}}^h$ can be derived for any arbitrary kernel function and some bandwidth $h, h \to 0$. We provide the derivation of these consistent estimators in Section 1.3.2.

Note that at the beginning of Section 1.3, we explicitly assume that the bandwidths are the same across two players. In addition, we assume that the bandwidths are the same across different kernel regression estimators of the choice probabilities.

This is done largely due to analytical convenience. In practice, we can easily allow the bandwidth to vary with the identities of players following the same MSE approximation above. In the same way, due to the linearity of the choice probabilities in the criterion function, we can also allow the bandwidth to vary across kernel regression estimators of the different choice probabilities. For brevity, we omit the derivation for these extensions.

1.3.2 Feasible Estimators

The previous subsection provides the asymptotic properties and the MSE of the estimator. To implement the plug-in bandwidth selector, we need to construct a consistent estimator of the constant terms, $\mathcal{B}$ and $\mathcal{B}^h$, for the optimal bandwidth h^* in (1.7). In addition, to draw inference from this estimator, we need to obtain consistent estimators of the variance for the model parameters.

In order to estimate the constant, we first choose an arbitrary value for it and obtain estimates of the model parameters, $\tilde{\boldsymbol{\theta}}_I$ using (1.5). Then, using $\tilde{\boldsymbol{\theta}}_I$, we construct a consistent estimator for the first order bias $\mathcal{B}$ and the higher-order bias $\mathcal{B}^h$. Note that $\mathcal{B} = \sum_{\upsilon=1}^{8} \kappa_\upsilon \mathcal{B}_\upsilon$ and $\mathcal{B}^h = \sum_{\upsilon=1}^{8} \kappa_\upsilon \mathcal{B}^h_\upsilon$, where explicit expressions for $\mathcal{B}_\upsilon$ and $\mathcal{B}^h_\upsilon$ can be found in Appendix S.C. For example, when $\upsilon = 5$, a plug-in estimator of the first-order bias, $\mathcal{B}_5$, is given by

$$\hat{\mathcal{B}}_5 = \hat{\Gamma}^{-1} \binom{n}{2}^{-1} \sum_{i \neq j} \left[\nabla_\theta \hat{\zeta}_{ij} \left(\tilde{\boldsymbol{\theta}}_I \right) \sum_{\iota_1 + \iota_2 = \iota, 0 < \iota_1, \iota_2 \leq \iota} \left[\int \int u_1^{\iota_1} u_2^{\iota_2} k \left(\mathbf{u} \right) d\mathbf{u} \right] \hat{\vartheta}_{5, \iota_1, \iota_2} \left(Z_{1i}, Z_{2i} \right) \right],$$

where $\sum_{i \neq j} = \sum_{i=1}^{n} \sum_{j=i+1}^{n}$. In addition, we can derive

$$\begin{aligned} \hat{\Gamma} &= \binom{n}{2}^{-1} \sum_{i \neq j} \left[\nabla_\theta \hat{\zeta}_{ij} \left(\tilde{\boldsymbol{\theta}}_I \right) \right] \left[\nabla_\theta \hat{\zeta}_{ij} \left(\tilde{\boldsymbol{\theta}}_I \right) \right]' \qquad (1.8) \\ \nabla_\theta \hat{\zeta}_{ij} \left(\tilde{\boldsymbol{\theta}}_I \right) &= \tau_{ij} \left(\tilde{\boldsymbol{\theta}}_I \right) \sum_{\upsilon=5}^{8} \kappa_\upsilon \nabla_\theta \hat{\varphi}_{\upsilon, ij} \left(\tilde{\boldsymbol{\theta}}_I \right) \end{aligned}$$

and $\nabla_\theta \hat{\varphi}_{\upsilon,ij}\left(\tilde{\boldsymbol{\theta}}_I\right) = \hat{f}_{\upsilon,ij}^{-1}\left(\tilde{\boldsymbol{\theta}}_I\right)\left(\nabla_\theta \hat{g}_{\upsilon,ij}\left(\tilde{\boldsymbol{\theta}}_I\right) - \hat{\varphi}_{\upsilon,ij}\left(\tilde{\boldsymbol{\theta}}_I\right)\nabla_\theta \hat{f}_{\upsilon,ij}\left(\tilde{\boldsymbol{\theta}}_I\right)\right)$ using the fourth-order kernel function and the bandwidth taking the form of $b = C_b n^{-1/2\iota+2+s}$, where s is the order of derivative. Furthermore, $\hat{\vartheta}_{5,\iota_1,\iota_2}(z_{1i}, z_{2i})$ can be derived accordingly following Taylor's expansion[24].

In addition, a plug-in estimator of the higher-order bias, $\mathcal{B}_5^h$, can be written as

$$\mathcal{B}_5^h = -\hat{\Gamma}^{-1}\hat{\sigma}_5^2\binom{n}{2}^{-1}\sum_{i\neq j}\frac{\nabla_\theta \hat{f}_{5,ij}\left(\tilde{\boldsymbol{\theta}}_I\right)}{\hat{f}_{5,ij}^{-2}\left(\tilde{\boldsymbol{\theta}}_I\right)}\hat{\mu}_{k^2},$$

where $\hat{\sigma}_\upsilon^2 = \frac{1}{n}\sum_{k=1}^n \hat{v}_{\upsilon,k}^2$. Also, $\hat{f}_{\upsilon,ij}\left(\tilde{\boldsymbol{\theta}}_I\right)$ and $\nabla_\theta \hat{f}_{\upsilon,ij}\left(\tilde{\boldsymbol{\theta}}_I\right)$ are the standard kernel estimators of the density and the derivative of the density, respectively. Here, we denote $\hat{\mu}_{k^2}$ as $\int\int k^2(\mathbf{u})\,d\mathbf{u}$, for some specific kernel functions k. The value of $\hat{\mu}_{k^2}$ varies across the kernel functions we choose. Having obtained estimators of $\mathcal{B}$ and $\mathcal{B}^h$, we construct an estimator for C_h as $\left(\frac{4\left(c'\mathcal{B}^h\right)^2}{2\iota(c'\mathcal{B})^2}\right)^{1/2\iota+4}$.

Next, with this estimator for C_h, we can construct an estimator for the standard error. Using the optimal bandwidth derived above, we can reestimate the model parameters $\hat{\boldsymbol{\theta}}$ using (1.5). With $\hat{\boldsymbol{\theta}}$, we construct a consistent estimator for the variance $\mathbb{V}(\hat{\boldsymbol{\theta}})$. We observe that the variance expression contains three components: the Hessian matrix Γ, the variance of each random component $\mathbb{V}[\nabla_\theta \xi_{n,\upsilon}]$, as well as the covariance $Cov[\nabla_\theta \xi_{n,\upsilon}\nabla_\theta \xi_{n,\upsilon'}]$, for $\upsilon \neq \upsilon'$. Consequently, as long as we can construct a consistent estimator of each component, we can derive a consistent estimator of the variance term. Natural candidates for estimates for each component are their kernel

[24]For example, for $\iota_1 = 4, \iota_2 = 0$,

$$\begin{aligned}\hat{\vartheta}_{5,\iota_1,\iota_2}(z_{1i}, z_{1i}) &= \partial^{(4)}\varphi(2\theta_1 - z_{1i}, 2\theta_2 - z_{2i})\, f(2\theta_1 - z_{1i}, 2\theta_2 - z_{2i})/\partial\theta_1^4 \\ &\quad + \partial^{(4)}\varphi(2\theta_1 - z_{1i}, 2\theta_2 - z_{2i})\, f(2\theta_1 - z_{1i}, 2\theta_2 - z_{2i})/\partial\theta_2^4.\end{aligned}$$

estimators, defined as follows:

$$\begin{aligned}
\hat{\Gamma} &= \binom{n}{2}^{-1} \sum_{i \neq j} \left[\nabla_\theta \hat{\zeta}_{ij}\left(\hat{\boldsymbol{\theta}}\right)\right]\left[\nabla_\theta \hat{\zeta}_{ij}\left(\hat{\boldsymbol{\theta}}\right)\right]', \\
\hat{\mathcal{V}}_{\upsilon\upsilon} &= \hat{\sigma}^2 \frac{1}{n} \sum_{k=1}^{n} \left[\hat{\chi}_{\upsilon,k} \hat{\chi}'_{\upsilon,k}\right], \\
\hat{\mathcal{V}}_{\upsilon\upsilon'} &= \hat{\sigma}^2 \frac{1}{n} \sum_{k=1}^{n} \left[\hat{\chi}_{\upsilon,k} \hat{\chi}'_{\upsilon',k}\right],
\end{aligned}$$

where we can write $\hat{\Gamma}$ with $\hat{\boldsymbol{\theta}}$, and we can write $\hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n} \hat{v}_k^2$. $\hat{\chi}_{\upsilon,k}\left(\hat{\boldsymbol{\theta}}\right)$ takes different values depending on the value of υ. For example, when $\upsilon = 1$, $\hat{\chi}_{1,k}\left(\hat{\boldsymbol{\theta}}\right) = (n-1)\sum_{j=1,j\neq k}^{n} \nabla_\theta \hat{\zeta}_{jk}\left(\hat{\boldsymbol{\theta}}\right)$ where $\nabla_\theta \hat{\zeta}_{jk}\left(\hat{\boldsymbol{\theta}}\right) = \tau_{jk}\left(\hat{\boldsymbol{\theta}}\right)\sum_{\upsilon=5}^{8} \kappa_\upsilon \nabla_\theta \hat{\varphi}_{\upsilon,jk}\left(\hat{\boldsymbol{\theta}}\right)$.

So far, we have been silent the initial bandwidth that is used in consistent estimators of the bandwidth constant. In addition, we have not mentioned about the initial bandwidth for the variance of the parameters. Since these estimators involve the derivatives of the density or derivatives of the choice probability, we will use bandwidth defined as one that minimizes the MSE of the estimated derivatives of the densities or the estimated derivatives of the choice probabilities, similar to those in Lemmas I.12 and I.13. For example, for the estimated derivatives of the densities, if we use the fourth-order kernel, $b = C_b n^{-1/2\iota+2+s} = n^{-1/10+s}$, where ι is the order of the kernel function and s is the order of the derivatives. Note that theoretically, the choice of the constants C_b could be rather arbitrary and will lead to a consistent estimator of the plug-in components, as long as the bandwidth satisfies that $b \to 0$ as $n \to \infty$. In practice, however, it is also important to explore the sensitivity of the estimators for different choices of C_b.

1.3.3 Practical Issues

To implement the proposed estimation procedure, one needs to specify the kernel function, the numerical values of the bandwidth constants and the trimming function.

We will discuss each of them in detail.

Choosing the Kernel Function

We begin by discussing the choice of the kernel function. Since Assumption 3 requires that the kernel function be symmetric, have the truncated support and satisfy the Lipschitz condition, we consider a set of the kernel functions that could satisfy Assumption 3, including any higher order Epanechenikov, Biweight or Triweight kernel functions. However, since the Gaussian kernel and the uniform kernel cannot satisfy this condition, they are excluded from our discussion. In our analysis, we use an Epanechenikov kernel function in our simulation and the empirical analysis.

Choosing the Bandwidth Constant

To specify the numerical plug-in bandwidth values, we need to specify the order of the kernel function and the constant term. For the order of the kernel function, previous studies (e.g., Horowitz and Härdle (1996) and Lewbel (1997)) have shown that: (i) estimates in simulation using a second-order kernel function are more stable than estimates derived using higher-order kernels; and (ii) a higher-order kernel can perform better only when the sample size is relatively large. On the other hand, asymptotically, the specification on the constant term of the bandwidth is less essential to estimation, while in practice, the constant term can largely affect the performance of the estimator (e.g., Honoré and Kyriazidou (2000)).

To obtain a $\sqrt{n}$-consistent estimator in our method, we require the bandwidth have to satisfy $nh^{2\iota} \to 0$ and $nh^4 \to \infty$. In other words, we need to consider a fourth-order or higher-order kernel to account for this bandwidth requirement. For example, if we choose a fourth-order kernel, the optimal bandwidth will take the form of $h^* = C_h n^{-1/\iota+2} = C_h n^{-1/6}$.

Moreover, in order to obtain the consistent estimator for the constants in the plug-in bandwidth or consistent estimator for the standard error, in fact, we can choose both second order kernel and higher order kernel. In addition, we can use any

bandwidth h such that $h \to 0$. Here, we choose the bandwidth for each component as shown in Section 1.3.2, proportional to the corresponding optimal bandwidth, similar to the result in Lemmas I.12 and I.13.

Choosing the Trimming Functions

Trimming plays a key role in our estimation procedure. Following the discussion in Ichimura and Todd (2007), we use two trimming functions τ_{ij} and G_{ij} to prevent our estimator from having undesirable properties during the estimation procedure.

The first trimming function τ_{ij} guarantees that our estimator will not suffer from a boundary bias problem. The boundary bias problem is commonly found when regressors have compact support (for more details, see Müller (1988, pp. 32-36)). In our analysis, τ_{ij} directly restricts the calculation of the choice probability to the interior of the observables. This trimming approach is standard in the literature.

The second trimming function G_{ij} ensures that the estimators of the choice probabilities are well defined. In practice, there are two candidate trimming schemes that one can use to achieve this goal. The first candidate trimming scheme follows from Lewbel (1997). However, it precludes Taylor's expansion. The second candidate scheme follows from Linton and Xiao (2001) and Buchinsky and Hahn (1998). This trimming scheme does allow for Taylor's expansion. In our analysis, we follow Linton and Xiao (2001) for analytical convenience.

1.4 Conclusion

This paper provides a new semiparametric identification and estimation strategy for the two-player entry game under a symmetry condition on unobservables. Given this symmetry condition, the identification strategy can identify the model parameters using observables with a weaker support condition than that in the existing literature. To some extent, it is a bounded support condition conditional on knowing that the parameters lie in a bounded space. This identification strategy leads

to an estimator with the $\sqrt{n}$-consistency. The findings complement the literature by providing assumptions that bypass the impossibility result of Khan and Nekipelov (2012) in entry game (or more general simultaneous discrete choice) models.

These new results obtained in this paper open several possible directions for extending the proposed method above. First, we are aware that the radial symmetry condition plays a key role in the rate of convergence improvement. As one extension, we will construct a statistic to test for this symmetry condition. Second, this present paper focuses on two-player entry games. Analogously, we may be able to extend our identification and estimation strategy to more than two players with caveats as discussed in Fox and Lazzati (2013). As a caveat to our results, note that $\sqrt{n}$-consistency is an asymptotic result, and does not say anything about the small sample performance of estimator. In the future, we plan to examine in greater details the small sample performance of the propose estimator in comparison with those of the existing methods. Finally, the identification and estimation strategy proposed here also relies on the independent markets assumption. Though the relaxation of this independence assumption is nontrivial, it might be also interesting to explore.

CHAPTER II

A Simulation Design

2.1 Introduction

In the present paper, we will evaluate the performance of the proposed semiparametric estimator in Zhou (2014a) (the first chapter of my dissertation) by using a Monte Carlo study. In particular, we will first illustrate the $\sqrt{n}$-consistency of the proposed semiparametric estimator (to be clear, what we mean by "illustrate the $\sqrt{n}$-consistency" is that in a small sample, the standard error of the semiparametric estimator decreases at an approximate $n^{-1/2}$ rate). Furthermore, we will check whether the proposed semiparametric estimator is more robust to non-normality (or, in general, an unknown distribution of unobservables) compared to other parametric approaches where the normality (or, in general, a distribution of unobservables) is often assumed to be known.

The first goal of this paper is motivated by the fact that in the first chapter of my dissertation, we propose a new semiparametric estimator that has been shown theoretically to have $\sqrt{n}$-consistency. In that paper, we find that one symmetry condition, called the radial symmetry condition, can possibly give additional identification power and lead to a $\sqrt{n}$-consistent estimator. It provides a possibility for the impossibility results as shown in Khan and Nekipelov (2012).[1] It is worth illustrating that, in

[1]Khan and Nekipelov (2012) show that an identification strategy built on the infinite support of

practice, the $\sqrt{n}$-consistency result holds across different symmetric distributions of the unobservables, as predicted by the theorems in Zhou (2014a) (the first chapter of my dissertation).

The second goal of this paper is motivated by the comparison between semiparametric and parametric estimators. Our estimator belongs to semiparametric estimators. It is well known that, in general, a semiparametric estimator is consistent when the distribution is not normal. Although this feature is widely discussed in the entry games literature (e.g., Fox and Lazzati (2013)), few papers have illustrated this feature for entry game models. To complement the literature, this paper considers different simulation designs to compare the parametric estimator proposed in Bresnahan and Reiss (1990, 1991a,b) to the semiparametric estimator proposed in the first chapter of my dissertation. More specifically, we try to test whether the semiparamtric estimator can perform the same as the parametric estimator when normality holds and whether the semiparametric estimator can improve upon the parametric one when normality is violated.

The organization of the present paper is as follows. Section 2.2 contributes to the model setup. Section 2.3 proposes a refined sample objective function to address the possible issues that could happen in the estimation. Section 2.4 discusses the kernel function choice, the bandwidth constant choice as well as the trimming function specification used in the simulation. Section 2.5 presents the results for semiparametric and parametric estimators. Section 2.6 concludes.

observables cannot lead to an estimator with $n^{-1/2}$ rate of convergence, which is a property that the semiparametric literature often attempts to attain.

2.2 Model Setup

The Monte Carlo study is based on a simple two-player entry game, which can be written as follows,

$$\begin{aligned} Y_{1i} &= \mathbb{I}\left(\alpha_1 - Z_{1i} + \Delta_1 Y_{2i} + \varepsilon_{1i} \geq 0\right), \\ Y_{2i} &= \mathbb{I}\left(\alpha_2 - Z_{2i} + \Delta_2 Y_{1i} + \varepsilon_{2i} \geq 0\right); \end{aligned}$$

where (Y_{1i}, Y_{2i}) is a vector of entry outcomes for firms, (Z_{1i}, Z_{2i}) is a vector of firm-market specific observed characteristics and $(\varepsilon_{1i}, \varepsilon_{2i})$ is a vector of unobserved characteristics. In the present paper, we still normalize the coefficients of scalar observables to -1, in order to be consistent with notations in the first chapter of my dissertation. Here $\mathbb{I}$ represents the indicator function.

Let $g\left(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ be a normal density function and $G\left(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ be the corresponding normal cumulative distribution function. More specifically, $g\left(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ can be written as

$$g\left(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = \frac{1}{\left(2\pi\right)^{k/2} \left|\boldsymbol{\Sigma}\right|^{1/2}} \exp\left(-\frac{1}{2}\left(\cdot - \boldsymbol{\mu}\right)' \boldsymbol{\Sigma}^{-1} \left(\cdot - \boldsymbol{\mu}\right)\right);$$

where $\boldsymbol{\mu} \in \mathbb{R}^k$ is a vector of means, $\boldsymbol{\Sigma}$ is a variance-covariance matrix and $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$. In this study, we consider two symmetric distributions of the unobserved characteristics as follows,

$$\text{Design 1:} \quad (\varepsilon_{1i}, \varepsilon_{2i}) \sim F_{\boldsymbol{\varepsilon}} = G\left(\boldsymbol{\varepsilon}; \mathbf{0}_2, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon},1}\right);$$

$$\text{Design 2:} \quad (\varepsilon_{1i}, \varepsilon_{2i}) \sim F_{\boldsymbol{\varepsilon}} = \sum_{b=1}^{4} \lambda_b G_b\left(\boldsymbol{z}; \boldsymbol{\mu}_b, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon},2}\right).$$

For Design 1, we consider the following variance-covariance matrix, $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon},1} = \sigma_\varepsilon^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ with variance $\sigma_\varepsilon^2 = 0.2$. For Design 2, we consider the individual variance-covariance

matrix as $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon},2} = \sigma_{\varepsilon}^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ with variance $\sigma_{\varepsilon}^2 = 0.04$.[2] Moreover, we set the weight as $\bar{\boldsymbol{\lambda}}_b = (0.25, 0.25, 0.25, 0.25)'$, and we denote a vector of means corresponding to each individual normal as $\bar{\boldsymbol{\mu}}_b$, which is specified as

$$\bar{\boldsymbol{\mu}}_b = \begin{pmatrix} (0.4, 0.4) \\ (-0.4, -0.4) \\ (0.4, -0.4) \\ (-0.4, 0.4) \end{pmatrix}.$$

This specification for the individual means guarantees that the four individual normal distributions are pairwise symmetric, in order to ensure that the mixture of these normals satisfies the radial symmetry condition as imposed in Assumption RS in the first chapter of my dissertation. In addition, this specification also enforces the overall centrality point for this four-modal mixture of normals at the origin, so that the centrality point in Design 2 is the same as the one in Design 1 as well. Moreover, we assume that the observed characteristics are drawn from a uniform distribution, which can be written as $Z_{1i}, Z_{2i} \sim Uniform(-1.2, 0.6)$. We set $(\alpha_1, \alpha_2) = (-0.2, -0.2)'$ and $(\Delta_1, \Delta_2) = (-0.2, -0.2)'$. We choose the sample size $n = (500, 1000, 2000)$. Each Monte Carlo design is based on 100 repetitions.

As mentioned in Section 2.1, our experiment is designed to illustrate that the semi-parametric estimator is $\sqrt{n}$-consistent and to compare this semiparametric estimator with an existing parametric estimator. By using the two different designs, where one is normally distributed, and another is not normally distributed, we can first test

[2]Note that the individual variance in Design 2 is much smaller than the variance in Design 1. The reason is that the variance matrix of mixture of normals depends on the dispersion of the individual mean and the individual variances (for more details, please refer to Frühwirth-Schnatter (2006, pp. 169-202)). Here we specify our mean and variance in Design 2 such that the overall variance of unobservables is equal to 0.2, which is the same as that in Design 1. By doing so, we can fix the variance the same across different distributions, so that we can compare the estimates across different shapes of the distributions.

whether the $\sqrt{n}$-consistency can hold across two different symmetric distributions. In addition, the existing parametric estimator has been shown to be consistent when the unobservables are normally distributed, while it is inconsistent when the unobservables are not normally distributed. By using two different designs, we hope to test whether the proposed semiparametric estimator is as good as the parametric estimator under the normality and whether the proposed semiparametric estimator can improve upon the parametric estimator when normality is violated.

2.3 Refined Sample Objective Function

In the estimation, we use a weighted sample objective function as follows.

$$\begin{aligned} \breve{Q}_n(\boldsymbol{\theta}) &= \begin{cases} \frac{\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\tau_{ij}(\boldsymbol{\theta})G_{ij}(\boldsymbol{\theta})\left[\sum_{v=1}^{8}\kappa_v\hat{\varphi}_v(\boldsymbol{Z}_i,\boldsymbol{Z}_j,\boldsymbol{\theta})\right]^2}{\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\tau_{ij}(\boldsymbol{\theta})G_{b,ij}(\boldsymbol{\theta})}, & if\ \boldsymbol{\theta}\in\boldsymbol{\Theta}^* \\ C_p; & if\ \boldsymbol{\theta}\in\boldsymbol{\Theta}/\boldsymbol{\Theta}^* \end{cases} \\ \text{and } \check{\boldsymbol{\theta}}_n &= \arg\min_{\boldsymbol{\theta}} \breve{Q}_n(\boldsymbol{\theta}); \end{aligned} \tag{2.1}$$

where $\boldsymbol{\Theta}^*$ is a bounded subset of the bounded parameter set $\boldsymbol{\Theta}$.[3] This weighted sample objective function is considered here because, without weighting, the sample objective proposed in the first chapter of my dissertation might be minimized at some value not equal to the true parameter value. Why? Recall the discussion of Theorem I.5 in the first chapter of my dissertation, when any reflection points of the observations are outside the support of the observables, we drop these observations in the calculation of the objective function. Given this feature, when the alternative parameters are near the truth, almost all the observations will be used in the sample objective function; whereas when the alternative parameters are far away from the truth, only a few observations might be used in the calculation, resulting in a lower value for the objective function. Reweighting can help to prevent against this possibility as we

[3]Here Θ^* is user-specified, which is rather ad hoc. In this section, we will discuss how to find the bounded parameter set Θ and further construct the subset Θ^*.

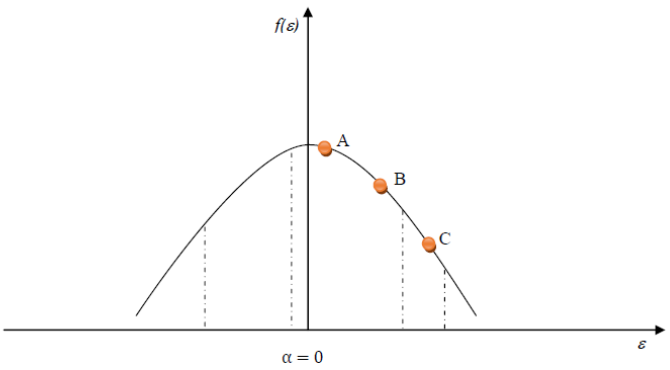

Figure 2.1: The probability density function of the unobservable

move parameters away from the truth. Furthermore, when the alternative parameters are close to the boundary of observables, very few observations are used and the objective function is approximately close to zero. In this case, reweighting may not be enough to prevent this possibility. To protect us further against such possibility, in addition to reweighting, we impose the penalty term C_p when the parameter search is close to the boundary of observables.

Below, we would like to illustrate why the two possibilities will happen. For simplicity, we apply our method to the one-dimensional case to illustrate these two possibilities. To further simplify our discussion, we assume that the data-generating process is $Y = \mathbb{I}(-Z + \varepsilon \geq 0)$, the true symmetric point is at the origin, that is, $\alpha = 0$. In addition, we assume that the observables are uniformly distributed with the support $[z_L, z_U]$, where $z_L < -z_U$.

In Figure 2.1, we consider three points, points A, B and C. Point A represents the alternative parameter a_A that is close to the truth; point B represents the alternative parameter a_B that is slightly far away from the truth; and point C represents the alternative parameter a_C that is even further away from the truth and close to the boundary of the support of observables. In Figure 2.2, correspondingly, when the alternative parameter is at point A, the observations used in the calculation of the objective function are in the range $S_A = [2a_A - z_U, z_U]$; when the alternative parameter

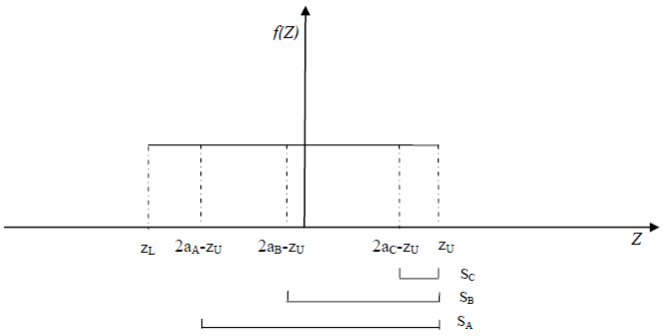

Figure 2.2: The probability density function of the observable

is at point B, the observations used in the calculation are in the range $S_B = [2a_B - z_U, z_U]$; and when the alternative parameter is at point C, the observations used in the calculation are in the range $S_C = [2a_C - z_U, z_U]$.

We expect that when the alternative parameter is at point A, the unweighted sample objective function is larger than that at the truth. Reweighting would not change the relative rank of the objective functions at the alternative and at the truth. At point B, the unweighted sample objective function, however, could be smaller than that at the truth due to a smaller number of observations that are summed over. We hope the reweighting could rescale up the sample objective function value at point B such that the value of weighted objective function compensates for the decreasing number of observations that are summed over and reflects the position of the alternative parameter values. At point C, due to an even smaller number of observations that are summed over, the unweighted sample objective function could be much lower than that at the truth and approximately close to zero. In this case, even though we scale up the unweighted sample objective function after reweighting a small number of the observations, the weighted sample objective function could still remain small and approximately close to zero. In order to prevent this possibility, we need to impose the penalty term for the points like point C. The same analogy can be applied to the two dimensional case in our context.

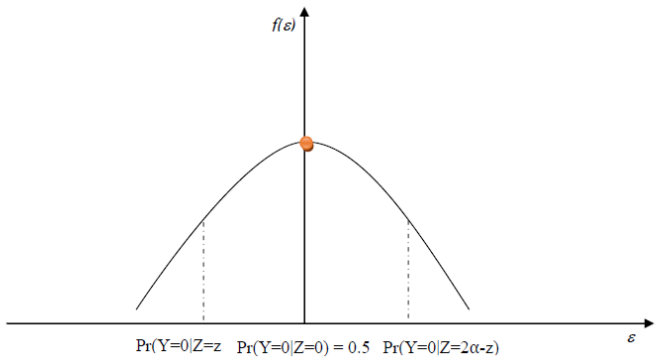

Figure 2.3: The probability at each point

Two practical issues arise when implementing our estimation procedure in practice. First, how can one find a bounded parameter space Θ and define the bounded subset Θ^* at the first place? Second, in practice, since it is possible to have global minimum point at the boundary of observables, is there any way to rule out these parameter values?

To address the first issue, one possibility is to start with partial identification without imposing any distributional assumptions (e.g., Ciliberto and Tamer (2009)). Then, we can use the identified set as the bounded parameter space Θ. As a starting point, we can set Θ^* as close as possible to Θ and impose the symmetry condition to establish point identification.

The second issue will occur when the identified set Θ (and in turn Θ^*) is close to the boundary of the support of observables. To address this issue, we try to use the symmetry condition again and rule out the global minimum at the boundary of observables. The rationale is that the global minimum point at the boundary of observables is not a symmetric point and thus does not satisfy the symmetric property. Using this fact, we could possibly rule out the global minimum point at the boundary of observables in our search procedure.

To illustrate this, we start our discussion in a one-dimensional case and later we will extend the discussion to the two-dimensional case. In the one-dimensional space,

like Figure 2.3, we observe that at the symmetric point, the choice probability of $Y = 0$ is equal to 0.5. When the points are below the symmetric point, the choice probability of $Y = 0$ is less than 0.5 and the choice probabilities of $Y = 0$ is greater than 0.5 when the points are above the symmetric point. Following the observation above, we find that, in identification, when the parameter is at the symmetric point, any observation point and its symmetrically reflected point that are contributed to the objective function satisfy the relation

$$\begin{aligned}
&\text{When } z < 0, \quad \Pr\left[Y = 0 | Z = z\right] \leq 0.5 \leq \Pr\left[Y = 0 | Z = 2\alpha - z\right]; \\
&\text{When } z > 0, \quad \Pr\left[Y = 0 | Z = 2\alpha - z\right] \leq 0.5 \leq \Pr\left[Y = 0 | Z = z\right].
\end{aligned}$$

In estimation, we can estimate $\Pr(Y = 0|Z = z)$ and $\Pr(Y = 0|Z = 2\alpha - z)$ to test the inequality restriction. It is easy to see that point C in Figure 2.1 cannot satisfy the inequality restriction above, that is, the choice probabilities at any point and its reflection point in the support S_c are all above 0.5.

Given this fact, it suggests that we could rule out point C as the symmetric point. Further, we can shrink the parameter search set Θ^* up to point C and estimate the parameter values again. We can continue the procedure iteratively until all possible global minimum points at the boundary of observables are ruled out.

Now, we extend our discussion in a two-dimensional case. The complication in the two-dimensional case arises due to the fact that we do not directly know the value of the choice probability of $Y = (0, 0)$ at the symmetric point as we do in the one-dimensional case. This choice probability is equal to 0.25 only when the correlation of the unobservables is equal to zero. In general, it could be below or above 0.25, depending on the shape of the joint distribution of unobservables. Although we cannot directly know this choice probability, we could recover it from data. We know that the choice probability at the symmetry point in the two-dimensional case is equal to the choice probability when $\boldsymbol{Z} = (0, 0)$. Following the same way as the one dimensional

case, in identification, the following inequality relationship will hold, that is, at the true parameter (α_1, α_2),

When $z_1 < 0,\ z_2 < 0$,

$$\Pr\left[(0,0)|\boldsymbol{Z} = (z_1, z_2)\right] \leq \Pr\left[(0,0)\,|\boldsymbol{Z} = (0,0)\right] \leq \Pr\left[(0,0)|\boldsymbol{Z} = (2\alpha_1 - z_1, 2\alpha_2 - z_2)\right];$$

When $z_1 > 0,\ z_2 > 0$,

$$\Pr\left[(0,0)|\boldsymbol{Z} = (2\alpha_1 - z_1, 2\alpha_2 - z_2)\right] \leq \Pr\left[(0,0)\,|\boldsymbol{Z} = (0,0)\right] \leq \Pr\left[(0,0)|\boldsymbol{Z} = (z_1, z_2)\right].$$

Now given this pair of inequality relationship, the second complication arises here since we can only use the inequality relationship to rule out the alternative parameters $a_1 < \alpha_1, a_2 < \alpha_2$ or $a_1 > \alpha_1, a_2 > \alpha_2$ (or, the homogeneous parameter, i.e. $\alpha_1 = \alpha_2$). However, we cannot rule out the alternative parameters $a_1 > \alpha_1, a_2 < \alpha_2$ or $a_1 < \alpha_1, a_2 > \alpha_2$. Admittedly, this is a limitation of our current estimation procedure.

Below, we propose two possible ways to address the issue presented above. First of all, we have shown in a separate notes that we can identify $2\alpha_1+\Delta_1$ and $2\alpha_2+\Delta_2$ using the radial symmetry condition and the choice probabilites of $(0,0)$ and $(1,1)$ together.[4] Then we could possibly plug in our estimates for (α_1, α_2) (or $(\alpha_1 + \Delta_1, \alpha_2 + \Delta_2)$) to verify the radial symmetry property in the identification of $(2\alpha_1 + \Delta_1, 2\alpha_2 + \Delta_2)$. We could rule out the estimates that can break down the radial symmetry property to identify $(2\alpha_1 + \Delta_1, 2\alpha_2 + \Delta_2)$. Second, throughout the main discussion of our paper, we focus on the equality restriction to build up our identification strategy. But instead, in principle, we could also use the inequality restriction to build up our identification strategy. Due to the dimensionality issue, the possible combination will involve too many comparison scenarios, which would substantially increase computational complexity. We will leave the detailed discussion for future studies.

[4]The detailed discussion is upon request.

2.4 Kernel Function and Trimming Function Specifications

Given the refined weighted sample objective function above, we need to further specify the unknown components, like kernel functions and trimming functions to implement our estimation procedure. In the first chapter of my dissertation, Theorems I.15 and I.16 hold as long as Assumption TR and Assumptions 3-4 are satisfied. It does not provide any specific guidance for the choice of kernel function, the bandwidth constant and the trimming functions. Below, we will discuss how to specify these unknown components to in order to implement our estimation procedure in practice.

2.4.1 Kernel Function Specification and Bandwidth Constant Choice

In this subsection, we will first discuss the specification for the kernel function in the estimation. To construct the weighted sample objective function in (2.1), we first need to estimate the choice probabilities $\hat{\varphi}_{\upsilon}\left(\boldsymbol{Z}_i, \boldsymbol{Z}_j, \boldsymbol{\theta}\right)$ for $\upsilon = 1, \ldots, 8$. For example, $\upsilon = 3$, our kernel estimator is

$$\begin{aligned}
\hat{\varphi}_{n,3}\left(\boldsymbol{z_i}, \boldsymbol{z_j}, \boldsymbol{\theta}\right) &= \hat{\varphi}_n\left(z_{1i}, z_{2j}\right) = \frac{\hat{g}_n\left(z_{1i}, z_{2j}\right)}{\hat{f}_n\left(z_{1i}, z_{2j}\right)}, \text{ with} \\
\hat{g}_n\left(z_{1i}, z_{2j}\right) &= \frac{1}{n-2}\sum\nolimits_{k=1, k\neq i\neq j}^{n} d_k K_n\left(\frac{Z_{1k} - z_{1i}}{h}, \frac{Z_{2k} - z_{2j}}{h}\right), \\
\hat{f}_n\left(z_{1i}, z_{2j}\right) &= \frac{1}{n-2}\sum\nolimits_{k=1, k\neq i\neq j}^{n} K_n\left(\frac{Z_{1k} - z_{1i}}{h}, \frac{Z_{2k} - z_{2j}}{h}\right);
\end{aligned}$$

where $K_n\left(\boldsymbol{u}\right) = \frac{1}{h^2} K\left(\Omega_t^{-1}\left(\boldsymbol{u}\right)\right)$ is a kernel function depending on the covariance matrix Ω_t and the bandwidth $h = h_n$. Here, we consider the forth-order Epanechnikov kernel function which takes the form of

$$k_4\left(u\right) = \frac{15}{8}\left(1 - \frac{7}{3}u^2\right) k\left(u\right),$$

where $k(u) = \frac{3}{4}(1-u^2) I(|u| \leq 1)$. We use the product kernels for the choice probability estimation. Note that any kernel functions that have higher order than the second-order kernel and have bounded support will satisfy the conditions for Theorems I.15 and I.16 in the first chapter of my dissertation. We choose an Epanechnikov higher-order kernel in our application, largely because a higher-order Epanechnikov kernel with an optimal bandwidth has been shown to yield the lowest possible asymptotic mean integrated squared error for density estimation (Hansen (2014)). Although this property may not be preserved for semiparametric estimation, we use it here as a starting point.[5]

Next, we will discuss the bandwidth constant choice. Recall that for the proposed semiparametric estimator we derive in the first chapter of my dissertation, the optimal bandwidth can be written as,

$$h^* = \left(\frac{4\left(c'\mathcal{B}^h\right)^2}{2\iota\left(c'\mathcal{B}\right)^2 n^2} \right)^{1/2\iota+4} = C_h n^{-1/\iota+2}; \tag{2.2}$$

where

$$\begin{aligned}
\mathcal{B}_\upsilon &= \mathbb{E}_{[i,j]}\left[\nabla_\theta \zeta_{ij}\left(\boldsymbol{\theta}^0\right) \sum_{\iota_1+\iota_2=\iota, 0<\iota_1,\iota_2\leq\iota} \left[\int\int u_1^{\iota_1} u_2^{\iota_2} K(\mathbf{u})\, d\mathbf{u} \right] \vartheta_{\upsilon,\iota_1,\iota_2}(\cdot,\cdot) \right], \\
\mathcal{B}_\upsilon^h &= \sigma_\upsilon^2 \mathbb{E}_{[i,j]}\left[-\frac{\nabla_\theta f_{\upsilon,ij}\left(\boldsymbol{\theta}^0\right)}{f_{\upsilon,ij}^{-2}\left(\boldsymbol{\theta}^0\right)} \right] \int\int K^2(\mathbf{u})\, d\mathbf{u};
\end{aligned}$$

and $\mathbb{E}_{[i,j]}$ is the expectation taken over i and j and $\vartheta_{\upsilon,\iota_1,\iota_2}$ are the corresponding bias components. When choosing a bandwidth constant for the plug-in bandwidth selector, four approaches are commonly used in the literature.

The first approach is called Rule of Thumb (ROT) bandwidth choice. This approach presumes that the unobservables and observables are normally distributed and commonly used in a simple one-dimensional case (e.g., Silverman (1986)). The

[5]We will perform robustness check for other higher-order Epanechnikov kernel functions and other kernel functions in future studies.

closed-form solution can be derived and can generate a numerical value of the optimal bandwidth constant directly. Although it is simple in the one-dimensional case, there are several caveats when using this approach in our context. One of most important caveats is that since both unobservables and observables are two-dimensional, beyond the normality assumption we also need to specify the correlation for unobservables and observables. This additional specification adds another source of noise in the first step and leads to less precise estimates for the bandwidth constant. In addition, this noise might be exaggerated when the true underlying distributions are not normal.

The second approach is called Two-step Plug-in bandwidth choice in Wand and Jones (1994). This approach tries to derive the optimal bandwidth constant nonparametrically. Here, we will explain it backwards. In the second step, they try to recover the bandwidth constant nonparametrically. That is, rather than simply assume normal distributions, they use the kernel function to estimate the choice probabilities and their derivatives directly. To guarantee that the kernel estimation works, this step, however, requires the initial bandwidth constant. Because of this, in the first step, they still need to assume normality to obtain the initial bandwidth constant. Note that this bandwidth constant C_b is different from C_h in (2.2) because C_b is derived to minimize the mean squared error of the choice probability estimator rather than the mean squared error of the parameter estimator.

The second approach, to some extent, is less restrictive than the first approach, since the second one incorporates the nonparametric estimation in the second step and it could possibly correct some misspecification in the first step. However, since it still needs to specify the underlying distribution in the first step, the caveat for the first approach will be applied here as well. In addition, the second approach will also incur a large computation burden due to the nonparametric estimation.

A third approach is purely nonparametric, where the researcher needs to specify a wide range of constants and to check how the semiparametric estimator performs

across these different constants (e.g., Honoré and Kyriazidou (2000)). The underlying rationale behind this approach is that, asymptotically, the performance of the estimator can be only affected by the rate of the bandwidth, $n^{-1/\iota+2}$, rather than the bandwidth constant, C_h. Though it sounds very appealing with fewer restrictions, this approach requires a large sample, in order to guarantee that it could mimic the asymptotic scenario where the bandwidth constant is not essential to the estimation. Because of this feature, the third approach is often used when we illustrate the theoretical methods, rather than examine the finite-sample property of the estimators. Like the second approach, the third approach also incurs a large computation burden as well, as we have to obtain our estimates for a range of constants.

The fourth approach is to choose the bandwidth constant (or, the bandwidth itself) subjectively by eye. As mentioned in Section 3 of Wand and Jones (1994), for the density estimator, this procedure starts by looking at several density estimators over a range of bandwidths and choosing the one that is "most pleasing" in some sense. One could try a large bandwidth first and decrease the bandwidth until fluctuations are more "random" than "structural". This approach can be used when one has reasons to believe that there is certain structure in the data, such as knowledge of the position of modes in density estimation. The two potential drawbacks of this method are: (1) it is not applicable when there is no prior knowledge available; (2) it can be very time-consuming to select the bandwidth by eye.

Among these four approaches, in our current experiment, we choose the fourth method as a starting point. Since in the simulation, we have prior knowledge for the true parameter values, we are able to select the bandwidth constant that can make the criterion function achieve its minimum as close as possible to the true parameter value. More specifically, in our design, we start with the bandwidth constant $C_h = 9$ and continuously decrease it to $C_h = 2$. We find that $C_h = 2$ is the smallest bandwidth constant that we can use and a further decrease on the bandwidth constant will

change the shape of the sample objective functions, which is not desirable. Thus, we adopt $C_h = 2$ in our estimation. As a final remark, although we can use the fourth bandwidth constant selection approach in the simulation, it is not feasible to implement our estimation procedure in a general context. Due to this infeasibility, a data-driven bandwidth selection approach, such as the first or the second approach, is more appealing. We will leave further development of data-driven bandwidth selection approaches for our estimation in future studies.

2.4.2 Trimming Function Specification

In this subsection, we will discuss the trimming function specification for the sample objective function in (2.1). We use the first trimming function τ to avoid the boundary bias problem. In addition, instead of using the second trimming function G proposed in the first chapter of my dissertation, in the current experiment, we set the estimated choice probability equal to zero, if it is below zero; and equal to one, if it is above one. We do this to avoid the worry about trimming out too many observations.

To specify the first trimming function τ that satisfies Assumption TR in the first chapter of my dissertation, we consider each element $\tau_\upsilon(\boldsymbol{z}, \tilde{\boldsymbol{z}}, \boldsymbol{\theta})$ for $\upsilon = 1, \ldots, 8$, as follows,

$$\tau_\upsilon(\boldsymbol{z}, \tilde{\boldsymbol{z}}, \boldsymbol{\theta}) = \left\{ \Pi_{p=1}^{2} \exp\left(-\frac{z_{\upsilon,p}^2(\boldsymbol{z}, \tilde{\boldsymbol{z}}, \boldsymbol{\theta})}{b_{\upsilon,p}^2\left(b_{\upsilon,p}^2 - z_{\upsilon,p}^2(\boldsymbol{z}, \tilde{\boldsymbol{z}}, \boldsymbol{\theta})\right)} \right) \mathbb{I}\left(|z_{\upsilon,p}(\boldsymbol{z}, \tilde{\boldsymbol{z}}, \boldsymbol{\theta})| \leq b_{\upsilon,p}\right) \right\}^{1/2};$$

where the trimming point is given by

$$b_{\upsilon,p}(\gamma) = \Phi^{-1}\left(1 - \frac{(1-(1-\gamma))^{1/8}}{2}\right);$$

where $\gamma \in [0, 1]$. If we specify $\gamma = 0.1$, for example, roughly ten percent of the combination of observations is trimmed out when the underlying distribution of observables is normal. This specific trimming function is the square root of the trimming function

proposed in Cattaneo, Crump, and Jansson (2013). This is done because we have eight choice probabilities to be estimated at the same time and each of them has two dimensions, which are pairwisely overlapped (e.g., the first dimension of φ_1 is the same as the first dimension of φ_3). We rescale the trimming function in Cattaneo, Crump, and Jansson (2013) to balance the weight on each dimension and each estimated choice probability.

This trimming function is introduced to avoid the boundary bias problem. The boundary bias problem is well discussed in the kernel estimation literature (for more details, see Müller (1988, pp. 32-36)). In our context, the boundary bias problem occurs in the kernel regression estimation of the choice probabilities at the boundary points of observables. When the estimated choice probabilities badly behaves at the boundary, it will make our sample objective function less accurate over the support of observables. As suggested by Wand and Jones (1994), one possible way to solve such issue is to remove these boundary points from estimation if these boundary observations are not essential to the estimation. Recall that our identification procedure will work as long as the support of observables contains the parameter set. Therefore, in our context, in principle, we can trim as many as possible observation points for estimation as long as the remaining support contains the parameter set. But in practice, how much observations can be trimmed out depends on the specific distribution of observables and the sample size.

Besides avoiding the boundary bias problem, the first trimming function τ will also remove the points at which the estimated choice probabilities are below zero or above one when the underlying distribution is unimodal. However, it cannot work for all the underlying reference distributions, such as mixtures of normals. As mentioned at the beginning of this section, we enforce the estimated choice probability to be between zero and one, by setting the estimated choice probability equal to zero, if it is below zero; and equal to one, if it is above one.

2.5 Results

In this section, we compare the semiparametric estimator proposed in the first chapter of my dissertation to the parametric estimator derived from Bresnahan and Reiss (1990, 1991a,b)(thereafter, BR estimator). To save the space, we leave the review about different estimators in the third chapter of my dissertation.

Table 2.1 summarizes the entry pattern of the two different simulation designs discussed in Section 2.2. It shows the percent of each entry outcome in each design. We specify the simulation design parameters such that the implied entry patterns across the two simulation designs are more or less the same. Any difference in the estimation across two designs is attributed to other aspects of the designs rather than the overall entry patterns. Furthermore, notice that in our specification, the unique equilibria $(0, 0)$ and $(1, 1)$ occur with lower frequency than $(0, 1)$ and $(1, 0)$. It illustrates that our estimation can work without requiring the the unique equilibria $(0, 0)$ and $(1, 1)$ be dominant in the sample.

Table 2.1: Entry Pattern for Designs 1 and 2 (percent)

	(0,0)	(1,1)	(1,0)	(0,1)
		Design 1		
$n = 500$	19.8420	20.1200	29.4620	30.5760
$n = 1000$	19.9690	19.9790	29.4540	30.5980
$n = 2000$	20.0035	19.9485	29.4270	30.6210
		Design 2		
$n = 500$	20.2320	19.7840	29.2300	30.7540
$n = 1000$	20.0690	19.6640	29.5960	30.6710
$n = 2000$	19.9625	19.6230	29.7150	30.6995

To evaluate the performance of the proposed estimator, we present the mean bias and the root mean squared error (RMSE). In addition, since these measures can be affected by outliers, we further consider the median bias and the median absolute deviation (MAD). Finally, we calculate the estimated standard errors to verify the $\sqrt{n}$-consistency of our estimator. Specifically, denoting the rth replication of the estimator

$\boldsymbol{\theta}$ as $\boldsymbol{\theta}(r)$, we explicitly define these statistics as follows: the mean bias ($R^{-1}\sum_{r=1}^{R}$ $\boldsymbol{\theta}(r)-\boldsymbol{\theta}^0$), the root mean squared error ($R^{-1}\sum_{r=1}^{R}\left(\boldsymbol{\theta}(r)-\boldsymbol{\theta}^0\right)^2$), the median bias (median$(\boldsymbol{\theta}(r)-\boldsymbol{\theta}^0)$) and the median absolute deviation ($median\left|\boldsymbol{\theta}(r)-\boldsymbol{\theta}^0\right|$) as well as the standard error $\sqrt{R^{-1}\sum_{r=1}^{R}\left(\boldsymbol{\theta}(r)-\bar{\boldsymbol{\theta}}\right)^2}$, where $\bar{\boldsymbol{\theta}}=R^{-1}\sum_{r=1}^{R}\boldsymbol{\theta}(r)$.

2.5.1 Semiparametric Estimators

In this subsection, we present the results for the semiparametric estimator proposed in the first chapter of my dissertation. We consider the homogeneous coefficients and use the built-in search procedure (fminsearchbnd) in Matlab to search the parameters by setting the initial points at the truth. Finally, we specify our bandwidth as $h=C_h n^{-1/6}$, where $C_h=2$ and we specify the value of γ equal to 0.15 in the simulations, so roughly fifteen percent of the combination of observations at the tail is trimmed out. We set Θ is in the 98 percent of the support of obseravables and set $\Theta^*_{\alpha}=[-0.6,0.2]$ and $\Theta^*_{\alpha+\Delta}=[-0.8,0]$ in our estimation.

Table 2.2 represents the estimates for the proposed semiparametric estimator in Design 1. First, we find that the relative standard error decreases roughly around $1/\sqrt{2}$ (≈ 0.7), as the sample size doubles, for both α and $\alpha+\Delta$, respectively. This result is consistent with our prediction by Theorem I.15 in the first chapter of my dissertation. Second, we find that α and $\alpha+\Delta$ have almost the same magnitude of biases but the biases are in different directions. This is due to assuming the symmetric distribution of the observables and placing the centrality point of the observables between the value of α and $\alpha+\Delta$. The magnitude of the bias could vary depending on different simulation designs. Third, we find that in terms of mean bias and median bias, the magnitude of bias is relative large since the $bias(\theta)/\sqrt{var(\theta)}$ measure is around 1 to 1.5. We suspect it is because the current bandwidth constant may not be optimal for the estimation and the sample size is not large enough to make the bandwidth constant choice not essential to the estimation. While we could

further show that the bias will decrease as we increase the sample size, it will require substantial computation time. In the following subsection, we consider a slightly different procedure which can still show the bias decreases with less computation time.

Table 2.2: Semiparametric Estimates for Design 1

		Mean Bias	RMSE	Median Bias	MAD	SD	RSD
				α = -0.2			
$C_h = 2$	$n = 500$	-0.0758	0.1120	-0.0739	0.0796	0.0829	
	$n = 1000$	-0.0707	0.0859	-0.0684	0.0684	0.0491	0.5923
	$n = 2000$	-0.0554	0.0644	-0.0523	0.0523	0.0330	0.6728
				$\alpha + \Delta$ = -0.4			
	$n = 500$	0.0850	0.1211	0.0810	0.0821	0.0867	
	$n = 1000$	0.0705	0.0931	0.0540	0.0540	0.0611	0.7047
	$n = 2000$	0.0581	0.0690	0.0580	0.0580	0.0375	0.6135

Table 2.3 represents the estimates of the proposed semiparametric estimator in Design 2. First, it shows that the standard error of the estimates under the mixture of normals still follows the $\sqrt{n}$-consistency as we increase the sample size, which is again consistent with our prediction of Theorem I.15 in the first chapter of my dissertation. Second, it shows that under the mixture of normals distribution, the semiparametric estimator tends to perform relatively worse than that under the unimodal of normal distribution, in terms of the mean bias and median bias. However, as we increase the sample size n to 2000, we find that the performances of the estimators under the different two distributions are relatively similar. This observation is consistent with our intuition that the semiparametric estimator tends to perform better and uniformly under a relatively large sample, since any possible noise in the kernel regression estimation of choice probabilities tends to have less effect at a relatively large sample.

The results in Design 1 and Design 2 suggest that for the proposed semiparametric estimator, the $\sqrt{n}$-consistency can hold across different symmetric distributions. The

Table 2.3: Semiparametric Estimates for Design 2

		Mean Bias	RMSE	Median Bias	MAD	SD	RSD
				α = -0.2			
C = 2	n = 500	-0.1215	0.1878	-0.0877	0.0963	0.1439	
	n = 1000	-0.1000	0.1333	-0.0762	0.0762	0.0886	0.6157
	n = 2000	-0.0703	0.0793	-0.0683	0.0683	0.0369	0.4163
				$\alpha + \Delta$ = -0.4			
	n = 500	0.1612	0.1981	0.1236	0.1236	0.1157	
	n = 1000	0.0990	0.1235	0.0855	0.0855	0.0741	0.6404
	n = 2000	0.0719	0.0879	0.0577	0.0577	0.0508	0.6856

semiparametric estimator tends to perform better when the sample size is relatively large. Though the proposed semiparametric estimators have the $\sqrt{n}$-consistency, the biases of the estimators are relatively large in the finite samples. We will leave the discussion for the bias in Section 2.5.3.

2.5.2 Parametric Estimators

In the following context, we compare the BR estimator with the proposed semiparametric estimator. When we consider the parametric estimator, in addition to α and $\alpha+\Delta$, we have two additional parameters to be estimated: the correlation of the bivariate normal, denoted as ρ, and the variance of the bivariate normal, denoted as σ^2. We consider the same search procedure as the semiparametric one for the four parameters here.

Table 2.4 represents the estimates of the BR estimator in Design 1. Note that under Design 1, for the parametric estimator, the underlying distribution of unobservables is correctly specified. We expect that the parametric estimator will satisfy the $\sqrt{n}$-consistency and also the parametric estimator will have a smaller mean bias and median bias. The results in Table 2.4 show that the estimates of the four parameters in the model achieve the $\sqrt{n}$-consistency. In addition, in terms of mean bias and median bias, the parametric estimates are much smaller than the semiparametric ones.

It suggests that the parametric estimator outperforms the semiparametric estimator when the model is correctly specified.

Table 2.4: Parametric Estimates for Design 1

	Mean Bias	RMSE	Median Bias	MAD	SD	RSD
			α = -0.2			
$n = 500$	0.0027	0.0460	0.0064	0.0311	0.0461	
$n = 1000$	0.0025	0.0404	0.0059	0.0271	0.0406	0.8792
$n = 2000$	0.0005	0.0253	0.0025	0.0169	0.0255	0.6276
			$\alpha + \Delta$ = -0.4			
$n = 500$	0.0047	0.0509	0.0049	0.0317	0.0509	
$n = 1000$	-0.0008	0.0357	-0.0003	0.0225	0.0359	0.7039
$n = 2000$	-0.0003	0.0238	-0.0010	0.0162	0.0239	0.6657
			$\sigma = \sqrt{0.2}$			
$n = 500$	0.0146	0.2752	0.0303	0.1761	0.2761	
$n = 1000$	0.0190	0.2234	0.0120	0.1563	0.2237	0.8101
$n = 2000$	0.0039	0.1489	0.0051	0.1010	0.1496	0.6689
			$\rho = 0$			
$n = 500$	0.0008	0.0406	-0.0020	0.0280	0.0408	
$n = 1000$	0.0023	0.0310	-0.0021	0.0168	0.0311	0.7605
$n = 2000$	0.0016	0.0188	0.0016	0.0125	0.0188	0.6048

Table 2.5 presents the estimates of the BR estimator in Design 2. Note that under Design 2, the parametric estimator incurs the misspecification issue, that is, the BR estimator misspecifies a mixture of normals as a unimodal normal distribution. We expect that the BR estimator will be inconsistent and the magnitude of the asymptotic bias will depend on the level of the misspecification. From Table 2.5, first, we find that the mean bias and median bias of the BR estimates in Design 2 are much larger than those in Design 1. It suggests that the BR estimator does suffer from the misspecification issue. Second, we find that the bias tends to be larger as we increase the sample size. It further indicates that the BR estimator has the misspecification issue.

Now, compared the BR estimator with the semiparametric estimator, we find that when the model is correctly specified, the BR estimator outperforms the semipara-

Table 2.5: Parametric Estimates for Design 2

	Mean Bias	RMSE	Median Bias	MAD	SD	RSD
			α = -0.2			
n = 500	-0.0004	0.0603	0.0008	0.0398	0.0606	
n = 1000	0.0076	0.0412	0.0059	0.0262	0.0407	0.6710
n = 2000	0.0134	0.0312	0.0113	0.0212	0.0283	0.6964
			$\alpha + \Delta$ = -0.4			
n = 500	-0.0113	0.0540	-0.0150	0.0297	0.0530	
n = 1000	-0.0171	0.0415	-0.0167	0.0255	0.0381	0.7176
n = 2000	-0.0204	0.0339	-0.0197	0.0230	0.0272	0.7138
			$\sigma = \sqrt{0.2}$			
n = 500	0.0224	0.0434	0.0190	0.0264	0.0374	
n = 1000	0.0149	0.0276	0.0146	0.0174	0.0234	0.6244
n = 2000	0.0121	0.0188	0.0114	0.0124	0.0144	0.6154
			$\rho = 0$			
n = 500	0.0333	0.3212	0.0587	0.2002	0.3211	
n = 1000	0.0638	0.2482	0.0726	0.1690	0.2411	0.7509
n = 2000	0.0876	0.1856	0.0801	0.1336	0.1645	0.6822

metric estimator in terms of lower bias and variance. When the model is misspecified, in terms of bias and variance, the BR estimator still seems better than the semiparametric estimator when the sample size is smaller, $n = 500$. However, the magnitude of the bias increases for BR estimator as the sample size increases, and the magnitude of the bias is much larger than that when the model is correctly specificied. It suggests that the BR estimator is inconsistent, when the model is misspecified. Following the observations in this comparison, though we can show the semiparametric estimator is consistent, it suggests again we need systematically examine the bias in the semiparametric estimator, which will be discussed in Section 2.5.3.

2.5.3 Discussion for the Bias on Semiparametric Estimators

Note that in Section 2.5.1, we find that the bias on the semiparametric estimator is relatively large. As shown in Theorem I.17 in the first chapter of my dissertation, we expect that as the sample size increases, the bias will decrease. In this section, we

will systematically check whether this is the case.

Ideally, we should directly check whether the bias decreases as the sample size increases. However, this could incur an exponential increase in the computation time, since the current sample objective function involves three summations related with the sample size. More specifically, in the current sample objective function, two summations in the outer-loop are attributed to the combination of the observations to construct the difference of the probabilities and one summation in the inter-loop is attributed to the choice probability estimation. As a result, when we double the sample size, the computation time will increase at least 8 times. Given that the current computation time, for $n = 2000$, is 8 hours for each repetition, as the sample size further increases, the increase in computation time will be substantial, which is not desirable. In order to avoid this computational curse of dimensionality, we propose an alternative way to show that the bias decreases as the sample size increases as below.

We suspect that the main source of the bias comes from the estimation of the choice probability. Following this conjecture, it suggests that we could possibly decrease the bias if we increase the number of observations in the kernel estimation of the choice probabilities but keep the combination of the observations to construct the difference of the choice probabilities as low as possible. As long as we can show that the bias decreases when the number of observations for the kernel estimation of choice probability increases, we can safely predict that the bias would decrease, when both the number of observations for kernel estimation and the number of combination of the observations increase. For illustrational purposes, we will conduct this experiment for Design 1. More specifically, to distinguish from the sample size n that we use in the main context, we use n_1 to denote the number of observations in the outer-loop (the combination of the observations) and use n_2 to denote the number of observations in the inter-loop (the kernel estimation of the choice probability). In particular, we

choose $n_1 = 500$ and $n_2 = (500,\ 1000, 2000, 4000, 8000, 20000)$.

Table 2.6 summarizes the results for our experiment (where we keep the same 500 observation points as we change n_2). First, we find that when we increase the number of observation n_2 from 500 to 8000, the bias for α decreases from -0.0758 to -0.0365, which is more than half of the original bias in a small sample; similarly, the bias for $\alpha + \Delta$ decreases from 0.0850 to 0.0411, which is also more than half of the original bias in a small sample. Second, as we further increase the sample size to 20000 to mimic the large sample, we find the bias for α decreases to -0.0232 and the bias for $\alpha + \Delta$ decreases to 0.0395. We expect that the bias would decrease towards zero as we further increase the number of observations substantially.

The results above partially suggest that the bias will decrease as the number of observations for the estimation of the choice probability at the first stage. Further, what we try to emphasize here is that since we only increase the number of observations in the inter-loop, the magnitude of decrease in bias will be even larger as we increase both the number of observations in the inter-loop and the number of the combinations for observations in the outer-loop.

Table 2.6: Semiparametric Estimates for Design 1 (Experiment: Large Sample)

		Mean Bias	RMSE	Median Bias	MAD
			α = -0.2		
$C_h = 2$	$n_2 = 500$	-0.0758	0.1120	-0.0739	0.0796
	$n_2 = 1000$	-0.0718	0.0873	-0.0691	0.0691
	$n_2 = 2000$	-0.0567	0.0657	-0.0537	0.0537
	$n_2 = 4000$	-0.0492	0.0567	-0.0446	0.0446
	$n_2 = 8000$	-0.0365	0.0532	-0.0408	0.0414
	$n_2 = 20000$	-0.0232	0.0525	-0.0358	0.0395
			$\alpha + \Delta$ = -0.4		
	$n_2 = 500$	0.0850	0.1211	0.0810	0.0821
	$n_2 = 1000$	0.0726	0.0982	0.0601	0.0601
	$n_2 = 2000$	0.0576	0.0690	0.0578	0.0578
	$n_2 = 4000$	0.0489	0.0570	0.0427	0.0427
	$n_2 = 8000$	0.0411	0.0453	0.0357	0.0357
	$n_2 = 20000$	0.0395	0.0406	0.0376	0.0376

Based on the results above, one may misinterpret that our estimation procedure can perform well only when we have medium-sized samples or large-sized samples. Here, we will argue that it might be the case for this particular design and cannot be generalized for all the designs. As an important concern, it is worth discussing why our estimator seems to perform worse with a small sample size in the current setting. To answer this question, it is important to examine what factor(s) would affect the first-step estimation (i.e., the estimation of the choice probabilities) and what factor(s) would affect the second-step estimation (i.e., the search procedure for the parameters)?

In the first step, even though we use all the observations for the estimation of the choice probability, the ratio of the unique entry outcomes to the total observations could affect the quality of the estimation. Why? Loosely speaking, in the kernel estimation, the accuracy of the choice probability at each point depends on the neighborhood observations that are similar to this point, that is, more weight is put at the neighborhood observations that are similar to such point. Using this fact, in our context, we expect that the more unique entry outcomes are, the better the estimated choice probability for the unique equilibrium is. Now recall the statistics in Table 2.1. In the current design, unique entry outcomes $(0, 0)$, for example, only account for roughly 20 percent of total observations. It suggests that when $n = 500$, for each point, there are only a few observations that are close to each point in the estimation, which could partially explain why our estimation procedure performs relatively worse in the small sample. We expect that the performance of our estimation procedure can be improved in the small sample in other designs that generate more unique equilibrium outcomes.

In the second step, compared to the parametric estimation, the estimation procedures for α and $\alpha + \Delta$ are separate. To some extent, we do not use all the information in the data at the same time for estimation. That could also explain why in the small

sample, our estimation could be worse. We leave the improvement in future studies.

2.6 Conclusion

In this paper, we consider two different simulation designs for unobserved characteristics in the entry game model, in order to illustrate the $\sqrt{n}$-consistency of the proposed semiparametric estimator and to compare this new semiparametric estimator with the existing parametric estimator.

We find that the proposed semiparametric estimator can approximately achieve $\sqrt{n}$-consistency across different distributions in small samples. It provides the evidence that is consistent with the prediction by Theorem I.15 derived in the first chapter of my dissertation. In addition, compared the proposed semiparametric estimator with an existing parametric estimator, BR estimator, the parametric estimator outperforms our semiparametric estimator when the model is correctly specified. However, the parametric estimator is inconsistent when the model is misspecified. Finally, we find a relatively large bias for the semiparametric estimator in the finite sample. We conduct an experiment to show that the bias will decrease as we increase only the number of observations in estimating choice probability. Given this experiment, we could possibly predict that the bias will decrease, as we increase the overall sample size in constructing the combination of observations and in estimating the choice probability.

The results in this paper suggest two important directions for us to further improve our estimation procedure. First, we need to further improve our sample objective function to avoid the global minimum point that could occur at the boundary of observables in the current sample objective function in the estimation. Second, given a relatively large bias we find that, in the semiparametric estimates, it might be worthwhile to propose a bias-correction estimator in the finite sample, though the bias will shrink as the sample size is large enough.

CHAPTER III

An Empirical Analysis

3.1 Introduction

Entry games are widely applied to a variety of empirical studies, including airline competition, technology adoption and the location choices of discount retailers. Although empirical studies commonly adopt parametric approaches by imposing a normality assumption, when the normality assumption fails to be satisfied, the estimator will be inconsistent. The larger the misspecification is, the greater the inaccuracy of the estimator is. In practice, misspecification can be a big concern for most empirical researchers. Since few alternative approaches are available, few studies have been done to systematically compare the estimators with and without invoking normality assumption for a particular application. Fortunately, in recent years, a sequence of papers have come up with new semiparametric estimators without invoking the normality assumption in entry game literature. Thus, this paper will attempt to compare semiparametric and parametric estimators in an example of location choices of discount retailers. In particular, we will compare the proposed semiparametric estimator in Zhou (2014a) (the first chapter of my dissertation) with two parametric estimators by Bresnahan and Reiss (1990, 1991a,b) and by Berry (1992).

We consider an entry game for two discount retailers, Kmart (K) and Walmart

(W), for markets $m = 1, \ldots, M$,

$$\begin{aligned} Y_{Km} &= \mathbb{I}\left(\alpha_K - \ Z_{Km} + Z_m\beta_K + \Delta_K Y_{Wm} + \varepsilon_{Km} \geq 0\right), \\ Y_{Wm} &= \mathbb{I}\left(\alpha_W - \ Z_{Wm} + Z_m\beta_W + \Delta_W Y_{Km} + \varepsilon_{Wm} \geq 0\right); \end{aligned}$$

where (Y_{Km}, Y_{Wm}) is a vector of entry outcomes for Kmart and Walmart at market m. (Z_{Km}, Z_{Wm}) is a vector of firm-market specific observed characteristics; Z_m is a vector of common market observed characteristics; $\boldsymbol{\varepsilon}_m = (\varepsilon_{Km}, \varepsilon_{Wm})$ is a vector of unobserved characteristics with an unknown distribution. We allow for any correlation between ε_{Km} and ε_{Wm}. For simplicity, we will write $Y_m = (Y_{Km}, Y_{Wm})$ and $X_m = (X_{Km}, X_{Wm})$ where $X_{Km} = (Z_{Km}, Z_m)$ and $X_{Wm} = (Z_{Wm}, Z_m)$. In this type of entry game, the discount retailers will enter a particular market ($Y_{pm} = 1$), for $p = K, W$ only if it is profitable to do so. Under the assumption of complete information, each discount retailer knows $(Y_m, X_m, \varepsilon_m)$, while the econometrician knows only (Y_m, X_m). Our objective is to recover the model parameters using data on retailers' entry decisions and the observed characteristics. The parameters that we are interested in are $((\alpha_K, \beta_K, \Delta_K), (\alpha_W, \beta_W, \Delta_W))$. The key parameters are (Δ_K, Δ_W), representing the competition effects between discount retailers, Kmart and Walmart. Instead of normalizing the variance of the unobservables as used in most parametric estimation, here we normalize the coefficients of the scalar firm-market specific observed characteristics to -1. This normalization allows us to estimate the parameters in both parametric and semiparametric estimation methods.

To recover the competition effects, different approaches invoke different underlying assumptions. It is very important to be aware of these assumptions when we compare different methods. The reasons of comparing these three methods are twofold. First of all, all these three methods require knowledge of the sign of competition effect in identification. Given this, we do not need to worry about the additional model

assumption across the methods. Second, the way of addressing multiple equilibria is quite similar among these three methods, especially between Bresnahan and Reiss (1990, 1991a,b) and Zhou (2014a).[1] Now, the only significant difference among these three methods is that the two parametric methods use the normality assumption to recover choice probabilities; and the semiparametric method relaxes this normality assumption and uses a nonparametric approach to recover the choice probabilities. Thus, if the estimates obtained by the parametric estimators are similar to those obtained by the semiparametric estimator, this could informally indicate normality is a valid approximation in this particular application. Conceptually, we can use the semiparametric estimator and the parametric estimator to construct a formal test of normality. We will leave this for a future study. In addition, there are other semiparametric estimators (e.g., Fox and Lazzati (2013)) that are quite similar to ours. For the time being, we will leave the comparison with other semiparametric estimators in future studies.

The cross-sectional data we use is drawn from Jia (2008) and the geographical information from Census. Several key features in the data allow us to compare these three methods we mentioned above. The first key feature is that the data contain 2065 observations (counties), which are sufficiently large for both parametric and semiparametric estimation. The second key feature is that this data naturally contains an excluded variable, that is, the distance from store to its headquarters. This exclusion restriction is a key requirement for the identification strategy in Zhou (2014a). Thus, with this excluded variable, we are able to compare these three approaches. Third, the data include rich information for both market characteristics and firm-market characteristics, which provides enough variation for estimation for all three approaches.

The remainder of the paper is organized as follows. Section 3.2 provides a short

[1] They all find that the choice probabilities of unique equilibrium contains enough information to recover the model parameters.

review of the three methods. Section 3.3 provides the background on the retailing industry. Section 3.4 describes the data source and the construction of variables. Section 3.5 examines the property of estimators across different approaches using the simulated data. Section 3.6 conducts an empirical analysis of discount retailers using different approaches to verify whether the normality assumption is a good approximation in this particular application. Section 3.7 concludes.

3.2 Review of Methods

In this section, we will review three approaches that provide point estimates of the competition effects in the literature. The first approach is proposed by Bresnahan and Reiss (1990, 1991a,b), the second approach is proposed by Berry (1992), and the third approach is proposed by Zhou (2014a). The first two approaches are parametric approaches that assume the normal distribution on unobserved characteristics. The third approach is a semiparametric approach that relaxes the normality assumption and instead recover the underlying distributions nonparametrically by using kernel estimation procedure. We now review each approach by summarizing its identification strategy and estimation strategy with the main objective function and the algorithm.

3.2.1 Bresnanhan and Reiss (1990, 1991a, b)

Bresnahan and Reiss (1990, 1991a,b) propose a parametric approach for entry games. Recall the two challenges to identification and estimation in the entry game literature: endogeneity and multiple equilibria. To help address the endogeneity problem, they assume a parametric distribution (commonly the normal distribution) for the unobservables and treat an unknown correlation and a standard error as additional parameters to estimate. To solve the multiple equilibria issue, the authors recognize that, when the signs of the competition effects are negative, there exist unique equilibria $(0, 0)$ and $(1, 1)$. Then they can uniquely recover

the choice probabilities $\Pr[(0,0)|X_m]$ and $\Pr[(1,1)|X_m]$ and pool the choice probabilities of multiple equilibria $\Pr[(0,1)|X_m]$ and $\Pr[(1,0)|X_m]$ together (that is, $\Pr[(0,1)|X_m] + \Pr[(1,0)|X_m] = 1 - \Pr[(0,0)|X_m] - \Pr[(1,1)|X_m]$). As a result, they use these choice probabilities to construct the likelihood function to recover the parameters of interest.

More specifically, Bresnahan and Reiss (1990, 1991a,b) propose the following log-likelihood function, which can be written as

$$\begin{aligned}\ln \mathcal{L} \;=\; & \sum\nolimits_{m=1}^{M} \mathbb{I}(Y_{Km}=0, Y_{Wm}=0) \Pr[(0,0)|X_m] \\ & + \mathbb{I}(Y_{Km}=1, Y_{Wm}=1) \Pr[(1,1)|X_m] \\ & + [\mathbb{I}(Y_{Km}=0, Y_{Wm}=1) + \mathbb{I}(Y_{Km}=1, Y_{Wm}=0)] (1 - \Pr[(0,0)|X_m] - \Pr[(1,1)|X_m]).\end{aligned}$$

The key observation behind this log-likelihood function is that the choice probabilities of the unique equilibria $(0,0)$ and $(1,1)$ contain all parameters of interest. So exploring information in the choice probabilities of unique equilibria is sufficient to identify the parameter values. This also suggests that in practice, we can recover the choice probabilities of the unique equilibria $(0,0)$ and $(1,1)$, without establishing the equilibrium selection rule for the occurrence of multiple equilibria. Note that in order to recover the choice probabilities, this method requires two more parameters in addition to the coefficients in the latent profit function: the correlation of unobserved characteristics, denoted as ρ, and the variance of unobserved characteristics, denoted as σ_ε^2. This objective function above naturally suggests the estimation algorithm as shown below.

Step 1. Start with the initial guess of the parameter values and draws from i.i.d. standard normal distribution, i.e., a vector of random variables $\{v_{Km}^t, v_{Wm}^t, v_m^t\}_{t=1}^T$.

Step 2. Given the initial guess of ρ and σ_ε^2, define $\varepsilon_{Km}^t = \sigma_\varepsilon \left(\sqrt{1-\rho} v_{Km}^t + \sqrt{\rho} v_m^t\right)$ and $\varepsilon_{Wm}^t = \sigma_\varepsilon \left(\sqrt{1-\rho} v_{Wm}^t + \sqrt{\rho} v_m^t\right)$.

Step 3. Given the initial guess of $((\alpha_K, \beta_K, \Delta_K), (\alpha_W, \beta_W, \Delta_W))$ and one simulation draw, for the observation at market m, get the predicted entry outcome. Repeat this step T times and obtain the predicted choice probabilities $\widehat{\Pr}[(0,0)|X_m]$ and $\widehat{\Pr}[(1,1)|X_m]$.

Step 4. Plug the estimated choice probabilities into the log-likelihood function. Search the parameter value such that it maximizes the log-likelihood function.

3.2.2 Berry (1992)

Berry (1992) proposes another parametric approach to recover the competition effects. To help address the endogeneity problem, the author still assumes a parametric distribution for the unobserved characteristics. To tackle the multiple equilibria issue, he focuses on the fact that the number of firms in markets is unique despite the existence of the multiple Nash equilibria. Using this fact, he proposes to recover the model parameters by using the predicted number of firms to best match the observed number of firms in each market. This constitutes the key moment conditions that provide the fundamental identification strategy in his paper.

A potential issue of such identification strategy is that it may not identify all the parameters in the model, that is, the number of moments could be less than the number of parameters. To overcome this issue, the author suggests two possible but exclusive solutions: either using the order of entry based on the predicted profit or assuming the order of firms' entry as additional information. In the context of the competition between Kmart and Walmart, it gives three possible specifications: (1) the model with the predicted order of entry based on the firm's predicted profit; (2) the model with the equilibrium most profitable for Kmart; (3) the model with the equilibrium most profitable for Walmart. Together with the main moments drawn from the unique number of firms, we will consider the estimators derived from these three specifications below.

We will review the procedure of constructing the moments to complete the estimation procedure using Berry's method. We start with the moments based on the unique number of firms and then we add additional moments either using the predicted order of firm entry or using the assumed order of firm entry.

3.2.2.1 Moments based on the unique number of firms

Given the model specification, denote $\pi_{pm}(X_{pm}, Y_{-pm}, \varepsilon_{pm}; \theta)$ as a latent profit for discount retailers $p = K, W$, which can be written as

$$\pi_{pm}(X_{pm}, Y_{-pm}, \varepsilon_{pm}; \theta) = \alpha_p - \ Z_{pm} + Z_m\beta_p + \Delta_p Y_{-pm} + \varepsilon_{pm}.$$

To start with, write $\varepsilon_m^t = (\varepsilon_{Km}^t, \varepsilon_{Wm}^t)$ and let $\{\varepsilon_m^t\}_{t=1}^T$ be a sequence of simulation draws $t = 1, \ldots, T$. Denote $\hat{n}_m$ as an estimator of the expected number of the firms in market m, which is defined as

$$\hat{n}_m(X_m, \varepsilon_m^t; \theta) = \max_{0 \leq n \leq P} \left(n : \#\{p : \pi_{pm}(X_{pm}, Y_{-pm}, \varepsilon_{pm}^t; \theta) \geq 0\} \geq n, \text{ where } n = y_{i1} + y_{i2}\right);$$

where $\hat{n}_m(X_m, \varepsilon_m^t; \theta)$ can be interpreted as the largest integer n such that at least n firms are profitable in an n-firm equilibrium. Here, note that Y_{-pm} in the function $\pi_{pm}(X_{pm}, Y_{-pm}, \varepsilon_m^t; \theta)$ represents, given the specific simulation draws, the equilibrium outcome following the Berry (1992) concept. It is not the realized outcome observed in the data. Note that here the key modification from Berry (1992) is that the latent profit depends on the entry of the competitor rather than the total number of firms in the market.[2,3]

[2]In Berry (1992), an unbiased estimator of the expected number of firms is

$$\hat{n}_m(X_m, \theta, u_m) = \max_{0 \leq n \leq K} (n : \#\{k : \pi_{ik}(X_m, n, u_m) \geq 0\} \geq n).$$

which is interpreted as, given the simulation draws, it is the largest integer n such that at least n firms are profitable in an n-firm equilibrium.

[3]There are two differences in the model specification. First, in Berry (1992), the latent profit

Next, we average across the simulation draws to get the averaged predicted number of firms in the market, which is defined as

$$\hat{N}_m(X_m, \varepsilon_m; \theta) = \frac{1}{T} \sum\nolimits_{t=1}^{T} \hat{n}_m(X_m, \varepsilon_m^t; \theta).$$

Finally, following Berry (1992), since $\hat{N}_m(X_m, \varepsilon_m; \theta)$ is an unbiased estimator of the expected number of firms in the market. We can define an estimating equation as

$$N_m = \hat{N}_m(X_m, \varepsilon_m; \theta) + \hat{\vartheta}_m;$$

where $\hat{\vartheta}_m$ is the predicted error, which is mean independent of the exogenous regressor at the true parameters value. We will use this equation to construct our moment condition. It is easy to check that the number of moments that are constructed based on $\hat{\vartheta}_m$ is equal to the number of exogenous regressors in the equation. Suppose that we have S common market observed characteristics (not counting the constant) with one excluded variable for each discount retailers. The resulting number of moments is equal to $S + 2$. When we allow for the parameters to be indexed by the identity of discount retailers, the total number of parameters is equal to $2S + 4 + 2$: $2S$ is the number of coefficients associated with the common market characteristics, 4 includes the constants and the competition effects in the two equations, and 2 represents the number of additional parameters as we use the simulated method of moments, that is the correlation term and the variance term similar to Bresnahan and Reiss (1990, 1991a,b). It is easy to see that the number of moments we have here is far below the number of parameters to be identified. Hence, we need additional moments.

depends on the total number of firms; here the latent profit depends on the entry of the competitor. Second, in Berry (1992), there are no firm-market specific observed characteristics; here there is one firm-market specific observed characteristic. While the expression here is slightly different from Berry (1992), the concept remains the same.

3.2.2.2 Moments based on the order of firm entry

Moments based on the predicted order of firm entry Now, we consider including the information on the predicted order of firm entry in the estimation. A natural question is why the information of the order of firm entry can provide information on the model parameters. The intuition is that a more profitable firm enters first, while a less profitable firm enters later. Under the true parameter values, the predicted order of entry can best match the order contained in the data. Following this logic, we provide the moments that are constructed from the order of firm entry.

First, given the simulation draw, the parameter values and the data observation, we define a function of ranking as $R_p\left(X_m, \varepsilon_m^t; \theta\right)$. Next, we then introduce an unbiased estimator of the probability of entry by the pth firm as

$$\hat{q}_p\left(X_m, \varepsilon_m^t; \theta\right) = \begin{cases} 1, & \text{if } \hat{n}_m(X_m, \varepsilon_m^t; \theta) \geq R_p\left(X_m, \varepsilon_m^t; \theta\right); \\ 0, & \text{otherwise.} \end{cases}$$

Similarly, the averaged estimator of the probability of entry over the simulation draws is defined as

$$\hat{Q}_p\left(X_m, \varepsilon_m; \theta\right) = \frac{1}{T} \sum\nolimits_{t=1}^{T} \hat{q}_p\left(X_m, \varepsilon_m^t; \theta\right).$$

Now, by using the information of the predicted order of firm entry, the additional estimating equation is

$$Y_{pm} = \hat{Q}_p\left(X_m, \varepsilon_m; \theta\right) + \hat{\vartheta}_{pm}.$$

Note that $\hat{\vartheta}_{pm}$ is mean independent of exogenous regressors, so that the number of moments that can be constructed is $2\left(S+2\right)$. Note that with additional moments, we could have an overidentification issue. If it occurs, we can select a subset of observed characteristics among all valid observed characteristics to construct the moments from this additional information.

Moments based on the assumed order of firm entry Finally, we construct the moments based on the assumed order of entry. In particular, we have two different models: one with the equilibrium most profitable for Kmart and the other with the equilibrium most profitable for Walmart. In this context, we consider the assumed order of the entry with the equilibrium most profitable for firm p as long as the profit $\pi_{pm} \geq 0$. Similarly, we have an unbiased estimator of the probability of entry by the pth firms is

$$\hat{q}_p\left(X_m, \hat{\varepsilon}_m^t; \theta\right) = \begin{cases} 1, & \text{if } \pi_{pm}^t \geq 0; \\ 0, & \text{otherwise.} \end{cases}$$

The rest of definition is similar to above. Again, we can construct additional $2(S+2)$ moments.

Given these descriptions, below we will specify the estimation procedure for baseline case, baseline case with the predicted order of firm entry and the assumed order of firm entry to outline the estimation algorithm in Berry (1992).

Baseline case: Here, we focus on the case where only information on the unique number of firms is used.

Step 1. Start with the initial guess of the parameter values and draw from i.i.d. standard normal distribution, i.e., a vector of random variables$\{v_{Km}^t, v_{Wm}^t, v_m^t\}_{t=1}^T$.

Step 2. Given the initial guess of ρ and σ_ε^2, define $\varepsilon_{Km}^t = \sigma_\varepsilon\left(\sqrt{1-\rho}v_{Km}^t + \sqrt{\rho}v_m^t\right)$ and $\varepsilon_{Wm}^t = \sigma_\varepsilon\left(\sqrt{1-\rho}v_{Wm}^t + \sqrt{\rho}v_m^t\right)$.

Step 3. (1) Given the parameter values, obtain the firm-market specific profit for each firm at each market m. (2) Repeat this step for all the simulation draws in Step 1. Calculate the average number of firms across simulation draws of firms whose profits are greater than zero. (This is the inter-loop for market m)

Step 4. Repeat Step 3 for all M markets. Get a sequence of the predicted number of firms. (This is the outer loop for all markets M)

Step 5. Calculate the predicted errors for all M markets and perform the GMM estimation, based on the fact that the predicted errors are independent of all the market characteristics and firm-market characteristics.

Baseline case + the predicted order of firm entry In this context, we add additional information by using the order of a firm entry that is based on the predicted profit, i.e., a more profitable firm enters earlier and a less profitable firm enters later.

Step 1, 2, 4 and 5 are the same as above. The only thing needs to be added is the rank of the firm entry in Step 3.

Step 3. (3) Rank the profit function based on a particular round of simulation draw. Define a new variable to indicate whether firm k will enter the market if the predicted total number of firms is greater than the ranking value of k. (4) Repeat (3) for all the simulation draws. Calculate the average number across simulation draws of firms who will enter the market.

Baseline case + the assumed order of firm entry In this context, additional information is included by assuming an arbitrary order among firms (e.g., favors Kmart or favors Walmart)

Step 1, 2 4 and 5 are the same as above. The only thing that need to added on is the assumed order of firm entry in Step 3.

Step 3. (3) Define a new variable to indicate whether firm p will be in the market if the predicted profit of firm p is greater or equal to zeros. (4) Repeat (3) for all the simulation draws. Calculate the average number across simulation draws of firms who will enter the market.

3.2.3 Zhou (2014a)

Here, we briefly describe the semiparametric identification and estimation strategy in Zhou (2014a) as a comparison to the two parametric methods above. Different

from two parametric methods above, the proposed semiparametric method does not impose any parametric distribution restriction on unobservables. As such, when using this method, simulating the unobservables from a specific distribution is not required. Rather, the choice probability is recovered from a kernel regression estimation. In particular, to help address the endogeneity problem, this paper uses the kernel approach to recover the joint distribution of unobservables. To handle the multiple equilibria problem, this paper only focuses on the unique equilibria that occur in the data and recovers the parameter values using these equilibria. Here, the way to handle the multiple equilibria is similar to Bresnahan and Reiss (1990, 1991a,b).

The weighted sample objective function used in the estimation is further proposed in Zhou (2014b) (the second chapter of my dissertation) as a way to improve the performance of the estimation in Zhou (2014a). The weighted sample objective function can be written as follows,

$$\begin{aligned} \breve{Q}_n(\boldsymbol{\theta}) &= \frac{\frac{1}{M(M-1)}\sum_{i=1}^{M}\sum_{j=1,j\neq i}^{M}\tau_{ij}(\boldsymbol{\theta})\,G_{ij}(\boldsymbol{\theta})\left[\sum_{v=1}^{8}\kappa_v\hat{\varphi}_v(\boldsymbol{X}_i,\boldsymbol{X}_j,\boldsymbol{\theta})\right]^2}{\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\tau_{ij}(\boldsymbol{\theta})\,G_{ij}(\boldsymbol{\theta})}; \quad (3.1) \\ \text{and } \check{\boldsymbol{\theta}}_n &= \arg\min_{\boldsymbol{\theta}_n}\breve{Q}_n(\boldsymbol{\theta}). \end{aligned}$$

The objective function naturally suggests the following estimation algorithm.

Step 1. For each combination by fixing observations i and j, guess the parameter values. Calculate the relative eight choice probabilities using the kernel regression estimation.

Step 2. Plug in the eight choice probabilities to calculate the weighted sample objective function $\breve{Q}_n(\boldsymbol{\theta})$.

Step 3. Search the parameter value such that it minimizes the sample objective function $\breve{Q}_n(\boldsymbol{\theta})$.

As a short summary, in terms of the estimation procedure, it is easy to see that the parametric approaches require simulating the random draws at the beginning

by assuming a known distribution, whereas the semiparametric approach does not require this step but rather needs using a nonparametric method, like kernel regression estimation. Below, we will apply these three approaches to discount retailing industry.

3.3 Review of Discount Retailing Industry

In this section, we will review the background of the discount retailing industry to better understand the construction of the data and interpretation on the estimates later, as we apply different methods to this particular industry. Discount stores, also known as "big box" stores, sell general merchandise items at a substantial discount compared to those sold in department stores. There are three dominant firms in discount retailing industry: Kmart, Walmart and Target. Below, we would like to introduce the history for each firm, respectively.

Kmart opened its first store at 1962 in Garden City, Michigan and originally served the Midwest. Its world headquarters was in Troy, Michigan, but after the purchase of Sears in 2005, the headquarters was relocated to Hoffman Estate near Chicago, Illinois. During the 1990's, Kmart struggled because of poor management and has been surpassed by Walmart as the largest discount retailer in the U.S.. In 2002, Kmart filed for bankruptcy protection and more than 300 stores were closed afterwards.

Walmart also opened its first store at 1962, in Rogers, Arkansas, four months after Kmart opened its first stores. Its world headquarters is in Bentonville, Arkansas. It basically focused on serving the South of U.S., primarily Arkansas, Kansas and Louisana. The expansion of Walmart was originally very slow at the beginning, aimed to suburban areas and tried to avoid direct competition. Around the 1990s, Walmart expanded very rapidly. Walmart is known for its "Everyday Low Price" strategy and "Always Low Prices, Always" slogan. During the 1990s, Walmart became the largest discount retailer in the U.S..

Interestingly, Target also opened its first store at 1962, in Roseville, Minnesota. Target originally expanded in the central areas of U.S.. Target differentiates itself from other retail stores by combining many of the best department store features — fashion, quality and service — with low prices. It places itself as an "upscale" discount retailer with higher-end products. Different from Kmart and Walmart, Target stores are also likely to locate in metropolitan areas in the Midwest. Due to these substantial differences, Target will be excluded from our analysis.

Discount retailing industry has drawn a great deal of attention from researchers in recent decades. Jia (2008) studies the strategic network of store locations of Kmart and Walmart. Zhu and Singh (2009) examine the importance of geographical differentiation in store location decisions of firms in the discount retailing industry. Holmes (2011) shows that the density of stores can increase competition among stores but reduce truck costs for the companies as a whole. Ellickson, Houghton, and Timmins (2013) discuss the effect of chain economies through the network. Orhun (2013) investigates geographic positioning choices of strategic firms and infers the tradeoff between locating close to favorable demand conditions and geographically shielding oneself from rivals. Our analysis and model specification are close to Jia (2008), which will be discussed in the following section.

3.4 Data

Two main data sources are used in our paper: one from Jia (2008), which contains opening and closing information on discount stores from 1988 to 1997, and the other from US Census which includes county level demographic information.

The first data source is drawn from the dataset in 1997 of Jia (2008). We will use this dataset in 1997 as our primary dataset, since it contains the most recent information for discount stores. Following Jia (2008), we define a market as a county. In her analysis and data construction, she excludes very high and very low populated

counties for her analysis. She argues that in sparsely populated counties, demand is not high enough to sustain multiple firms, while in largely populated counties that might have multiple self-contained shopping areas, consumers are less likely to travel across the county to shop at discount stores and other competitors might exist as well. Thus, we have data on 2065 out of all 3140 counties in the US.

Note that Target, the third largest discount retailer, is excluded in the analysis in Jia (2008), because that Target stores are commonly in markets with larger populations and significantly higher income than the markets for Kmart and Walmart, as shown in the literature (e.g., Jia (2008) and Zhu and Singh (2009)). As a result, there are few observations for Target in the counties that comprise the dataset in Jia (2008). Therefore, like Jia (2008), we only focus on the competition of Kmart and Walmart in our analysis.

The second data source is the geographic data for U.S. counties from the US Census Bureau. This data set contains detailed latitude and longitude information for each county. Using this information, we use the counties where Kmart and Walmart headquarters are located as centers and calculate the distances to headquarters using the Haversine formula (for more details, see Zhu and Singh (2009, pp. 19)).

Table 3.1: Entry Pattern

	# of counties
(0, 0)	978
(K, 0)	105
(0, W)	694
(K, W)	288
Obs	2065

Table 3.1 represents the competition configuration across counties between two discount retailers in 1997. We use letter K and W to represent the presence of the discount retailers in a particular county. Following this definition, (0, 0) represents counties with no discount retailers; (K, 0) represents counties in which only Kmart is

present and (K, W) represents the counties with both Kmart and Walmart. We observe that in almost half of counties, neither Kmart nor Walmart enters. In addition, we observe that in almost fifteen percent of counties, both Kmart and Walmart enter. So overall, more than half of the entry outcomes are unique equilibria shown in the data. Note that with one year of data, we cannot examine dynamic entry and exit of discount retailers. This means that, only the stores that are operated in 1997, no matter how long they have been open, are used in our analysis. In addition, a county that has never had a discount retailer is treated the same as a county in which the only discount retailer closed at least one month ahead of the end of year 1997.

Table 3.2: Summary Statistics: Log Value

Variable	Obs	Mean	Std. Dev.	Min	Max
Ln (Population)	2065	2.98	0.67	1.54	4.37
Ln (Per capita retail sales in 1987)	2065	8.20	0.47	5.08	10.66
Urban population ratio in 1990	2065	0.33	0.24	0.00	1.00
Ln (Dist to HQ) (Kmart)	2065	6.28	0.61	3.82	8.37
Ln (Dist to HQ) (Walmart)	2065	6.24	0.63	3.03	8.29

Table 3.2 represents the summary statistics of observables. Note that except for the distances to headquarters, the data is directly from Jia (2008). We first derive the summary statistics of observables including the distances from stores to their headquarters in terms of levels, which is identical to Table II in Jia (2008). However, we present the summary statistics of observables in terms of the logarithm because they are directly used in the estimation. In particular, we include the log of population, the log of per capita retail sales and urban population ratio as well as the log of distance to headquarters for both Kmart and Walmart, respectively. This table provides fundamental information on the simulated data in Section 3.5.

3.5 Simulated Data Illustration

Before we consider the results using real data, it is important to evaluate the performance of the different estimators. To achieve this goal, we will compare different estimators using the simulated data under the normality assumption of unobservables.

3.5.1 Model Setup

Here, we consider a two-player entry game, which is drawn from the following data-generating process:

$$\begin{aligned} Y_{Km} &= \mathbb{I}\left(\alpha_K - \ Z_{Km} + Z_m\beta_K + \Delta_K Y_{Wm} + \varepsilon_{Km} \geq 0\right), \\ Y_{Wm} &= \mathbb{I}\left(\alpha_W - \ Z_{Wm} + Z_m\beta_W + \Delta_W Y_{Km} + \varepsilon_{Wm} \geq 0\right); \end{aligned}$$

where $(\varepsilon_{Km}, \varepsilon_{Wm})' = N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \sigma_\varepsilon^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$ where $\sigma_\varepsilon^2 = 20$ and $\rho = 0.5$; and in addition,

$$\begin{pmatrix} Z_{Km} \\ Z_{Wm} \\ Z_{m,1} \\ Z_{m,2} \\ Z_{m,3} \end{pmatrix} = N\left(\begin{bmatrix} 6 \\ 6 \\ 3 \\ 8 \\ 0.3 \end{bmatrix}, \sigma_Z^2 \begin{bmatrix} 0.36 & & \cdots & & 0 \\ & 0.36 & & & \\ \vdots & & 0.36 & & \vdots \\ & & & 0.25 & \\ 0 & & \cdots & & 0.9 \end{bmatrix}\right)$$

where $\sigma_Z^2 = 15$. We specify the model parameters $(\alpha_K, \beta_K, \Delta_K) = (-24, 1.8, 2, 1.5, -1)$ and $(\alpha_W, \beta_W, \Delta_W) = (-16, 2, 1.8, 1.6, -1)$, which follows the estimates in Jia (2008). Note that we adjust the standard errors of observables and unobservables to best match the entry pattern in the original data. In this study, we consider the experiment with the sample size $n = 1200$ and the repetition $R = 100$. In addition, for the parametric estimators, in each repetition, we consider $T = 1500$ independent ran-

dom draws for unobserved characteristics. We verify the entry pattern and observed characteristics for the simulated data, which matches with Table 1 and Table 2.

3.5.2 Results

In this subsection, we compare the following estimators: "Zhou" (the semiparamtric estimator proposed by Zhou (2014a)); "BR" (the parametric estimator proposed by Bresnahan and Reiss (1990, 1991a,b)); "Berry" (the parametric estimator proposed by Berry (1992) using the unique number of firms in the markets and the predicted order of firm entry); "Favors Kmart" (the parametric estimator proposed by Berry (1992) using the unique number of firms in the markets and the assumed order of the firm entry with the equilibrium most profitable for Kmart); "Favors Walmart" (the estimator proposed by Berry (1992) using the unique number of firms in the markets and the assumed order of the firm entry the equilibrium most profitable for Walmart).

To evaluate the estimators, we consider five statistics: mean bias, the root mean squared error (RMSE), the median bias and the median absolute deviation (MAD) as well as the estimated standard error. Specifically, denoting the rth replication of the estimator $\boldsymbol{\theta}$ as $\boldsymbol{\theta}(r)$, we explicitly define these statistics as follows: the mean bias $(R^{-1}\sum_{r=1}^{R}\boldsymbol{\theta}(r)-\boldsymbol{\theta}^0)$, the root mean squared error $(R^{-1}\sum_{r=1}^{R}\left(\boldsymbol{\theta}(r)-\boldsymbol{\theta}^0\right)^2)$, the median bias $(\text{median}(\boldsymbol{\theta}(r)-\boldsymbol{\theta}^0))$ and the median absolute deviation $(median\left|\boldsymbol{\theta}(r)-\boldsymbol{\theta}^0\right|)$ as well as the standard error $\sqrt{R^{-1}\sum_{r=1}^{R}\left(\boldsymbol{\theta}(r)-\bar{\boldsymbol{\theta}}\right)^2}$, where $\bar{\boldsymbol{\theta}}=R^{-1}\sum_{r=1}^{R}\boldsymbol{\theta}(r)$.

Table 3.3 presents the estimates across these different estimation methods. Note that when the normality assumption holds and there is no misspecification for parametric estimators, we find that the estimates are more or less the same across the different approaches. In terms of absolute value, the mean bias and median bias for the semiparametric estimates are larger than those for the parametric estimates. Also, the root mean squared error and median absolute deviation are also larger for the semiparametric estimates. In other words, when the model is correctly specified,

Table 3.3: Parameter Estimates from Different Methods (Simulated Data)

		Zhou	BR	Berry	Favors Kmart	Favors Walmart
Kmart's Profit						
Log population	Mean Bias	0.1670	0.0375	-0.0970	-0.1346	-0.0807
	RMSE	0.5400	0.2887	0.6267	0.5978	0.5969
	Median Bias	0.2505	0.0692	-0.1102	-0.0910	-0.0860
	MAD	0.4334	0.1485	0.5685	0.4373	0.4186
	SD	0.5200	0.2878	0.6226	0.5856	0.5947
Log Retail Sales/Capita	Mean Bias	0.3148	0.0672	0.0440	0.0351	0.0370
	RMSE	0.6111	0.3462	0.2126	0.1896	0.2038
	Median Bias	0.5373	0.0320	0.0132	0.0156	0.0336
	MAD	0.5735	0.1454	0.1018	0.0918	0.1085
	SD	0.4767	0.3415	0.2091	0.1874	0.2016
Urban Ratio	Mean Bias	0.4801	0.2853	-0.0134	-0.0135	-0.0074
	RMSE	0.6083	0.3889	0.1027	0.1147	0.1055
	Median Bias	-0.3637	0.1816	-0.0126	-0.0204	-0.0055
	MAD	0.5660	0.1816	0.0585	0.0513	0.0500
	SD	0.5572	0.2657	0.1024	0.1146	0.1058
Constant	Mean Bias	0.1170	0.2038	0.1285	0.1459	0.1170
	RMSE	0.6952	0.3508	0.5456	0.4757	0.4916
	Median Bias	-0.6802	0.1823	0.1104	0.1325	0.1670
	MAD	0.6802	0.2052	0.4215	0.3110	0.3148
	SD	0.3124	0.2870	0.5332	0.4553	0.4801
Competition Effect	Mean Bias	0.4831	0.1497	0.0870	0.1187	0.0614
	RMSE	0.9255	0.3394	0.4948	0.4531	0.4503
	Median Bias	0.3139	0.1333	0.0493	0.1512	0.0432
	MAD	0.6371	0.1932	0.3291	0.2979	0.2308
	SD	0.7937	0.3063	0.4898	0.4397	0.4485

(Continues)

		Zhou	BR	Berry	Favors Kmart	Favors Walmart
Walmart's Profit						
Log population	Mean Bias	0.7487	0.1541	-0.0336	-0.0427	-0.0151
	RMSE	0.7941	0.3330	0.6477	0.5803	0.6264
	Median Bias	0.8478	0.2208	-0.0094	0.0398	0.0247
	MAD	0.8478	0.2541	0.5812	0.3667	0.4895
	SD	0.2661	0.2968	0.6504	0.5819	0.6297
Log Retail Sales/Capita	Mean Bias	0.8834	-0.1672	0.0227	0.0251	0.0152
	RMSE	0.8978	0.3522	0.1515	0.1470	0.1681
	Median Bias	0.9796	-0.1155	0.0035	0.0209	0.0200
	MAD	0.9796	0.1412	0.0627	0.0504	0.0848
	SD	0.1611	0.3117	0.1506	0.1456	0.1683
Urban Ratio	Mean Bias	0.1034	-0.0935	-0.0105	-0.0065	-0.0028
	RMSE	0.5665	0.3261	0.0687	0.0708	0.0662
	Median Bias	0.1799	-0.0105	-0.0131	-0.0050	-0.0084
	MAD	0.4436	0.1109	0.0369	0.0365	0.0318
	SD	0.5601	0.3141	0.0682	0.0709	0.0665
Constant	Mean Bias	0.4916	0.3298	0.0395	0.0786	0.0682
	RMSE	0.7499	0.4143	0.4147	0.4036	0.3745
	Median Bias	-0.7349	0.3205	0.0347	0.0706	0.0927
	MAD	0.7349	0.3239	0.2289	0.2627	0.2139
	SD	0.2622	0.2523	0.4151	0.3981	0.3703
Competition Effect	Mean Bias	0.2799	0.1013	0.0671	0.1077	0.0465
	RMSE	0.7614	0.3150	0.5073	0.4992	0.4851
	Median Bias	0.0011	0.1043	-0.0086	0.0861	0.0320
	MAD	0.4153	0.1824	0.3949	0.3518	0.3174
	SD	0.7120	0.2999	0.5057	0.4901	0.4855

the parametric estimators outperform the semiparametric estimator, consistent with the simulation studies by Zhou (2014b). Finally, we also find that in general, the Berry estimators perform better than the BR estimator. It might be possible that given the current number of the simulation draws for each repition, the simulation error in ML estimation exists in the BR estimator. Given these results, we will move to the real data analysis, we expect that when the normality condition holds, the different estimators will provide similar results.

3.6 An Empirical Illustration

In this section, we will compare the three estimation methods using the data of discount retailers, Kmart and Walmart. In particular, we are interested in the following entry game model of discount retailers,

$$\begin{aligned} Y_{Km} &= \mathbb{I}\left(\alpha_K - Z_{Km} + Z_m\beta_K + \beta_K^d Midwest + \Delta_K Y_{Wm} + \varepsilon_{Km} \geq 0\right), \\ Y_{Wm} &= \mathbb{I}\left(\alpha_W - Z_{Wm} + Z_m\beta_W + \beta_W^d South + \Delta_W Y_{Km} + \varepsilon_{Wm} \geq 0\right); \end{aligned}$$

where (Y_{Km}, Y_{Wm}) is a vector of entry outcomes in county m; (Z_{Km}, Z_{Wm}) is a vector of the distances from the county to each store's headquarters; Z_m is a list of market characteristics, including the log of population, the log of retail sales per capita, and the urban population ratio. In addition, we also consider the effect of the regional location $Midwest$ for Kmart and the effect of regional location $South$ for Walmart. $(\varepsilon_{Km}, \varepsilon_{Wm})$ are allowed to be correlated with an unknown distribution. Although $(\varepsilon_{Km}, \varepsilon_{Wm})$ only need to be independent of the excluded regressor in identification, we follow Jia (2008) and others in assuming that the unobservables are independent of other regressors.

Given this model specification, we assume that Kmart and Walmart make independent decisions across markets, that is, when they decide whether to enter a

particular market, they do not take into account of entry decisions in other markets. This imposed myopia eliminates network or chain effects, which are salient to the discount retailing industry. It is unavoidable to discuss how excluding these effects would change the point estimates. To answer this question, we need to examine what these network effects or chain effects reflect in a firm's profit or operation. Holmes (2011) finds that, the density of network can help reduce the truck cost when setting up the stores next to each other. Jia (2008) suggests that nearby stores split the costs of operations, delivery and advertising to achieve scale economies. In addition, the nearby stores also share knowledge of local markets and learn from one another's managerial success. All these factors suggest that having stores nearby benefits the operation in a nearby market and that the benefit declines with the distance. Besides this, the evidence also suggests that the network structure or chain structure reflects their managerial effort and improves their management. Now, following the discussion by Ellickson and Misra (2011), from an economic perspective, the constant terms summarize the managerial effects that are not explained by other observables. Presumably, the network structure or chain structure is uncorrelated with other observables. When we ignore the network or chain structure, the estimated constant terms might be upward biased as it now contains the positive average chain effects in addition to the average managerial effects for other unobserved characteristics. Based on this discussion, we might expect that other coefficients might remain the same as Jia (2008) if the normality assumption is a reasonable approximation but the constants might be upward biased (i.e., decrease in term of absolute values) as we do not take into account of the network in our model specification. We will verify this using our estimates.

Table 3.4 presents the estimates across different methods. From Table 3.4, we find that the signs of all coefficients are the same across the different approaches and the magnitudes of the coefficients are roughly the same except for the competition

Table 3.4: Parameter Estimates from Different Methods (Real Data)

	Zhou	BR	Berry	Favors Kmart	Favors Walmart
Kmart's Profit					
Log population	1.72	1.58	1.86	1.84	1.83
	(0.12)	(0.07)	(0.29)	(0.10)	(0.09)
Log Retail Sales/Capita	1.92	1.78	2.16	2.16	2.14
	(0.06)	(0.07)	(0.16)	(0.05)	0.06
Urban Ratio	1.32	1.28	1.49	1.50	1.48
	(0.16)	(0.08)	(0.36)	(0.20)	(0.11)
Midwest	0.51	0.42	0.55*	0.55	0.53
	(0.18)	(0.10)	(0.30)	(0.12)	(0.06)
Constant	-19.63	-20.49	-19.12	-19.03	-19.04
	(0.58)	(1.09)	(0.96)	(0.43)	(0.26)
Competition Effect	-0.76	-0.96	-0.89	-0.88	-0.83
	(0.60)	(0.15)	(0.17)	(0.09)	(0.08)
Walmart's Profit					
Log population	1.92	2.14	2.02	2.00	1.95
	(0.18)	(0.07)	(0.29)	(0.20)	(0.18)
Log Retail Sales/Capita	1.96	1.99	1.83	1.81	1.83
	(0.15)	(0.21)	(0.14)	(0.07)	0.05
Urban Ratio	1.56	1.54	1.65	1.70	1.70
	(0.13)	(0.07)	(0.30)	(0.13)	(0.13)
South	0.95	0.99	1.04	1.05	1.22
	(0.14)	(0.06)	(0.34)	(0.22)	(0.18)
Constant	-14.11	-14.89	-15.92	-15.76	-15.75
	(0.61)	(0.33)	(0.85)	(0.38)	(0.44)
Competition Effect	-2.89	-0.78	-0.88	-0.91	-1.01
	(1.04)	(0.14)	(0.18)	(0.06)	(0.07)

Note: All coefficients are statistically significant at 5 percentage level except *

effects.[4] To some extent, the results informally suggest that normality seems a fairly reasonable approximation in this particular application.

Note that even though all the estimators have a closed-form solution for the standard error, it is very computationally demanding to calculate them. For the semiparametric estimator, it involves both the kernel estimation for the choice probabilities and the different orders of derivatives of the choice probabilities. In addition, for the parametric estimators, it also involves the numerical derivatives. So here we consider bootstrapped standard errors to avoid computation burden. We construct bootstrap 1200 out of 2065 with 100 replications. Note that we could also consider the bootstrap with clusters, which we will leave for a future study.

Compared with the results of Jia (2008), we find that all coefficients except the constant terms are more or less the same as those in Jia (2008) but the constant terms are smaller than the ones in Jia (2008). Recall that as we discussed at the beginning, when ignoring the network or store chains, we expect that other coefficients remain the same but the constant terms will be upward biased as it contains the additional averaged positive network effects. The estimates in our results now are consistent with our prediction.

Finally, we will discuss the economic implications for our estimates. First of all, the constant terms reflect the profitability that cannot be explained by observed factors. From the estimates, we find that Walmart is slightly more profitable (with higher coefficients) whereas Kmart is less profitable (with lower coefficients). This is consistent with Kmart having a systematic management problem after 1990s. Second, we find that urban population ratio favors for Walmart rather than Kmart, even though at the beginning, Walmart targeted to more suburban areas. Finally, we

[4] As the search for the competition effects is a separate procedure from other coefficients, we could possibly only investigate issues associated with competition effects without worrying about other coefficients. One possible reason for this exceptional estimate for the competition effects could be due to the fact, that the proportion of entry $(1, 1)$ is relative smaller than the proportion of no entry $(0, 0)$. In other words, there are less variations to estimate competition effects. But we need more investigation before we come to the conclusion.

find that in terms of the competition effects, Walmart has a larger effect on Kmart than that Kmart does on Walmart. It is also consistent with the discussion that Walmart is a dominant player and Kmart is relatively weak. As a final remark, since our estimates are based on data in 1997, it can only provide partial information to infer competition between Kmart and Walmart today, since there was a large scale closure of Kmart stores around 2002. We might expect that the competition effect of Walmart on Kmart is even larger today than what we estimate here.

3.7 Conclusion

In this paper, we compare the estimates for the static entry game of complete information between the semiparametric approach and the parametric approaches in the discount retailing industry. In particular, using the data from Jia (2008), we find that the estimates using semiparametric approach are quite similar to those of the parametric approaches. This informally suggests that the normality assumption seems a fairly reasonable approximation for the underlying distribution of unobservables in the entry game of discount retailers. In addition, we find that compared with other studies, like Jia (2008), our parametric estimates are very close to hers except the constant. It is consistent with our prediction that when ignoring the network or store chains, we expect that other coefficients remain the same but the constant terms will be upward biased as it contains the positive averaged network effects.

APPENDIX

APPENDIX A

A.1 Extension

A.1.1 Extension I: Multivariate Covariates under Heteroskedasticity

In this section, we will discuss a more general case with multivariate covariates under heteroskedasticity. To facilitate our analysis, we define a vector of observables $X_{pi} = (Z_{pi}, W_{pi})$ for each player $p = 1, 2$, where as before Z_{pi} is a scalar observable, and we decompose $W_{pi} = \left(\breve{Z}_{pi}, Z_i\right)$, where $\breve{Z}_{pi}$ are firm-market characteristics other than Z_{pi}, and Z_i are the market characteristics. Finally, we use (β_1, β_2) as their conformable coefficients with (Z_{1i}, Z_{2i}), which is still normalized as $(-1, -1)$, and we use (Λ_1, Λ_2) as the coefficients associated with (W_{1i}, W_{2i}). In this section, we provide the identification results with multivariate covariates under the heteroskedasticity assumption. Note that we maintain Assumptions S throughout this appendix. We will modify Assumptions R, ER and RS accordingly and add additional assumptions for other regressors to accommodate the multivariate covariates, that is, beyond Assumption S in the main context, we assume that the following regularity conditions hold.

Assumption RMH (Random Sampling with Multivariate Covariates under Heteroskedasticity): *An independent sample* $\{\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{\varepsilon}_i\}_{i=1}^{n}$ *is drawn from the population.*

Assumption RMH still assumes that firms make independent decisions across markets.

Assumption ERMH (Exclusion Restriction with Multivariate Covariates under Heteroskedasticity) *Suppose that a vector of observed characteristics* (Z_{1i}, Z_{2i}) *satisfies that:*

(i) (Z_{1i}, Z_{2i}) *is independent of* $(\varepsilon_{1i}, \varepsilon_{2i})$ *conditional on* (W_{1i}, W_{2i})

(ii) the scaler covariate Z_{pi} *enters only the payoff function for player* p*, but not the payoff function for other players.*

Assumption ERMH is a generalized version of Assumption ER given the multivariate regressors, which allows the excluded regressors to be conditionally independent of unobservables. In particular, we only need one excluded regressor in our identification.

Assumption RSMH (Radial Symmetry with Multivariate Covariates under Heteroskedasticity): *The conditional distribution of the unobserved characteristics* $(\varepsilon_1, \varepsilon_2)$ *is continuous over the support* $\mathcal{S}_{\boldsymbol{\varepsilon}}$ *conditioning on* (W_{1i}, W_{2i}) and *is radially symmetric around* (α_1, α_2)*; that is,*

$$f\left(\varepsilon_1, \varepsilon_2 | W_{1i}, W_{2i}\right) = f\left(2\alpha_1 - \varepsilon_1, 2\alpha_2 - \varepsilon_2 | W_{1i}, W_{2i}\right).$$

Assumption RSMH requires that the symmetry points remain the same conditioning on the observables. Note that radial symmetry implies $\mathbb{E}\left(\varepsilon_1\right) = \alpha_1$ and $\mathbb{E}\left(\varepsilon_2\right) = \alpha_2$, where α_1 and α_2 are the respective medians of the respective conditional distributions.

Given these assumptions, we can now construct the identifying restriction.

Lemma A.1. *Suppose that* $\boldsymbol{W}_{\upsilon} = \boldsymbol{W}_{\upsilon'} = \boldsymbol{w} = (w_1, w_2)$*, for* $\upsilon \neq \upsilon'$*. For any two vectors* $\boldsymbol{z} = (z_1, z_2)$*,* $\tilde{\boldsymbol{z}} = (\tilde{z}_1, \tilde{z}_2) \in \mathcal{S}_{\boldsymbol{z}|\boldsymbol{w}}$*, consider*

$$\begin{aligned}
\boldsymbol{Z}_1 &= (z_1, z_2)\,; & \boldsymbol{Z}_5 &= (2(\alpha_1 + w_1\Lambda_1) - z_1, 2(\alpha_2 + w_2\Lambda_2) - z_2)\,;\\
\boldsymbol{Z}_2 &= (\tilde{z}_1, \tilde{z}_2)\,; & \boldsymbol{Z}_6 &= (2(\alpha_1 + w_1\Lambda_1) - \tilde{z}_1, 2(\alpha_2 + w_2\Lambda_2) - \tilde{z}_2)\,;\\
\boldsymbol{Z}_3 &= (z_1, \tilde{z}_2)\,; & \boldsymbol{Z}_7 &= (2(\alpha_1 + w_1\Lambda_1) - z_1, 2(\alpha_2 + w_2\Lambda_2) - \tilde{z}_2)\,;\\
\boldsymbol{Z}_4 &= (\tilde{z}_1, z_2)\,; & \boldsymbol{Z}_8 &= (2(\alpha_1 + w_1\Lambda_1) - \tilde{z}_1, 2(\alpha_2 + w_2\Lambda_2) - z_2)\,.
\end{aligned}$$

Given that Assumptions RMH, S and ERMH hold, define

$$\begin{aligned}
& B_0(\boldsymbol{z}, \tilde{\boldsymbol{z}}; \boldsymbol{\alpha}, \boldsymbol{\Lambda})\\
= \;& \Pr((0,0)\,|\boldsymbol{X}_1) + \Pr((0,0)\,|\boldsymbol{X}_2) - \Pr((0,0)\,|\boldsymbol{X}_3) - \Pr((0,0)\,|\boldsymbol{X}_4);\\
& B_0(2(\boldsymbol{\alpha} + \boldsymbol{w}\boldsymbol{\Lambda}) - \boldsymbol{z}, 2(\boldsymbol{\alpha} + \boldsymbol{w}\boldsymbol{\Lambda}) - \tilde{\boldsymbol{z}}; \boldsymbol{\alpha}, \boldsymbol{\Lambda})\\
= \;& \Pr((0,0)\,|\boldsymbol{X}_5) + \Pr((0,0)\,|\boldsymbol{X}_6) - \Pr((0,0)\,|\boldsymbol{X}_7) - \Pr((0,0)\,|\boldsymbol{X}_8).
\end{aligned}$$

By Assumption RSMH, we have

$$B_0(\boldsymbol{z}, \tilde{\boldsymbol{z}}; \boldsymbol{\alpha}, \boldsymbol{\Lambda}) - B_0(2(\boldsymbol{\alpha} + \boldsymbol{w}\boldsymbol{\Lambda}) - \boldsymbol{z}, 2(\boldsymbol{\alpha} + \boldsymbol{w}\boldsymbol{\Lambda}) - \tilde{\boldsymbol{z}}; \boldsymbol{\alpha}, \boldsymbol{\Lambda}) = 0.$$

In addition, we can also consider the identifying restriction for the competition effects.

Lemma A.2. *Suppose that* $\boldsymbol{W}_{\upsilon} = \boldsymbol{W}_{\upsilon'} = \boldsymbol{w} = (w_1, w_2)$*, for* $\upsilon \neq \upsilon'$*. For any two*

vectors $\boldsymbol{z} = (z_1, z_2)$, $\tilde{\boldsymbol{z}} = (\tilde{z}_1, \tilde{z}_2) \in \mathcal{S}_{\boldsymbol{z}|\boldsymbol{w}}$, *consider*

$$\begin{aligned}
\boldsymbol{Z}_1 &= (z_1, z_2); \qquad \boldsymbol{Z}_5 = (2(\alpha_1 + \Delta_1 + w_1\Lambda_1) - z_1, 2(\alpha_2 + \Delta_2 + w_2\Lambda_2) - z_2); \\
\boldsymbol{Z}_2 &= (\tilde{z}_1, \tilde{z}_2); \qquad \boldsymbol{Z}_6 = (2(\alpha_1 + \Delta_1 + w_1\Lambda_1) - \tilde{z}_1, 2(\alpha_2 + \Delta_2 + w_2\Lambda_2) - \tilde{z}_2); \\
\boldsymbol{Z}_3 &= (z_1, \tilde{z}_2); \qquad \boldsymbol{Z}_7 = (2(\alpha_1 + \Delta_1 + w_1\Lambda_1) - z_1, 2(\alpha_2 + \Delta_2 + w_2\Lambda_2) - \tilde{z}_2); \\
\boldsymbol{Z}_4 &= (\tilde{z}_1, z_2); \qquad \boldsymbol{Z}_8 = (2(\alpha_1 + \Delta_1 + w_1\Lambda_1) - \tilde{z}_1, 2(\alpha_2 + \Delta_2 + w_2\Lambda_2) - z_2).
\end{aligned}$$

Given that Assumptions RMH, S and ERMH hold, define

$$\begin{aligned}
& B_1(\boldsymbol{z}, \tilde{\boldsymbol{z}}; \boldsymbol{\alpha}, \boldsymbol{\Delta}, \boldsymbol{\Lambda}) \\
= \; & \Pr((1,1)|\boldsymbol{X}_1) + \Pr((1,1)|\boldsymbol{X}_2) - \Pr((1,1)|\boldsymbol{X}_3) - \Pr((1,1)|\boldsymbol{X}_4); \\
& B_1(2(\boldsymbol{\alpha} + \boldsymbol{\Delta} + \boldsymbol{w}\boldsymbol{\Lambda}) - \boldsymbol{z}, 2(\boldsymbol{\alpha} + \boldsymbol{\Delta} + \boldsymbol{w}\boldsymbol{\Lambda}) - \tilde{\boldsymbol{z}}; \boldsymbol{\alpha}, \boldsymbol{\Delta}, \boldsymbol{\Lambda}) \\
= \; & \Pr((1,1)|\boldsymbol{X}_5) + \Pr((1,1)|\boldsymbol{X}_6) - \Pr((1,1)|\boldsymbol{X}_7) - \Pr((1,1)|\boldsymbol{X}_8).
\end{aligned}$$

By Assumption RSMH, we have

$$B_1(\boldsymbol{z}, \tilde{\boldsymbol{z}}; \boldsymbol{\alpha}, \boldsymbol{\Delta}, \boldsymbol{\Lambda}) - B_1(2(\boldsymbol{\alpha} + \boldsymbol{\Delta} + \boldsymbol{w}\boldsymbol{\Lambda}) - \boldsymbol{z}, 2(\boldsymbol{\alpha} + \boldsymbol{\Delta} + \boldsymbol{w}\boldsymbol{\Lambda}) - \tilde{\boldsymbol{z}}; \boldsymbol{\alpha}, \boldsymbol{\Delta}, \boldsymbol{\Lambda}) = 0.$$

Now, given these two fundamental identifying restrictions, we now introduce the definition of the identification.

Definition A.3. (Radial Symmetry - Discrete Response Identification) Let $\boldsymbol{a} = (a_1, a_2) \in \Theta_\alpha$, and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2) \in \Theta_{\boldsymbol{\Lambda}}$. Let

$$T(\boldsymbol{a}, \boldsymbol{\lambda}) = \left\{ (\boldsymbol{z}, \tilde{\boldsymbol{z}}) \;\middle|\; \begin{array}{l} B_0(\boldsymbol{z}, \tilde{\boldsymbol{z}}; \boldsymbol{\alpha}, \boldsymbol{\Lambda}) \neq B_0(2(\boldsymbol{a} + \boldsymbol{w}\boldsymbol{\lambda}) - \boldsymbol{z}, 2(\boldsymbol{a} + \boldsymbol{w}\boldsymbol{\lambda}) - \tilde{\boldsymbol{z}}; \boldsymbol{\alpha}, \boldsymbol{\Lambda}); \\ \boldsymbol{z}, \tilde{\boldsymbol{z}} \in \mathcal{S}_{\boldsymbol{z}|\boldsymbol{w}}, 2(\boldsymbol{a} + \boldsymbol{w}\boldsymbol{\lambda}) - \boldsymbol{z}, 2(\boldsymbol{a} + \boldsymbol{w}\boldsymbol{\lambda}) - \tilde{\boldsymbol{z}} \in \mathcal{S}_{\boldsymbol{z}|\boldsymbol{w}}. \end{array} \right\}$$

(i)We say that $(\boldsymbol{a}, \boldsymbol{\lambda})$ is RSDR identified relative to $(\boldsymbol{\alpha}, \boldsymbol{\Lambda})$ if

$$\Pr\left(\left(\boldsymbol{Z}, \tilde{\boldsymbol{Z}}\right) \in T(\boldsymbol{a}, \boldsymbol{\lambda}) | \boldsymbol{W} = \boldsymbol{w}\right) > 0.$$

(ii) in addition, we say that $(\boldsymbol{a}, \boldsymbol{\lambda})$ is RSDR point identified if for all $(\boldsymbol{a}, \boldsymbol{\lambda}) \neq (\boldsymbol{\alpha}, \boldsymbol{\Lambda})$,

$$\Pr\left(\left(\boldsymbol{Z}, \tilde{\boldsymbol{Z}}\right) \in T\left(\boldsymbol{a}, \boldsymbol{\lambda}\right) | \boldsymbol{W} = \boldsymbol{w}\right) > 0.$$

Definition A.4. (Radial Symmetry - Discrete Response Identification) Let $\boldsymbol{a} = (a_1, a_2) \in \Theta_\alpha$, $\delta = (\delta_1, \delta_2) \in \Theta_\Delta$ and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2) \in \Theta_{\boldsymbol{\Lambda}}$. Let

$$T\left(\boldsymbol{a}, \boldsymbol{\delta}, \boldsymbol{\lambda}\right) = \left\{ (\boldsymbol{z}, \tilde{\boldsymbol{z}}) \left| \begin{array}{l} B_1\left(\boldsymbol{z}, \tilde{\boldsymbol{z}}; \boldsymbol{\alpha}, \boldsymbol{\Delta}, \boldsymbol{\Lambda}\right) \\ \neq B_1\left(2\left(\boldsymbol{\alpha} + \boldsymbol{\delta} + \boldsymbol{w}\boldsymbol{\lambda}\right) - \boldsymbol{z}, 2\left(\boldsymbol{\alpha} + \boldsymbol{\delta} + \boldsymbol{w}\boldsymbol{\lambda}\right) - \tilde{\boldsymbol{z}}; \boldsymbol{\alpha}, \boldsymbol{\Delta}, \boldsymbol{\Lambda}\right); \\ \boldsymbol{z}, \tilde{\boldsymbol{z}} \in \mathcal{S}_{\boldsymbol{z}|\boldsymbol{w}}, 2\left(\boldsymbol{a} + \boldsymbol{\delta} + \boldsymbol{w}\boldsymbol{\lambda}\right) - \boldsymbol{z}, 2\left(\boldsymbol{a} + \boldsymbol{\delta} + \boldsymbol{w}\boldsymbol{\lambda}\right) - \tilde{\boldsymbol{z}} \in \mathcal{S}_{\boldsymbol{z}|\boldsymbol{w}}. \end{array} \right. \right\}$$

(i)We say that $(\boldsymbol{a}, \boldsymbol{\delta}, \boldsymbol{\lambda})$ is RSDR identified relative to $(\boldsymbol{\alpha}, \boldsymbol{\Delta}, \boldsymbol{\Lambda})$ if

$$\Pr\left(\left(\boldsymbol{Z}, \tilde{\boldsymbol{Z}}\right) \in T\left(\boldsymbol{a}, \boldsymbol{\delta}, \boldsymbol{\lambda}\right) | \boldsymbol{W} = \boldsymbol{w}\right) > 0.$$

(ii) in addition, we say that $(\boldsymbol{a}, \boldsymbol{\delta}, \boldsymbol{\lambda})$ is RSDR point identified if for all $(\boldsymbol{a}, \boldsymbol{\delta}, \boldsymbol{\lambda}) \neq (\boldsymbol{\alpha}, \boldsymbol{\Delta}, \Lambda)$,

$$\Pr\left(\left(\boldsymbol{Z}, \tilde{\boldsymbol{Z}}\right) \in T\left(\boldsymbol{a}, \boldsymbol{\delta}, \boldsymbol{\lambda}\right) | \boldsymbol{W} = \boldsymbol{w}\right) > 0.$$

Assumption SVMH (Sufficient Variation with Multivariate Covariates under Heteroskedasticity) Given any set $S \subset \mathcal{S}_{\boldsymbol{Z}|\boldsymbol{W}}$ and a vector $\boldsymbol{a} = (a_1, a_2)$, define the symmetrically reflected set

$$S'\left(S, \boldsymbol{a}\right) = \{(z_1', z_2') \ \ | (z_1', z_2') = \left(2\left(a_1 + w_1\lambda_1\right) - z_1, 2\left(a_2 + w_2\lambda_2\right) - z_2\right), (z_1, z_2) \in S \ \ \};$$

Similarly, in addition, given a vector $\boldsymbol{\delta} = (\delta_1, \delta_2)$, define the symmetrically reflected set

$$\begin{aligned} & S^{'}\left(S, \boldsymbol{a}+\boldsymbol{\delta}\right) \\ = \; & \left\{(z_1', z_2') \;\; | (z_1', z_2') = \left(2\left(a_1+\delta_1+w_1\lambda_1\right)-z_1, 2\left(a_2+\delta_2+w_2\lambda_2\right)-z_2\right), (z_1, z_2) \in S\right\}; \end{aligned}$$

Suppose that

(i) The points $(\alpha_1 + w_1\Lambda_1, \alpha_2 + w_2\Lambda_2)$ and $(\alpha_1 + \Delta_1 + w_1\Lambda_1, \alpha_2 + \Delta_2 + w_2\Lambda_2)$ are in the interior of the support $\mathcal{S}_{\boldsymbol{Z}|\boldsymbol{W}}$;

(ii) The random vector $\boldsymbol{Z} = (Z_1, Z_2)$ is absolutely continuously distributed with the positive density $f_{(Z_1,Z_2)|W}(\cdot,\cdot)$ over the support of $\mathcal{S}_{\boldsymbol{Z}|\boldsymbol{W}}$, with respect to the Lebesgue measure;

(iii) For all $\boldsymbol{a} \in \mathcal{S}_{\boldsymbol{Z}|\boldsymbol{W}}$ such that $\boldsymbol{a} \neq \boldsymbol{\alpha}$, there exists a measurable set $S \subset \mathcal{S}_{\boldsymbol{Z}|\boldsymbol{W}}$ such that $S^{'}(S, \boldsymbol{a}) \subset \mathcal{S}_{\boldsymbol{Z}|\boldsymbol{W}}$ and

$$f_{\varepsilon}(z_1, z_2) \neq f_{\varepsilon}(z_1', z_2') \text{ a.e. for } (z_1, z_2) \in S, (z_1', z_2') \in S^{'}(S, \boldsymbol{a}).$$

Moreover, for all $\boldsymbol{a} + \boldsymbol{\delta} \in \mathcal{S}_{\boldsymbol{Z}|\boldsymbol{W}}$ such that $\boldsymbol{a} + \boldsymbol{\delta} \neq \boldsymbol{\alpha} + \boldsymbol{\Delta}$, there exists a measurable set $S \subset \mathcal{S}_{\boldsymbol{Z}|\boldsymbol{W}}$ such that $S^{'}(S, \boldsymbol{a} + \boldsymbol{\delta}) \subset \mathcal{S}_{\boldsymbol{Z}}$ and

$$f_{\varepsilon}(z_1, z_2) \neq f_{\varepsilon}(z_1', z_2') \text{ a.e. for } (z_1, z_2) \in S, (z_1', z_2') \in S^{'}(S, \boldsymbol{a} + \boldsymbol{\delta}).$$

Theorem A.5. *Suppose that Assumptions RMH, S, ERMH, RSMH and SVMH hold. Then, $\boldsymbol{\alpha}, \boldsymbol{\Lambda}$ and $\boldsymbol{\Delta}$ are point identified.*

Theorem A.6. *Suppose that Assumptions RMH, S, ERMH, RSMH and SVMH-(i)(ii) hold, and the distribution of unobservables $(\varepsilon_{1i}, \varepsilon_{2i})$ is unimodal. Then, $\boldsymbol{\alpha}, \boldsymbol{\Lambda}$ and $\boldsymbol{\Delta}$ are point identified.*

A.1.2 Extension II: Random Coefficients Model

In this section, we extend our analysis to the complete information entry games with the random coefficients. In this context, we can allow for the competition effects depending on the unobserved heterogeneity, similar to Dunker, Hoderlein, and Kaido (2013). More specifically, we consider a simple entry game with random coefficients,

$$\begin{aligned} Y_{1i}^* &= -Z_{1i} + \tilde{\Delta}_{1i} Y_{2i} + \varepsilon_{1i}; \\ Y_{2i}^* &= -Z_{2i} + \tilde{\Delta}_{2i} Y_{1i} + \varepsilon_{2i}; \\ Y_{pi} &= \begin{cases} 1, & \text{if } Y_{pi}^* \geq 0 \\ 0, & \text{otherwise} \end{cases} \quad \text{for } p = 1, 2. \end{aligned}$$

To keep the following analysis simple, we still use the scalar observed characteristic and normalize the coefficients associated with the scalar observables to $(-1, -1)$. This normalization is also used in Lewbel and Tang (2013) for the incomplete information entry games with random coefficients.

Note that we can rewrite the random coefficients $\left(\tilde{\Delta}_{1i}, \tilde{\Delta}_{2i}\right)$ as follows: $\tilde{\Delta}_{1i} = \Delta_1 + \varrho_{1i}$ where $\Delta_1 \equiv \mathbb{E}\left[\tilde{\Delta}_{1i}\right]$ and $\varrho_{1i} \equiv \tilde{\Delta}_{1i} - \mathbb{E}\left[\tilde{\Delta}_{1i}\right]$; similarly, $\tilde{\Delta}_{2i} = \Delta_2 + \varrho_{2i}$ where $\Delta_2 \equiv \mathbb{E}\left[\tilde{\Delta}_{2i}\right]$ and $\varrho_{2i} \equiv \tilde{\Delta}_{2i} - \mathbb{E}\left[\tilde{\Delta}_{2i}\right]$. This formalization suggests that we can possibly transform the random coefficient models to the fixed coefficient models. We can first identify the models up to a constant mean and then trace out the part of the distributions associated with random coefficients. Here, our identification strategy for the random coefficient model (RCM) is almost the same as the one in the main context except that we apply a stronger symmetry condition, called elliptical symmetry. We will explain it in greater detail once we introduce assumptions. Given this, to present our identification strategy, we keep Assumption R in the main context, and assume that the following regularity conditions hold.

Assumption SRC (Sign in RCM) $\tilde{\Delta}_{1i} < 0, \tilde{\Delta}_{2i} < 0.$

Assumption SRC also guarantees that there exist the unique entry outcomes $(0,0)$ and $(1,1)$, similar to Assumption S.

Assumption ERRC (Exclusion Restriction in RCM) *Suppose that*

(i) (Z_{1i}, Z_{2i}) *is independent of* $(\varepsilon_{1i}, \varepsilon_{2i})$*; and* (Z_{1i}, Z_{2i}) *is independent of* $(\varrho_{1i}, \varrho_{2i})$.

(ii) the scalar covariate Z_{pi} *enters only the payoff function of player* p*, but not the payoff function of the other player.*

Assumption ERRC assumes that the excluded regressor is also independent of the unobserved heterogeneity of the random coefficients, in addition to the one that is assumed in Assumption ER in the main context.

Assumption ESRC (Elliptical Symmetry in RCM) Suppose that

(i) the distribution of (ϱ_1, ϱ_2) is elliptically symmetric, that is, $(\varrho_1, \varrho_2) \sim E_d(\mathbf{0}, \Sigma_\varrho, \phi_\varrho)$.

(ii) the distribution of $(\varepsilon_1, \varepsilon_2)$ is elliptically symmetric, that is, $(\varepsilon_1, \varepsilon_2) \sim E_d(\boldsymbol{\alpha}, \Sigma_\varepsilon, \phi_\varepsilon)$.

In addition, we assume that $\Sigma_\varrho = \Sigma_\varepsilon = \Sigma$.

Assumption ESRC assumes that both unobserved heterogeneity in the random coefficients and unobserved heterogeneity in the profit function are all elliptically symmetric. Note that in $E_d(\boldsymbol{\mu}, \Sigma, \phi)$, $\boldsymbol{\mu}$ represents the mean of a random vector, Σ represents the variance-covariance matrix (also known as dispersion matrix) (this is not necessarily equal to the variance-covariance matrix $(\varrho_{1i}, \varrho_{2i})$ (or $(\varepsilon_{1i}, \varepsilon_{2i})$), and ϕ is referred to as the characteristic generator of the corresponding random vector (for more details, see Hult and Lindskog (2002) and Fang, Kotz, and Ng (1990)).

Assumption IRC (Independence in RCM) Suppose that $(\varrho_{1i}, \varrho_{2i})$ is independent of $(\varepsilon_{1i}, \varepsilon_{2i})$.

Assumption IRC assumes that the unobservables in the profit function is independent of unobserved heterogeneity in the random coefficients. Note that Assumption

IRC can help us simplify the analysis in the lemma presented below. In fact, we can allow a certain dependence between $(\varepsilon_{1i}, \varepsilon_{2i})$ and $(\varrho_{1i}, \varrho_{2i})$, as shown in Hult and Lindskog (2002). But to keep our analysis simple, we omit the derivation and discussion here (For more details, see Hult and Lindskog (2002)).

Lemma A.7. *By Assumption ESRC and IRC, the distribution of* $(\varrho_{1i} + \varepsilon_{1i}, \varrho_{2i} + \varepsilon_{2i})$ *is elliptical symmetric, that is* $(\varrho_{1i} + \varepsilon_{1i}, \varrho_{2i} + \varepsilon_{2i}) \sim E_d(\boldsymbol{\alpha}, \Sigma, \phi^*)$ *where* $\phi^* = \phi_\varrho \phi_\varepsilon$.

Lemma A.7 provides the fundamental results that allow us to transform the identification of random coefficient models to one with fixed coefficients, as we represented in the main context of the paper. Since elliptical symmetry implies the radial symmetry, essentially, the analysis in the main context can be applied here directly. On the other hand, one can easily check that, the result in Lemma A.7 does not necessarily hold for the radial symmetry, which is why we impose a stronger symmetry condition here.

Given Lemma A.7, we can outline our identification steps for the random coefficient models. We first identify (α_1, α_2) using the conditional choice probability of $(0, 0)$, and identify (Δ_1, Δ_2) using the conditional choice probability of $(1, 1)$. Next, given the fixed coefficients have been identified, we will identify the portion of the joint distribution $(\varepsilon_{1i}, \varepsilon_{2i})$ from the conditional choice probability of $(0, 0)$. Similarly, we can identify the portion of $(\varrho_{1i} + \varepsilon_{1i}, \varrho_{2i} + \varepsilon_{2i})$ from the conditional choice probability of $(1, 1)$. Finally, given the portion of the distribution of $(\varepsilon_{1i}, \varepsilon_{2i})$ and $(\varrho_{1i} + \varepsilon_{1i}, \varrho_{2i} + \varepsilon_{2i})$, we can identify the part of joint distribution $(\varrho_{1i}, \varrho_{2i})$.

As a final remark, the analysis can be directly extended by allowing multivariate covariates with heteroskedasticity, following Appendix A.1.1. We can allow for the correlation between observables with unobserved heterogeneity in the random coefficients, and the correlation between observables with unobserved heterogeneity in the profit function, like the existing literature (e.g., Fox and Lazzati (2013), Kline (2012) and Dunker, Hoderlein, and Kaido (2013)). For brevity, we omit the discussion.

A.2 Proofs for Identification

In this appendix, we will show the proofs of theorems for identification.

Proof of Lemma I.1: Lemma I.1 can be shown by direct calculation. Without loss of generality, we assume that $z_1 < \tilde{z}_1$, $z_2 < \tilde{z}_2$. We observe that under Assumptions R, S and ER, we can write the difference of the choice probabilities as

$$\begin{aligned}
& B_0\left(\boldsymbol{z}, \tilde{\boldsymbol{z}}; \boldsymbol{\alpha}\right) \\
= \; & \Pr[(0,0)\,|\boldsymbol{Z}_1] + \Pr[(0,0)\,|\boldsymbol{Z}_2] - \Pr[(0,0)\,|\boldsymbol{Z}_3] - \Pr[(0,0)\,|\boldsymbol{Z}_4] \\
= \; & \int_{-\infty}^{z_2}\int_{-\infty}^{z_1} f\left(\varepsilon_1, \varepsilon_2\right) d\left(\varepsilon_1, \varepsilon_2\right) + \int_{-\infty}^{\tilde{z}_2}\int_{-\infty}^{\tilde{z}_1} f\left(\varepsilon_1, \varepsilon_2\right) d\left(\varepsilon_1, \varepsilon_2\right) \\
& - \int_{-\infty}^{\tilde{z}_2}\int_{-\infty}^{z_1} f\left(\varepsilon_1, \varepsilon_2\right) d\left(\varepsilon_1, \varepsilon_2\right) - \int_{-\infty}^{z_2}\int_{-\infty}^{\tilde{z}_1} f\left(\varepsilon_1, \varepsilon_2\right) d\left(\varepsilon_1, \varepsilon_2\right) \\
= \; & \int_{z_2}^{\tilde{z}_2}\int_{z_1}^{\tilde{z}_1} f\left(\varepsilon_1, \varepsilon_2\right) d\varepsilon_1 d\varepsilon_2;
\end{aligned}$$

$$\begin{aligned}
& B_0\left(2\boldsymbol{\alpha} - \boldsymbol{z}, 2\boldsymbol{\alpha} - \tilde{\boldsymbol{z}}; \boldsymbol{\alpha}\right) \\
= \; & \Pr[(0,0)\,|\boldsymbol{Z}_5] + \Pr[(0,0)\,|\boldsymbol{Z}_6] - \Pr[(0,0)\,|\boldsymbol{Z}_7] - \Pr[(0,0)\,|\boldsymbol{Z}_8] \\
= \; & \int_{2\alpha_2 - \tilde{z}_2}^{2\alpha_2 - z_2}\int_{2\alpha_1 - \tilde{z}_1}^{2\alpha_1 - z_1} f\left(\varepsilon_1, \varepsilon_2\right) d\left(\varepsilon_1, \varepsilon_2\right) \\
= \; & \int_{z_2 - 2\alpha_2}^{\tilde{z}_2 - 2\alpha_2}\int_{z_1 - 2\alpha_1}^{\tilde{z}_1 - 2\alpha_1} f\left(-\varepsilon_1, -\varepsilon_2\right) d\left(-\varepsilon_1, -\varepsilon_2\right) \\
= \; & \int_{z_2}^{\tilde{z}_2}\int_{z_1}^{\tilde{z}_1} f\left(2\alpha_1 - \varepsilon_1, 2\alpha_2 - \varepsilon_2\right) d\left(\varepsilon_1, \varepsilon_2\right);
\end{aligned}$$

where the last three equalities follow from direct calculation and the properties of integrals. Now by Assumption RS, that is, $f\left(\varepsilon_1, \varepsilon_2\right) = f\left(2\alpha_1 - \varepsilon_1, 2\alpha_2 - \varepsilon_2\right)$, the desired result follows. *Q.E.D.*

Proof of Lemma I.2: Similar to Lemma I.1, Lemma I.2 can also be shown by direct calculation. Without loss of generality, we continue to assume that $z_1 < \tilde{z}_1$, $z_2 < \tilde{z}_2$. It follows from Assumptions R, S, ER, we have

$$\begin{aligned}
& B_1(\boldsymbol{z}, \tilde{\boldsymbol{z}}; \boldsymbol{\alpha}, \boldsymbol{\Delta}) \\
= \; & \Pr[(1,1)|\boldsymbol{Z}_1] + \Pr[(1,1)|\boldsymbol{Z}_2] - \Pr[(1,1)|\boldsymbol{Z}_3] - \Pr[(1,1)|\boldsymbol{Z}_4] \\
= \; & \int_{-\infty}^{z_2-\Delta_2} \int_{-\infty}^{z_1-\Delta_1} f(\varepsilon_1, \varepsilon_2)\, d(\varepsilon_1, \varepsilon_2) + \int_{-\infty}^{\tilde{z}_2-\Delta_2} \int_{-\infty}^{\tilde{z}_1-\Delta_1} f(\varepsilon_1, \varepsilon_2)\, d(\varepsilon_1, \varepsilon_2) \\
& - \int_{-\infty}^{\tilde{z}_2-\Delta_2} \int_{-\infty}^{z_1-\Delta_1} f(\varepsilon_1, \varepsilon_2)\, d(\varepsilon_1, \varepsilon_2) - \int_{-\infty}^{z_2-\Delta_2} \int_{-\infty}^{\tilde{z}_1-\Delta_1} f(\varepsilon_1, \varepsilon_2)\, d(\varepsilon_1, \varepsilon_2) \\
= \; & \int_{z_2-\Delta_2}^{\tilde{z}_2-\Delta_2} \int_{z_1-\Delta_1}^{\tilde{z}_1-\Delta_1} f(\varepsilon_1, \varepsilon_2)\, d(\varepsilon_1, \varepsilon_2);
\end{aligned}$$

$$\begin{aligned}
& B_1(2(\boldsymbol{\alpha}+\boldsymbol{\Delta}) - \boldsymbol{z}, 2(\boldsymbol{\alpha}+\boldsymbol{\Delta}) - \tilde{\boldsymbol{z}}; \boldsymbol{\alpha}, \boldsymbol{\Delta}) \\
= \; & \Pr[(1,1)|\boldsymbol{Z}_5] + \Pr[(1,1)|\boldsymbol{Z}_6] - \Pr[(1,1)|\boldsymbol{Z}_7] - \Pr[(1,1)|\boldsymbol{Z}_8] \\
= \; & \int_{2\alpha_2-(\tilde{z}_2-\Delta_2)}^{2\alpha_2-(z_2-\Delta_2)} \int_{2\alpha_1-(\tilde{z}_1-\Delta_1)}^{2\alpha_1-(z_1-\Delta_1)} f(\varepsilon_1, \varepsilon_2)\, d(\varepsilon_1, \varepsilon_2) \\
= \; & \int_{(z_2-\Delta_2)-2\alpha_2}^{(\tilde{z}_2-\Delta_2)-2\alpha_2} \int_{(z_1-\Delta_1)-2\alpha_1}^{(\tilde{z}_1-\Delta_1)-2\alpha_1} f(-\varepsilon_1, -\varepsilon_2)\, d(\varepsilon_1, \varepsilon_2) \\
= \; & \int_{z_2-\Delta_2}^{\tilde{z}_2-\Delta_2} \int_{z_1-\Delta_1}^{\tilde{z}_1-\Delta_1} f(2\alpha_1 - \varepsilon_1, 2\alpha_2 - \varepsilon_2)\, d(\varepsilon_1, \varepsilon_2)
\end{aligned}$$

Now by Assumption RS, that is, $f(\varepsilon_1, \varepsilon_2) = f(2\alpha_1 - \varepsilon_1, 2\alpha_2 - \varepsilon_2)$, the desired result follows. *Q.E.D.*

Proof of Theorem I.5: Note that the proofs of Part (i) and (ii) are almost identical, here we only explicitly give the proof of Part (i). To prove the desired result in Part (i), that is, $\Pr\left((\boldsymbol{Z}, \tilde{\boldsymbol{Z}}) \in T(\boldsymbol{a})\right) > 0$, it is equivalent to prove for any

$\boldsymbol{a} \neq \boldsymbol{\alpha}$, $\Pr\left((\boldsymbol{Z}, \tilde{\boldsymbol{Z}}) \in T^c(\boldsymbol{a})\right) < 1$, where we denote $T^c(\boldsymbol{a})$ as the complement set of $T(\boldsymbol{a})$, which can be written as $T^c(\boldsymbol{a}) = T_1^c(\boldsymbol{a}) + T_2^c(\boldsymbol{a})$, $T_1^c(\boldsymbol{a}) \cap T_2^c(\boldsymbol{a}) = \emptyset$

$$
\begin{aligned}
T_1^c(\boldsymbol{a}) &= \left\{ (\boldsymbol{z}, \tilde{\boldsymbol{z}}) \in \mathcal{S}_{\boldsymbol{Z}} | \quad (2\boldsymbol{a} - \boldsymbol{z}), (2\boldsymbol{a} - \tilde{\boldsymbol{z}}) \in \mathcal{S}_{\boldsymbol{Z}} \text{ and } B(\boldsymbol{z}, \tilde{\boldsymbol{z}}; \boldsymbol{\alpha}) = B(2\boldsymbol{a} - \boldsymbol{z}, 2\boldsymbol{a} - \tilde{\boldsymbol{z}}; \boldsymbol{\alpha}) \right\}; \\
T_2^c(\boldsymbol{a}) &= \left\{ (\boldsymbol{z}, \tilde{\boldsymbol{z}}) \in \mathcal{S}_{\boldsymbol{Z}} | \quad (2\boldsymbol{a} - \boldsymbol{z}), (2\boldsymbol{a} - \tilde{\boldsymbol{z}}) \notin \mathcal{S}_{\boldsymbol{Z}} \right\}.
\end{aligned}
$$

In the following analysis, we will show that $\Pr\left((\boldsymbol{Z}, \tilde{\boldsymbol{Z}}) \in T^c(\boldsymbol{a})\right) = 1$ leads to a contradiction. Suppose that $\Pr\left((\boldsymbol{Z}, \tilde{\boldsymbol{Z}}) \in T^c(\boldsymbol{a})\right) = 1$. It suggests that $\Pr(T_1^c(\boldsymbol{a})) + \Pr(T_2^c(\boldsymbol{a})) = 1$ with $\Pr\left((\boldsymbol{Z}, \tilde{\boldsymbol{Z}}) \in T_1^c(\boldsymbol{a})\right) \geq 0$, $\Pr\left((\boldsymbol{Z}, \tilde{\boldsymbol{Z}}) \in T_2^c(\boldsymbol{a})\right) \geq 0$. There are three possible cases under which, the above statement will hold. In the following context, we will explicitly discuss these three possible cases.

Case 1: $\Pr\left((\boldsymbol{Z}, \tilde{\boldsymbol{Z}}) \in T_1^c(\boldsymbol{a})\right) = 0$ and $\Pr\left((\boldsymbol{Z}, \tilde{\boldsymbol{Z}}) \in T_2^c(\boldsymbol{a})\right) = 1$. This can only occur when alternative parameter $\boldsymbol{a}$ is outside the support of $\mathcal{S}_{\boldsymbol{z}}$, which directly contradicts Assumption SV-(i), that is, the support of regressors contains the parameter space.

Case 2: $\Pr\left((\boldsymbol{Z}, \tilde{\boldsymbol{Z}}) \in T_1^c(\boldsymbol{a})\right) > 0$, $\Pr\left((\boldsymbol{Z}, \tilde{\boldsymbol{Z}}) \in T_2^c(\boldsymbol{a})\right) > 0$, $\Pr(T_1^c(\boldsymbol{a})) + \Pr(T_2^c(\boldsymbol{a})) = 1$. The strategy to show the contradiction in Case 2 is as follows: (a) we recall that for all $\boldsymbol{z}, \tilde{\boldsymbol{z}} \in T_1^c(\boldsymbol{a})$, the difference of choice probabilities is equal to zero; (b) we fixed the $\tilde{\boldsymbol{z}}$ at some fixed values in a small neighborhood, we take the derivatives with respect to z_1 and z_2, respectively; (c) we show that $\Pr(T_1^c(\boldsymbol{a})) + \Pr(T_2^c(\boldsymbol{a})) = 1$ with $\Pr\left((\boldsymbol{Z}, \tilde{\boldsymbol{Z}}) \in T_1^c(\boldsymbol{a})\right) > 0$, $\Pr\left((\boldsymbol{Z}, \tilde{\boldsymbol{Z}}) \in T_2^c(\boldsymbol{a})\right) > 0$ will lead to a contradiction.

Step (a): consider $\boldsymbol{z}, \tilde{\boldsymbol{z}} \in \mathcal{S}_{\boldsymbol{Z}}$, such that $B(\boldsymbol{z}, \tilde{\boldsymbol{z}}; \boldsymbol{\alpha}) - B(2\boldsymbol{a} - \boldsymbol{z}, 2\boldsymbol{a} - \tilde{\boldsymbol{z}}; \boldsymbol{\alpha}) = 0$,

$(2\boldsymbol{a}-\boldsymbol{z})\in\mathcal{S}_{\boldsymbol{Z}}$, $(2\boldsymbol{a}-\tilde{\boldsymbol{z}})\in\mathcal{S}_{\boldsymbol{Z}}$, which can be written as

$$\begin{aligned} B_0\left(\boldsymbol{z},\tilde{\boldsymbol{z}};\boldsymbol{\alpha}\right) &= \int_{z_1}^{\tilde{z}_1}\int_{z_2}^{\tilde{z}_2} f\left(\varepsilon_1,\varepsilon_2\right)d\left(\varepsilon_1,\varepsilon_2\right); \\ B_0\left(2\boldsymbol{a}-\boldsymbol{z},2\boldsymbol{a}-\tilde{\boldsymbol{z}};\boldsymbol{\alpha}\right) &= \int_{2a_1-z_1}^{2a_1-\tilde{z}_1}\int_{2a_2-z_2}^{2a_2-\tilde{z}_2} f\left(\varepsilon_1,\varepsilon_2\right)d\left(\varepsilon_1,\varepsilon_2\right). \end{aligned}$$

Step (b): now, by Assumption SV-(ii), for any $\tilde{\boldsymbol{z}}^*=(z_1^*,z_2^*)$ in the neighborhood $\mathcal{N}\left(\tilde{\boldsymbol{z}}_0^*,\epsilon_0\right)$ with some arbitrarily small $\epsilon_0>0$, where $\tilde{\boldsymbol{z}}_0^*\in\mathcal{S}_{\boldsymbol{Z}}$ and $2\boldsymbol{a}-\tilde{\boldsymbol{z}}_0^*\in\mathcal{S}_{\boldsymbol{Z}}$, we can take the derivative with respect to z_1 and z_2 on both sides of equations,

$$\begin{aligned} \frac{\partial^2 B_0\left(\boldsymbol{z},\tilde{\boldsymbol{z}}^*;\boldsymbol{\alpha}\right)}{\partial z_1\partial z_2} &= f_\varepsilon(z_1,z_2); \\ \frac{\partial^2 B_0\left(2\boldsymbol{a}-\boldsymbol{z},2\boldsymbol{a}-\tilde{\boldsymbol{z}}^*;\boldsymbol{\alpha}\right)}{\partial z_1\partial z_2} &= f_\varepsilon\left(2a_1-z_1,2a_2-z_2\right). \end{aligned}$$

Step (c): $\Pr\left(T_1^c\left(\boldsymbol{a}\right)\right)+\Pr\left(T_2^c\left(\boldsymbol{a}\right)\right)=1$ with two cases $\Pr\left(\left(\boldsymbol{Z},\tilde{\boldsymbol{Z}}\right)\in T_1^c\left(\boldsymbol{a}\right)\right)>0$, $\Pr\left(\left(\boldsymbol{Z},\tilde{\boldsymbol{Z}}\right)\in T_2^c\left(\boldsymbol{a}\right)\right)>0$, implies that given the neighborhood $\mathcal{N}\left(\tilde{\boldsymbol{z}}_0^*,\epsilon_0\right)$, by SV-(ii), for any Lebesgue measurable set $S\subset S_{\boldsymbol{Z}}$ with positive measure such that $S'\left(S,\boldsymbol{a}\right)\subset S_{\boldsymbol{Z}}$,

$$f_\varepsilon(z_1,z_2)=f_\varepsilon\left(2a_1-z_1,2a_2-z_2\right)\ \text{a.e. for}\ \left(z_1,z_2\right)\in S;$$

that is, given arbitrary values $\boldsymbol{a}$, the densities are the same for all $(z_1,z_2)\in S$ and for all measurable sets, which contradicts Assumption SV-(iii).

Case 3: $\Pr\left(\left(\boldsymbol{Z},\tilde{\boldsymbol{Z}}\right)\in T_1^c\left(\boldsymbol{a}\right)\right)=1$, $\Pr\left(\left(\boldsymbol{Z},\tilde{\boldsymbol{Z}}\right)\in T_2^c\left(\boldsymbol{a}\right)\right)=0$. This is an extreme example of Case 2 and the proof can follow the proof of Case 2 directly. *Q.E.D.*

Proof of Theorem I.6: The proof of Theorem I.6 is similar to the proof of Theorem I.5 . Similar to Theorem I.5, we follow to define $T^c\left(\boldsymbol{a}\right)=T_1^c\left(\boldsymbol{a}\right)+T_2^c\left(\boldsymbol{a}\right)$, $T_1^c\left(\boldsymbol{a}\right)\cap T_2^c\left(\boldsymbol{a}\right)=\emptyset$. To show the desired result, it is also equivalent to show that $\Pr\left(T^c\left(\boldsymbol{a}\right)\right)<1$. Here we will show it by contradiction. Suppose that for any $\boldsymbol{a}\neq\boldsymbol{\alpha}$, $\Pr\left(T^c\left(\boldsymbol{a}\right)\right)=1$. Now, given that the distribution of the unobservables is unimodal,

it directly implies that $\Pr\left(T_1^c\left(\boldsymbol{a}\right)\right) = 0$. It follows that $\Pr\left(T^c\left(\boldsymbol{a}\right)\right) = 1$ implies $\Pr\left(T_2^c\left(\boldsymbol{a}\right)\right) = 1$. This becomes the same as Case 1 in Theorem I.5. By the same argument used in Case 1 of the proof for Theorem I.5, $\Pr\left(T_2^c\left(\boldsymbol{a}\right)\right) = 1$ contradicts Assumption SV(i) directly, which gives the desired result. *Q.E.D.*

A.3 Proofs for Estimation

This appendix collects the proofs for the theorems relating to the properties of the estimator presented in Section 1.3. Throughout this appendix, following the definition of the choice probabilities in the main context, we can define the kernel estimators of these conditional choice probabilities. These conditional choice probabilities and their kernel estimators are the building blocks of our estimator. Note that only the choice probabilities for $\upsilon = 5, \ldots, 8$ explicitly contain the parameters of interest. This implies that the derivatives of $\sum_{\upsilon=1}^{8} \kappa_\upsilon \varphi_\upsilon \left(\boldsymbol{z}, \tilde{\boldsymbol{z}}, \boldsymbol{\theta}\right)$ and their kernel estimators depend only on the derivatives of the choice probabilities for $\upsilon = 5, \ldots, 8$. This fact is critical for understanding some properties of our estimator.

Next, we will use the following theorems to prove the results. Let MI refer to Markov's inequality, CSI to the Cauchy-Schwarz inequality, DCT to the Dominated Convergence Theorem, LLN to Khintchine's law of large numbers and CLT to the Lindberg-Levy central limit theorem. Let o and O denote a sequence of the real numbers and o_p and O_p denote the order in probability of a sequence of random variables. Moreover, for simplicity, we will use $\sum_{i \neq j}$ to abbreviate $\sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n}$, use $\sum_{i \neq j \neq k}$ to abbreviate $\sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \sum_{k=1, k \neq i \neq j}^{n}$. We will write $\mathbb{E}_{[i]} = \int dF\left(x_i\right)$ and $\mathbb{E}_{[i,j]} = \int dF\left(x_i\right) dF\left(x_j\right)$ without further explanations.

In addition, because $\sum_{\upsilon=1}^{8} \kappa_\upsilon \varphi_\upsilon \left(\boldsymbol{Z}_i, \boldsymbol{Z}_j, \boldsymbol{\theta}^0\right) = 0$, for all $\boldsymbol{Z}_i, \boldsymbol{Z}_j$, then any term that contains $\sum_{\upsilon=1}^{8} \kappa_\upsilon \varphi_\upsilon \left(\boldsymbol{Z}_i, \boldsymbol{Z}_j, \boldsymbol{\theta}^0\right)$ will be equal to zero. We summarize some properties of these terms at below. We use these results directly in the proof without additional explanation.

$$
\begin{aligned}
\frac{1}{n\left(n-1\right)} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \left(\sum_{\upsilon=1}^{8} \kappa_\upsilon \varphi_\upsilon \left(\boldsymbol{Z}_i, \boldsymbol{Z}_j, \boldsymbol{\theta}^0\right) \right) h(\cdot) &= 0; \\
\frac{1}{n-1} \sum_{j=1, i \neq j}^{n} \left(\sum_{\upsilon=1}^{8} \kappa_\upsilon \varphi_\upsilon \left(\boldsymbol{z}_i, \boldsymbol{Z}_j, \boldsymbol{\theta}^0\right) \right) h(\cdot) &= 0; \\
\mathbb{E}\left[\left(\sum_{\upsilon=1}^{8} \kappa_\upsilon \varphi_\upsilon \left(\boldsymbol{Z}_i, \boldsymbol{Z}_j, \boldsymbol{\theta}^0\right) \right) h(\cdot) \right] &= 0;
\end{aligned}
$$

where $h(\cdot)$ can be any arbitrary function. For notational convenience, we write $\zeta(\boldsymbol{z}_i, \boldsymbol{z}_j, \boldsymbol{\theta}) = \tau(\boldsymbol{z}_i, \boldsymbol{z}_j, \boldsymbol{\theta}) \sum_{v=1}^{8} \kappa_v \varphi_v(\boldsymbol{z}_i, \boldsymbol{z}_j, \boldsymbol{\theta})$ in the following context and supplementary appendix.

Moreover, we consider two trimming functions. The first trimming component is

$$\tau_{ij}(\boldsymbol{\theta}) = \left(\Pi_{v=1}^{8} \tau_v(\boldsymbol{z}_i, \boldsymbol{z}_j, \boldsymbol{\theta})\right)^{1/8}.$$

This trimming component deals with the problem of the boundary bias. In addition, we introduce a second trimming component, which takes the form of[1]

$$G_{ij} = G\left(\min_{v} \hat{\varphi}_v(\boldsymbol{z}_i, \boldsymbol{z}_j, \boldsymbol{\theta})\right),$$

where G is a smooth trimming function similar to the one used in Linton and Xiao (2001). This trimming component helps us deal with the issue of the estimated choice probabilities that are below zero or above one by using a higher-order kernel function. For expositional purposes, we derive the property of the objective function $Q_n(\boldsymbol{\theta})$ with the first trimming component as follows

$$Q_n(\boldsymbol{\theta}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \tau_{ij}(\boldsymbol{\theta}) \left[\sum_{v=1}^{8} \kappa_v \hat{\varphi}_v\left(\boldsymbol{Z}_i, \tilde{\boldsymbol{Z}}_j, \boldsymbol{\theta}\right)\right]^2.$$

Our main asymptotic results will be based on $Q_n(\boldsymbol{\theta})$, ignoring the second trimming component for a while. In the end, we will treat this second trimming component separately and show that it is asymptotically negligible and does not affect our asymptotic results as we derive for $Q_n(\boldsymbol{\theta})$.

In summary, the rest of appendix is organized as follows. Appendix A.3.1 con-

[1]Note that we only consider an additional trimming on the estimated choice probability to accommodate the usage of higher-order kernel. In fact, the higher order kernel may possibly affect the estimated density of observables. We omit this for simplicity's sake but we are aware that the additional trimming can help in the estimation and the additional possible trimming also deserves more discussion in the higher-order MSE approximation.

tributes identification of parameters. Appendix A.3.2 is devoted to the consistency of the estimator. Appendix A.3.3 proves that the estimator is $\sqrt{n}$-consistent and asymptotically normal. Appendix A.3.4 describes the mean squared error approximation of the estimator, which provides the basis for the bandwidth selection in this paper. Appendix A.3.5 shows the effect of trimming. We collect the proofs of Lemmas associated with Appendix A.3 in the supplementary material Appendix S.C to save space. Throughout Appendix A.3, we assume that Assumptions R, S, ER, RS and SV hold.

A.3.1 Identification (Population Objective Function)

Note that as we show in Appendix S.C.1, following Assumption TR, we restrict the calculation of choice probability at the interior of $\mathcal{S}_{\boldsymbol{Z}}$, $\mathcal{S}_{\boldsymbol{Z}}^o$. By doing so, we remove a possible boundary bias problem from the kernel regression estimator (for more details, see Imbens and Ridder (2009)) and guarantee the asymptotic property of the kernel regression estimator for the choice probability. More specifically, we consider $\tau(\boldsymbol{\theta})$ restricts that $(\boldsymbol{z}, \tilde{\boldsymbol{z}})$ be in $\mathcal{S}_{\boldsymbol{Z}}^o$ in the population objective function. This is valid under our identification assumption as long as we allow that the interior of support is also wider than the parameter space.

Proof of Theorem I.11: It is straightforward to verify (i) holds from the quadratic form of the population objective function. To show (ii), we first show existence, which requires that we verify that $Q(\boldsymbol{\theta})$ achieves its minimum at $\boldsymbol{\theta}^0$. This can be directly obtained from Lemma I.1 and Lemma I.2 and Theorem I.5. Next, we show uniqueness, that is, $Q(\boldsymbol{\theta}^*) > Q\left(\boldsymbol{\theta}^0\right) = 0$ for all $\boldsymbol{\theta}^* \neq \boldsymbol{\theta}^0$. For simplicity, we suppress the dependence on $\boldsymbol{Z}, \tilde{\boldsymbol{Z}}$ and write $\varphi_v(\boldsymbol{\theta}) = \varphi_v\left(\boldsymbol{Z}, \tilde{\boldsymbol{Z}}, \boldsymbol{\theta}\right)$, and $\tau(\boldsymbol{\theta}) = \tau(\boldsymbol{z}, \tilde{\boldsymbol{z}}, \boldsymbol{\theta})$.

Observe that

$$\begin{aligned}
& Q\left(\boldsymbol{\theta}^{*}\right)-Q\left(\boldsymbol{\theta}^{0}\right) \\
= & \mathbb{E}\left[\tau\left(\boldsymbol{\theta}^{*}\right)\sum\nolimits_{\upsilon=1}^{8}\kappa_{\upsilon}\varphi_{\upsilon}\left(\boldsymbol{\theta}^{*}\right)\right]^{2}-\mathbb{E}\left[\tau\left(\boldsymbol{\theta}^{0}\right)\sum\nolimits_{\upsilon=1}^{8}\kappa_{\upsilon}\varphi_{\upsilon}\left(\boldsymbol{\theta}^{0}\right)\right]^{2} \\
= & 2\mathbb{E}\left[\tau\left(\boldsymbol{\theta}^{0}\right)\sum\nolimits_{\upsilon=1}^{8}\kappa_{\upsilon}\varphi_{\upsilon}\left(\boldsymbol{\theta}^{0}\right)\right] \\
& \times\mathbb{E}\left[\tau\left(\boldsymbol{\theta}^{*}\right)\sum\nolimits_{\upsilon=1}^{8}\kappa_{\upsilon}\varphi_{\upsilon}\left(\boldsymbol{\theta}^{*}\right)-\tau\left(\boldsymbol{\theta}^{0}\right)\sum\nolimits_{\upsilon=1}^{8}\kappa_{\upsilon}\varphi_{\upsilon}\left(\boldsymbol{\theta}^{0}\right)\right] \qquad \text{(A.1)} \\
& +\mathbb{E}\left[\tau\left(\boldsymbol{\theta}^{*}\right)\sum\nolimits_{\upsilon=1}^{8}\kappa_{\upsilon}\varphi_{\upsilon}\left(\boldsymbol{\theta}^{*}\right)-\tau\left(\boldsymbol{\theta}^{0}\right)\sum\nolimits_{\upsilon=1}^{8}\kappa_{\upsilon}\varphi_{\upsilon}\left(\boldsymbol{\theta}^{0}\right)\right]^{2}. \qquad \text{(A.2)}
\end{aligned}$$

In the following, we will show the terms (A.1) and (A.2), respectively.

For term (A.1), we know that $\tau\left(\boldsymbol{\theta}^{0}\right)\sum_{\upsilon=1}^{8}\kappa_{\upsilon}\varphi_{\upsilon}\left(\boldsymbol{z},\tilde{\boldsymbol{z}},\boldsymbol{\theta}^{0}\right)=0$ for all $\boldsymbol{z},\tilde{\boldsymbol{z}}$, so term (A.1) is zero.

For term (A.2), we first define the set $T\left(\boldsymbol{\theta}^{*}\right)$ as follows (similar to Definition I.3 and Definition I.4)

$$T\left(\boldsymbol{\theta}^{*}\right)=\left\{\left(\boldsymbol{z},\tilde{\boldsymbol{z}}\right)\in\mathcal{S}_{\boldsymbol{Z}}^{o}|\sum\nolimits_{\upsilon=1}^{8}\kappa_{\upsilon}\varphi_{\upsilon}\left(\boldsymbol{z},\tilde{\boldsymbol{z}},\boldsymbol{\theta}^{*}\right)\neq0;2\boldsymbol{\theta}^{*}-\boldsymbol{z},2\boldsymbol{\theta}^{*}-\tilde{\boldsymbol{z}}\in\mathcal{S}_{\boldsymbol{Z}}^{o}\right\};$$

and in an analogous way, we define its complementary sets $T^{c}\left(\theta^{*}\right)=T_{1}^{c}\left(\boldsymbol{\theta}^{*}\right)\cup T_{2}^{c}\left(\boldsymbol{\theta}^{*}\right)$ and $T_{1}^{c}\left(\boldsymbol{\theta}^{*}\right)\cap T_{2}^{c}\left(\boldsymbol{\theta}^{*}\right)=\emptyset$, which are defined as

$$\begin{aligned}
T_{1}^{c}\left(\boldsymbol{\theta}^{*}\right) &= \left\{\left(\boldsymbol{z},\tilde{\boldsymbol{z}}\right)\in\mathcal{S}_{\boldsymbol{Z}}^{o}|\sum\nolimits_{\upsilon=1}^{8}\kappa_{\upsilon}\varphi_{\upsilon}\left(\boldsymbol{z},\tilde{\boldsymbol{z}},\boldsymbol{\theta}^{*}\right)=0;2\boldsymbol{\theta}^{*}-\boldsymbol{z},2\boldsymbol{\theta}^{*}-\tilde{\boldsymbol{z}}\in\mathcal{S}_{\boldsymbol{Z}}^{o}\right\}; \\
T_{2}^{c}\left(\boldsymbol{\theta}^{*}\right) &= \left\{\left(\boldsymbol{z},\tilde{\boldsymbol{z}}\right)\in\mathcal{S}_{\boldsymbol{Z}}^{o}|\,2\boldsymbol{\theta}^{*}-\boldsymbol{z},2\boldsymbol{\theta}^{*}-\tilde{\boldsymbol{z}}\notin\mathcal{S}_{\boldsymbol{Z}}^{o}\right\}.
\end{aligned}$$

Next, we decompose the expectation in term (A.2) into three integral regions

$$\begin{aligned}
& \mathbb{E}\left[\tau\left(\boldsymbol{\theta}^{*}\right)\textstyle\sum_{\upsilon=1}^{8}\kappa_{\upsilon}\varphi_{\upsilon}\left(\boldsymbol{Z},\tilde{\boldsymbol{Z}},\boldsymbol{\theta}^{*}\right)-\tau\left(\boldsymbol{\theta}^{0}\right)\sum_{\upsilon=1}^{8}\kappa_{\upsilon}\varphi_{\upsilon}(\boldsymbol{Z},\tilde{\boldsymbol{Z}},\boldsymbol{\theta}^{0}))\right]^{2} \\
= & \textstyle\int_{T\left(\boldsymbol{\theta}^{*}\right)}\left[\tau\left(\boldsymbol{\theta}^{*}\right)\sum_{\upsilon=1}^{8}\kappa_{\upsilon}\varphi_{\upsilon}\left(\boldsymbol{r},\boldsymbol{s},\boldsymbol{\theta}^{*}\right)-\tau\left(\boldsymbol{\theta}^{0}\right)\sum_{\upsilon=1}^{8}\kappa_{\upsilon}\varphi_{\upsilon}\left(\boldsymbol{r},\boldsymbol{s},\boldsymbol{\theta}^{0}\right)\right]^{2}dF\left(\boldsymbol{r},\boldsymbol{s}\right) \quad \text{(A.3)} \\
+ & \textstyle\int_{T_{1}^{c}\left(\boldsymbol{\theta}^{*}\right)}\left[\tau\left(\boldsymbol{\theta}^{*}\right)\sum_{\upsilon=1}^{8}\kappa_{\upsilon}\varphi_{\upsilon}\left(\boldsymbol{r},\boldsymbol{s},\boldsymbol{\theta}^{*}\right)-\tau\left(\boldsymbol{\theta}^{0}\right)\sum_{\upsilon=1}^{8}\kappa_{\upsilon}\varphi_{\upsilon}\left(\boldsymbol{r},\boldsymbol{s},\boldsymbol{\theta}^{0}\right)\right]^{2}dF\left(\boldsymbol{r},\boldsymbol{s}\right) \quad \text{(A.4)} \\
+ & \textstyle\int_{T_{2}^{c}\left(\boldsymbol{\theta}^{*}\right)}\left[\tau\left(\boldsymbol{\theta}^{*}\right)\sum_{\upsilon=1}^{8}\kappa_{\upsilon}\varphi_{\upsilon}\left(\boldsymbol{r},\boldsymbol{s},\boldsymbol{\theta}^{*}\right)-\tau\left(\boldsymbol{\theta}^{0}\right)\sum_{\upsilon=1}^{8}\kappa_{\upsilon}\varphi_{\upsilon}\left(\boldsymbol{r},\boldsymbol{s},\boldsymbol{\theta}^{0}\right)\right]^{2}dF\left(\boldsymbol{r},\boldsymbol{s}\right). \quad \text{(A.5)}
\end{aligned}$$

For term (A.3), we observe that for all $(\boldsymbol{r}, \boldsymbol{s}) \in T(\boldsymbol{\theta}^*)$, we have $\tau(\boldsymbol{\theta}^*) \neq 0$, $\sum_{v=1}^{8} \kappa_v \varphi_v(\boldsymbol{r}, \boldsymbol{s}, \boldsymbol{\theta}^*) \neq 0$; $\tau(\boldsymbol{\theta}^0) = 0$ or $\sum_{v=1}^{8} \kappa_v \varphi_v(\boldsymbol{r}, \boldsymbol{s}, \boldsymbol{\theta}^0) = 0$. It implies that, for the integral region $T(\boldsymbol{\theta}^*)$, the integrand is strictly positive, that is

$$\left[\tau(\boldsymbol{\theta}^*) \sum_{v=1}^{8} \kappa_v \varphi_v(\boldsymbol{r}, \boldsymbol{s}, \boldsymbol{\theta}^*) - \tau(\boldsymbol{\theta}^0) \sum_{v=1}^{8} \kappa_v \varphi_v(\boldsymbol{r}, \boldsymbol{s}, \boldsymbol{\theta}^0)\right]^2 > 0.$$

In addition, from Theorem I.5, we show that $\Pr(T(\boldsymbol{\theta}^*)) > 0$ (that is, the model parameters are identified). Hence, term (A.3) is strictly positive. In addition, for term (A.4), we know that similar to the proof of Theorem 2.1, for all $(\boldsymbol{r}, \boldsymbol{s}) \in T_1^c(\boldsymbol{\theta}^*)$, we have $\tau(\boldsymbol{\theta}^*) \neq 0$ but $\sum_{v=1}^{8} \kappa_v \varphi_v(\boldsymbol{r}, \boldsymbol{s}, \boldsymbol{\theta}^*) = 0$ and $\tau(\boldsymbol{\theta}^0) \sum_{v=1}^{8} \kappa_v \varphi_v(\boldsymbol{r}, \boldsymbol{s}, \boldsymbol{\theta}^0) = 0$, which means term (A.4) equals zero. Furthermore, for term (A.5), for all $(\boldsymbol{r}, \boldsymbol{s}) \in T_2^c(\boldsymbol{\theta}^*)$, we have $\tau(\boldsymbol{\theta}^*) = 0$ and though $\sum_{v=1}^{8} \kappa_v \varphi_v(\boldsymbol{r}, \boldsymbol{s}, \boldsymbol{\theta}^*)$ is undefined, we have that $\tau(\boldsymbol{\theta}^0) \sum_{v=1}^{8} \kappa_v \varphi_v(\boldsymbol{r}, \boldsymbol{s}, \boldsymbol{\theta}^0) = 0$. Then, term (A.5) is equal to zero. These three results immediately implies that, for all $\boldsymbol{\theta}^* \neq \boldsymbol{\theta}^0$,

$$\mathbb{E}\left[\tau(\boldsymbol{\theta}^*) \sum_{v=1}^{8} \kappa_v \left(\varphi_v(\boldsymbol{r}, \boldsymbol{s}, \boldsymbol{\theta}^*) - \tau(\boldsymbol{\theta}^0) \sum_{v=1}^{8} \kappa_v \varphi_v(\boldsymbol{r}, \boldsymbol{s}, \boldsymbol{\theta}^0)\right)\right]^2 > 0.$$

Therefore, (A.1) equals to zero while (A.2) is strictly positive giving the desired result. *Q.E.D.*

A.3.2 Consistency

Below, we show the consistency of the estimator.

Proof of Theorem I.14: To show the consistency, we will apply Theorem 2.1 from Newey and McFadden (1994) (also see Theorem A-1 in Andrews (1994) for a similar condition), which is standard in M-estimation and requires that the following conditions hold: (A1) $Q(\boldsymbol{\theta})$ is uniquely minimized at $\boldsymbol{\theta}^0$; (A2) the parameter space Θ is compact; (A3) $Q(\boldsymbol{\theta})$ is continuous; and (A4) $Q_n(\boldsymbol{\theta})$ converges uniformly in

probability to $Q(\boldsymbol{\theta})$.

Condition (A1) holds from Theorem I.11. Condition (A2) is satisfied by construction of the parameter space Θ in Assumption 1. Condition (A3) is straightforward to verify from the continuity of the quadratic function and the choice probabilities φ_υ. For Condition (A4), following Hong and Tamer (2003), we first introduce an infeasible sample objective function $\bar{Q}(\boldsymbol{\theta})$,

$$\bar{Q}_n(\boldsymbol{\theta}) = \frac{1}{n(n-1)} \sum\nolimits_{i \neq j} \tau_{ij}(\boldsymbol{\theta}) \left[\sum\nolimits_{\upsilon=1}^{8} \kappa_\upsilon \varphi_\upsilon(\boldsymbol{z}_i, \boldsymbol{z}_j, \boldsymbol{\theta}) \right]^2.$$

Then by the triangle inequality, it follows that

$$\left|Q_n(\boldsymbol{\theta}) - Q(\boldsymbol{\theta})\right| \leq \left|Q_n(\boldsymbol{\theta}) - \bar{Q}_n(\boldsymbol{\theta})\right| + \left|\bar{Q}_n(\boldsymbol{\theta}) - Q(\boldsymbol{\theta})\right|;$$

so that it is sufficient to show that (i) $\sup_{\boldsymbol{\theta}\in\Theta} \left|Q_n(\boldsymbol{\theta}) - \bar{Q}_n(\boldsymbol{\theta})\right| = o_p(1)$, and in addition, (ii) $\sup_{\boldsymbol{\theta}\in\Theta} \left|\bar{Q}_n(\boldsymbol{\theta}) - Q(\boldsymbol{\theta})\right| = o_p(1)$. We will discuss these two results sequentially.

First, consider (i). We observe that

$$\begin{aligned}
& \sup_{\boldsymbol{\theta}\in\Theta} \left|Q_n(\boldsymbol{\theta}) - \bar{Q}_n(\boldsymbol{\theta})\right| \\
= & \sup_{\boldsymbol{\theta}\in\Theta} \left| \frac{1}{n(n-1)} \sum\nolimits_{i \neq j} \tau_{ij} \left(\left[\sum\nolimits_{\upsilon=1}^{8} \kappa_\upsilon \hat{\varphi}_\upsilon(\boldsymbol{z}_i, \boldsymbol{z}_j, \boldsymbol{\theta}) \right]^2 - \left[\sum\nolimits_{\upsilon=1}^{8} \kappa_\upsilon \varphi_\upsilon(\boldsymbol{z}_i, \boldsymbol{z}_j, \boldsymbol{\theta}) \right]^2 \right) \right| \\
\leq & \frac{1}{n(n-1)} \sum\nolimits_{i \neq j} \tau_{ij} \sup_{\boldsymbol{\theta}\in\Theta} \left| \left[\sum\nolimits_{\upsilon=1}^{8} \kappa_\upsilon \hat{\varphi}_\upsilon(\boldsymbol{z}_i, \boldsymbol{z}_j, \boldsymbol{\theta}) \right]^2 - \left[\sum\nolimits_{\upsilon=1}^{8} \kappa_\upsilon \varphi_\upsilon(\boldsymbol{z}_i, \boldsymbol{z}_j, \boldsymbol{\theta}) \right]^2 \right| \\
\leq & \sup_{\boldsymbol{z}_i, \boldsymbol{z}_j \in \mathcal{S}_{\boldsymbol{Z}}^o} \sup_{\boldsymbol{\theta}\in\Theta} \left| \left[\sum\nolimits_{\upsilon=1}^{8} \kappa_\upsilon \hat{\varphi}_\upsilon(\boldsymbol{z}_i, \boldsymbol{z}_j, \boldsymbol{\theta}) \right]^2 - \left[\sum\nolimits_{\upsilon=1}^{8} \kappa_\upsilon \varphi_\upsilon(\boldsymbol{z}_i, \boldsymbol{z}_j, \boldsymbol{\theta}) \right]^2 \right| \\
\leq & \sup_{\boldsymbol{z}_i, \boldsymbol{z}_j \in \mathcal{S}_{\boldsymbol{Z}}^o} \sup_{\boldsymbol{\theta}\in\Theta} C \left| \sum\nolimits_{\upsilon=1}^{8} \kappa_\upsilon \left(\hat{\varphi}_{\upsilon,ij}(\boldsymbol{z}_i, \boldsymbol{z}_j, \boldsymbol{\theta}) - \varphi_{\upsilon,ij}(\boldsymbol{z}_i, \boldsymbol{z}_j, \boldsymbol{\theta}) \right) \right| \\
\leq & C \sum\nolimits_{\upsilon=1}^{8} \sup_{\boldsymbol{z}_i, \boldsymbol{z}_j \in \mathcal{S}_{\boldsymbol{Z}}^o} \sup_{\boldsymbol{\theta}\in\Theta} \left| \hat{\varphi}_{\upsilon,ij}(\boldsymbol{z}_i, \boldsymbol{z}_j, \boldsymbol{\theta}) - \varphi_{\upsilon,ij}(\boldsymbol{z}_i, \boldsymbol{z}_j, \boldsymbol{\theta}) \right| \\
= & o_p(1);
\end{aligned}$$

where the first inequality hold from the triangle inequality and the supremum of the sum is less than the sum of supremum; and the second inequality by the property of supremum; the third inequality follows by the fact that uniformly over $\boldsymbol{z}_i, \boldsymbol{z}_j \in \mathcal{S}_{\boldsymbol{Z}}^o$ and $\boldsymbol{\theta} \in \Theta$,

$$\begin{aligned}
& \left[\sum_{v=1}^{8} \kappa_v \hat{\varphi}_{v,ij}(\boldsymbol{\theta})\right]^2 - \left[\sum_{v=1}^{8} \kappa_v \varphi_{v,ij}(\boldsymbol{\theta})\right]^2 \\
= & \left(\sum_{v=1}^{8} \kappa_v \left(\hat{\varphi}_{v,ij}(\boldsymbol{\theta}) + \varphi_{v,ij}(\boldsymbol{\theta})\right)\right)\left(\sum_{v=1}^{8} \left(\hat{\varphi}_{v,ij}(\boldsymbol{\theta}) - \varphi_{v,ij}(\boldsymbol{\theta})\right)\right) \\
\leq & \left(2\sum_{v=1}^{8} \left|\varphi_{v,ij}(\boldsymbol{\theta})\right| + o_p(1)\right)\left(\sum_{v=1}^{8} \left(\hat{\varphi}_{v,ij}(\boldsymbol{\theta}) - \varphi_{v,ij}(\boldsymbol{\theta})\right)\right) \\
\leq & C\left(\sum_{v=1}^{8} \left(\hat{\varphi}_{v,ij}(\boldsymbol{\theta}) - \varphi_{v,ij}(\boldsymbol{\theta})\right)\right);
\end{aligned}$$

following from Lemma I.13 and Assumption 2 with the bounded $\varphi_{v,ij}(\boldsymbol{\theta})$;[2] the fourth inequality can be obtained by applying the triangle inequality again; the last equality directly follows from Lemma I.13.

Next, we will show that (ii) holds, that is, $\sup_{\theta\in\Theta}\left|\bar{Q}_n(\boldsymbol{\theta}) - Q(\boldsymbol{\theta})\right|$. By the LLN (following from Theorem A in Section 5.4, Serfling (1980)), we directly obtain the pointwise convergence of $\bar{Q}_n(\boldsymbol{\theta})$, $\bar{Q}_n(\boldsymbol{\theta}) = Q(\boldsymbol{\theta}) + o_p(1)$. Then we can conclude the uniformity by showing stochastic equicontinuity, $\sup_{\theta,\tilde{\theta}\in\Theta,\left|\theta-\tilde{\theta}\right|<\epsilon}\left|\bar{Q}_n(\boldsymbol{\theta}) - \bar{Q}_n\left(\tilde{\boldsymbol{\theta}}\right)\right| = o_p(1)$. Following Andrews (1994), the stochastic equicontinuity can be shown by verifying that $\bar{Q}_n(\boldsymbol{\theta})$ is in the type II class of function, that is, the function satisfies the Lipschitz condition, $\left|\bar{Q}_n(\boldsymbol{\theta}) - \bar{Q}_n\left(\tilde{\boldsymbol{\theta}}\right)\right| \leq C\left\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\right\|$. It is straightforward to verify that it holds from the continuity of the quadratic form of the objective function and the continuity of the conditional choice probability with bounded first derivative.

Therefore, combining all the results above gives the desired result. *Q.E.D.*

[2]Note that in Lemma I.13 we show the uniformity using the arguments $(z_{v,1}, z_{v,2})$. As we use the generic expression on the choice probabilities, alternatively, we can show the uniformity instead using the arguments $(\boldsymbol{z}, \tilde{\boldsymbol{z}}, \boldsymbol{\theta})$ as we use here.

A.3.3 Root–n Consistency and Asymptotic Normality

To show $\sqrt{n}$-consistency and asymptotic normality of the estimator, it is important to examine properties of the objective functions. For this purpose, Appendix A.3.3.1 collects lemmas devoted to showing key properties of the objective function. Then, in Appendix A.3.3.2, we follow to show $\sqrt{n}$-consistency and asymptotic normality of the estimator in the main context.

Recall that the sample objective function $Q_n(\boldsymbol{\theta})$ is

$$Q_n(\boldsymbol{\theta}) = \frac{1}{n(n-1)} \sum\nolimits_{i \neq j} \tau_{ij}(\boldsymbol{\theta}) \left[\sum\nolimits_{v=1}^{8} \hat{\varphi}_v(\boldsymbol{Z}_i, \boldsymbol{Z}_j, \boldsymbol{\theta}) \right]^2.$$

To ease notation, we suppress the dependence on $\boldsymbol{Z}, \tilde{\boldsymbol{Z}}$ and let $\varphi_{v,ij}(\boldsymbol{\theta})$ abbreviate $\varphi_v(\boldsymbol{Z}_i, \boldsymbol{Z}_j, \boldsymbol{\theta})$ and $\hat{\varphi}_{v,ij}(\boldsymbol{\theta})$ abbreviate $\hat{\varphi}_v(\boldsymbol{Z}_i, \boldsymbol{Z}_j, \boldsymbol{\theta})$. Then, following Sherman (1994), we decompose $Q_n(\boldsymbol{\theta})$ as follows,

$$\begin{aligned} Q_n(\boldsymbol{\theta}) &= \frac{1}{n(n-1)} \sum\nolimits_{i \neq j} \tau_{ij}(\boldsymbol{\theta}) \left[\sum\nolimits_{v=1}^{8} \kappa_v \varphi_{v,ij}(\boldsymbol{\theta}) \right]^2 \\ &\quad + \frac{2}{n(n-1)} \sum\nolimits_{i \neq j} \tau_{ij}(\boldsymbol{\theta}) \left[\sum\nolimits_{v=1}^{8} \kappa_v \varphi_{v,ij}(\boldsymbol{\theta}) \right] \left[\sum\nolimits_{v=1}^{8} \kappa_v \left(\hat{\varphi}_{v,ij}(\boldsymbol{\theta}) - \varphi_{v,ij}(\boldsymbol{\theta}) \right) \right] \\ &\quad + \frac{1}{n(n-1)} \sum\nolimits_{i \neq j} \tau_{ij}(\boldsymbol{\theta}) \left[\sum\nolimits_{v=1}^{8} \kappa_v \left(\hat{\varphi}_{v,ij}(\boldsymbol{\theta}) - \varphi_{v,ij}(\boldsymbol{\theta}) \right) \right]^2 \\ &= Q_{n,1}(\boldsymbol{\theta}) + Q_{n,2}(\boldsymbol{\theta}) + Q_{n,3}(\boldsymbol{\theta}). \end{aligned}$$

In an analogous way, we decompose $Q_n\left(\boldsymbol{\theta}^0\right)$ following the same steps. It follows that

$$Q_n(\boldsymbol{\theta}) - Q_n\left(\boldsymbol{\theta}^0\right) = Q_{n,1}(\boldsymbol{\theta}) - Q_{n,1}\left(\boldsymbol{\theta}^0\right) + Q_{n,2}(\boldsymbol{\theta}) - Q_{n,1}\left(\boldsymbol{\theta}^0\right) + Q_{n,3}(\boldsymbol{\theta}) - Q_{n,3}\left(\boldsymbol{\theta}^0\right)$$

where

$$
\begin{aligned}
Q_{n,1}(\boldsymbol{\theta}) - Q_{n,1}(\boldsymbol{\theta}^0) &= (n(n-1))^{-1} \sum\nolimits_{i \neq j} \tau_{ij}(\boldsymbol{\theta}) \left[\sum\nolimits_{v=1}^{8} \kappa_v \varphi_{v,ij}(\boldsymbol{\theta}) \right]^2 \\
&\quad - (n(n-1))^{-1} \sum\nolimits_{i \neq j} \tau_{ij}(\boldsymbol{\theta}^0) \left[\sum\nolimits_{v=1}^{8} \kappa_v \varphi_{v,ij}(\boldsymbol{\theta}^0) \right]^2;
\end{aligned}
$$

$$
\begin{aligned}
Q_{n,2}(\boldsymbol{\theta}) - Q_{n,2}(\boldsymbol{\theta}^0) &= 2(n(n-1))^{-1} \sum\nolimits_{i \neq j} \tau_{ij}(\boldsymbol{\theta}) \left[\sum\nolimits_{v=1}^{8} \kappa_v \varphi_{v,ij}(\boldsymbol{\theta}) \right] \\
&\quad \times \left[\sum\nolimits_{v=1}^{8} \kappa_v \left(\hat{\varphi}_{v,ij}(\boldsymbol{\theta}) - \varphi_{v,ij}(\boldsymbol{\theta}) \right) \right] \\
&\quad -2(n(n-1))^{-1} \sum\nolimits_{i \neq j} \tau_{ij}(\boldsymbol{\theta}^0) \left[\sum\nolimits_{v=1}^{8} \kappa_v \varphi_{v,ij}(\boldsymbol{\theta}^0) \right] \\
&\quad \times \left[\sum\nolimits_{v=1}^{8} \kappa_v \left(\hat{\varphi}_{v,ij}(\boldsymbol{\theta}^0) - \varphi_{v,ij}(\boldsymbol{\theta}^0) \right) \right];
\end{aligned}
$$

$$
\begin{aligned}
Q_{n,3}(\boldsymbol{\theta}) - Q_{n,3}(\boldsymbol{\theta}^0) &= (n(n-1))^{-1} \sum\nolimits_{i \neq j} \tau_{ij}(\boldsymbol{\theta}) \left[\sum\nolimits_{v=1}^{8} \kappa_v \left(\hat{\varphi}_{v,ij}(\boldsymbol{\theta}) - \varphi_{v,ij}(\boldsymbol{\theta}) \right) \right]^2 \\
&\quad - (n(n-1))^{-1} \sum\nolimits_{i \neq j} \tau_{ij}(\boldsymbol{\theta}^0) \left[\sum\nolimits_{v=1}^{8} \kappa_v \left(\hat{\varphi}_{v,ij}(\boldsymbol{\theta}^0) - \varphi_{v,ij}(\boldsymbol{\theta}^0) \right) \right]^2;
\end{aligned}
$$

As will be shown below, $Q_{n,1}(\boldsymbol{\theta}) - Q_{n,1}(\boldsymbol{\theta}^0)$ will be attributed to the Hessian matrix; $Q_{n,2}(\boldsymbol{\theta}) - Q_{n,2}(\boldsymbol{\theta}^0)$ will be devoted to the key components of the asymptotic normality of the estimator; and $Q_{n,3}(\boldsymbol{\theta}) - Q_{n,3}(\boldsymbol{\theta}^0)$ will become asymptotically negligible. In the following context, we will examine each of these three terms in turn.

A.3.3.1 Lemmas and Propositions

We begin with $Q_{n,1}(\boldsymbol{\theta}) - Q_{n,1}(\boldsymbol{\theta}^0)$, which will gives the Hessian matrix in our analysis.

Proposition A.8. *Suppose that Assumptions R, S, ER, RS, SV. Then, (i)* $Q_{n,1}(\boldsymbol{\theta}) = Q_1(\boldsymbol{\theta}) + o_p(1)$ *and* $Q_{n,1}(\boldsymbol{\theta}^0) = Q_1(\boldsymbol{\theta}^0) + o_p(1)$*; (ii) uniformly over* $O_p(\varsigma_n)$ *neighborhood of* $\boldsymbol{\theta}^0$

$$
Q_{n,1}(\boldsymbol{\theta}) - Q_{n,1}(\boldsymbol{\theta}^0) = \frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}^0)' \Gamma (\boldsymbol{\theta} - \boldsymbol{\theta}^0) + o_p\left(\|\boldsymbol{\theta} - \boldsymbol{\theta}^0\|^2 \right)
$$

where $\Gamma = \mathbb{E}\left[\left(\tau_{ij}\left(\boldsymbol{\theta}^0\right)\sum_{\upsilon=5}^{8}\kappa_\upsilon\nabla_\theta\varphi_{\upsilon,ij}\left(\boldsymbol{\theta}^0\right)\right)\left(\tau_{ij}\left(\boldsymbol{\theta}^0\right)\sum_{\upsilon=5}^{8}\kappa_\upsilon\nabla_\theta\varphi_{\upsilon,ij}\left(\boldsymbol{\theta}^0\right)\right)'\right]$.

Proof of Proposition A.8: We find that result (i) directly follows from Theorem A in Section 5.4 of Serfling (1980). That is, by the symmetry of the linear combination of $\sum_{\upsilon=1}^{8}\kappa_\upsilon\varphi_{\upsilon,ij}(\boldsymbol{\theta})$ and $\mathbb{E}\left|\sum_{\upsilon=1}^{8}\kappa_\upsilon\varphi_{\upsilon,ij}(\boldsymbol{\theta})\right| < \infty$, $Q_{n,1}(\boldsymbol{\theta}) = Q_1(\boldsymbol{\theta}) + o_p(1)$ and likewise, for $Q_{n,1}\left(\boldsymbol{\theta}^0\right)$. Hence, the following analysis focuses on result (ii). We start by expanding $Q_{n,1}(\boldsymbol{\theta})$ around $\boldsymbol{\theta}^0$ up to the second derivative. Uniformly over an $O_p(\varsigma_n)$ neighborhood of $\boldsymbol{\theta}^0$,

$$
\begin{aligned}
& Q_{n,1}(\boldsymbol{\theta}) - Q_{n,1}\left(\boldsymbol{\theta}^0\right) \\
= \;& \left(n(n-1)\right)^{-1}\sum\nolimits_{i\neq j}\tau_{ij}(\boldsymbol{\theta})\left[\sum\nolimits_{\upsilon=1}^{8}\kappa_\upsilon\varphi_{\upsilon,ij}(\boldsymbol{\theta})\right] \\
& -\left(n(n-1)\right)^{-1}\sum\nolimits_{i\neq j}\tau_{ij}\left(\boldsymbol{\theta}^0\right)\left[\sum\nolimits_{\upsilon=1}^{8}\kappa_\upsilon\varphi_{\upsilon,ij}\left(\boldsymbol{\theta}^0\right)\right] \\
= \;& \left(\boldsymbol{\theta}-\boldsymbol{\theta}^0\right)'\left[\nabla_\theta\left(n(n-1)\right)^{-1}\sum\nolimits_{i\neq j}\tau_{ij}(\boldsymbol{\theta})\left[\sum\nolimits_{\upsilon=1}^{8}\kappa_\upsilon\varphi_{\upsilon,ij}(\boldsymbol{\theta})\right]^2\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0}\right] \\
& +\frac{1}{2}\left(\boldsymbol{\theta}-\boldsymbol{\theta}^0\right)'\left[\nabla_{\theta\theta}\left(n(n-1)\right)^{-1}\sum\nolimits_{i\neq j}\tau_{ij}(\boldsymbol{\theta})\left[\sum\nolimits_{\upsilon=1}^{8}\kappa_\upsilon\varphi_{\upsilon,ij}(\boldsymbol{\theta})\right]^2\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0}\right]\left(\boldsymbol{\theta}-\boldsymbol{\theta}^0\right) \\
& +o_p\left(\left\|\boldsymbol{\theta}-\boldsymbol{\theta}^0\right\|^2\right) \\
= \;& \left(\boldsymbol{\theta}-\boldsymbol{\theta}^0\right)'R_n + \left(\boldsymbol{\theta}-\boldsymbol{\theta}^0\right)'\Gamma_n\left(\boldsymbol{\theta}-\boldsymbol{\theta}^0\right)' + o_p\left(\left\|\boldsymbol{\theta}-\boldsymbol{\theta}^0\right\|^2\right);
\end{aligned}
$$

where the last equality follows by writing R_n as the first derivative and Γ_n as the second derivative; in addition, for all $\boldsymbol{\theta}$ in $o_p(1)$-neighborhood of $\boldsymbol{\theta}^0$, the bounded third derivative and LLN gives that the third term. This suggests that we need to show (i) $R_n = o_p(1)$ and (ii) $\Gamma_n = \Gamma + o_p(1)$ to obtain the desired result.

First, we notice that

$$
\begin{aligned}
R_n &= \nabla_\theta\left(n(n-1)\right)^{-1}\sum\nolimits_{i\neq j}\tau_{ij}(\boldsymbol{\theta})\left[\sum\nolimits_{\upsilon=1}^{8}\kappa_\upsilon\varphi_{\upsilon,ij}(\boldsymbol{\theta})\right]^2\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0} \\
&= 2\left(n(n-1)\right)^{-1}\sum\nolimits_{i\neq j}\tau_{ij}\left(\boldsymbol{\theta}^0\right)\sum\nolimits_{\upsilon=1}^{8}\kappa_\upsilon\varphi_{\upsilon,ij}\left(\boldsymbol{\theta}^0\right)\left[\sum\nolimits_{\upsilon=1}^{8}\kappa_\upsilon\nabla_\theta\varphi_{\upsilon,ij}\left(\boldsymbol{\theta}^0\right)\right] \\
&\quad +\left(n(n-1)\right)^{-1}\sum\nolimits_{i\neq j}\nabla_\theta\tau_{ij}\left(\boldsymbol{\theta}^0\right)\left[\sum\nolimits_{\upsilon=1}^{8}\kappa_\upsilon\varphi_{\upsilon,ij}\left(\boldsymbol{\theta}^0\right)\right]^2 \\
&= R_{n,1} + R_{n,2}.
\end{aligned}
$$

Note that for $R_{n,1}$, we have

$$\begin{aligned}
\mathbb{E}\left[\tau_{ij}\left(\boldsymbol{\theta}^0\right)\sum_{\upsilon=1}^{8}\kappa_\upsilon\varphi_{\upsilon,ij}\left(\boldsymbol{\theta}^0\right)\left(\sum_{\upsilon=1}^{8}\kappa_\upsilon\nabla_\theta\varphi_{\upsilon,ij}\left(\boldsymbol{\theta}^0\right)\right)\right] &= 0; \\
\mathbb{V}\left[\tau_{ij}\left(\boldsymbol{\theta}^0\right)\sum_{\upsilon=1}^{8}\kappa_\upsilon\varphi_{\upsilon,ij}\left(\boldsymbol{\theta}^0\right)\left(\sum_{\upsilon=1}^{8}\kappa_\upsilon\nabla_\theta\varphi_{\upsilon,ij}\left(\boldsymbol{\theta}^0\right)\right)\right] &= 0;
\end{aligned}$$

following from the property of the term containing $\sum_{\upsilon=1}^{8}\kappa_\upsilon\varphi_{\upsilon,ij}\left(\boldsymbol{\theta}^0\right)$ from the identifying restriction. Then, applying MI, we can show $R_{n,1}=o_p\left(1\right)$. By the same reasoning, we can similarly show that $R_{n,2}=o_p\left(1\right)$ as well.

Next, we show that $\Gamma_n=\Gamma+o_p\left(1\right)$. The direct calculation implies that

$$\begin{aligned}
\Gamma_n &= \left(n\left(n-1\right)\right)^{-1}\sum\nolimits_{i\neq j}\left[\tau_{ij}\left(\boldsymbol{\theta}^0\right)\left[\sum_{\upsilon=1}^{8}\kappa_\upsilon\nabla_\theta\varphi_{\upsilon,ij}\left(\boldsymbol{\theta}^0\right)\sum_{\upsilon=1}^{8}\kappa_\upsilon\nabla_\theta\varphi_{\upsilon,ij}\left(\boldsymbol{\theta}^0\right)\right]'\right] \\
&\quad+\left(n\left(n-1\right)\right)^{-1}\sum\nolimits_{i\neq j}\left[\tau_{ij}\left(\boldsymbol{\theta}^0\right)\sum_{\upsilon=1}^{8}\kappa_\upsilon\varphi_{\upsilon,ij}\left(\boldsymbol{\theta}^0\right)\sum_{\upsilon=1}^{8}\kappa_\upsilon\nabla_{\theta\theta}\varphi_{\upsilon,ij}\left(\boldsymbol{\theta}^0\right)\right] \\
&\quad+\left(n\left(n-1\right)\right)^{-1}\sum\nolimits_{i\neq j}\left[\nabla_\theta\tau_{ij}\left(\boldsymbol{\theta}^0\right)\sum_{\upsilon=1}^{8}\kappa_\upsilon\varphi_{\upsilon,ij}\left(\boldsymbol{\theta}^0\right)\left[\sum_{\upsilon=1}^{8}\kappa_\upsilon\nabla_\theta\varphi_{\upsilon,ij}\left(\boldsymbol{\theta}^0\right)\right]\right] \\
&\quad+\left(n\left(n-1\right)\right)^{-1}\sum\nolimits_{i\neq j}\left[\nabla_{\theta\theta}\tau_{ij}\left(\boldsymbol{\theta}^0\right)\left[\sum_{\upsilon=1}^{8}\kappa_\upsilon\varphi_{\upsilon,ij}\left(\boldsymbol{\theta}^0\right)\right]^2\right] \\
&= \Gamma_{n,1}+\Gamma_{n,2}+\Gamma_{n,3}+\Gamma_{n,4}.
\end{aligned}$$

Similar to the argument used for R_n, we can show that $\Gamma_{n,2}=o_p\left(1\right)$ by the fact that

$$\begin{aligned}
\mathbb{E}\left[\tau_{ij}\left(\boldsymbol{\theta}^0\right)\sum_{\upsilon=1}^{8}\kappa_\upsilon\varphi_{\upsilon,ij}\left(\boldsymbol{\theta}^0\right)\sum_{\upsilon=1}^{8}\kappa_\upsilon\nabla_{\theta\theta}\varphi_{\upsilon,ij}\left(\boldsymbol{\theta}^0\right)\right] &= 0; \\
\mathbb{V}\left[\tau_{ij}\left(\boldsymbol{\theta}^0\right)\sum_{\upsilon=1}^{8}\kappa_\upsilon\varphi_{\upsilon,ij}\left(\boldsymbol{\theta}^0\right)\sum_{\upsilon=1}^{8}\kappa_\upsilon\nabla_{\theta\theta}\varphi_{\upsilon,ij}\left(\boldsymbol{\theta}^0\right)\right] &= 0.
\end{aligned}$$

Similarly, we can show that $\Gamma_{n,3}=o_p\left(1\right)$ and $\Gamma_{n,4}=o_p\left(1\right)$. Finally, denoting

$$\Gamma=\mathbb{E}\left[\tau_{ij}\left(\boldsymbol{\theta}^0\right)\sum_{\upsilon=5}^{8}\kappa_\upsilon\nabla_\theta\varphi_{\upsilon,ij}\left(\boldsymbol{\theta}^0\right)\left[\sum_{\upsilon=5}^{8}\kappa_\upsilon\nabla_\theta\varphi_{\upsilon,ij}\left(\boldsymbol{\theta}^0\right)\right]'\right].$$

and applying MI implies that $\Gamma_{n,1}=\Gamma+o_p\left(1\right)$. Combining these two parts gives the desired result. *Q.E.D.*

Second, we focus on the term $Q_{n,2}(\boldsymbol{\theta}) - Q_{n,2}(\boldsymbol{\theta}^0)$. The key is to decompose $Q_{n,2}(\boldsymbol{\theta})$ and $Q_{n,2}(\boldsymbol{\theta}^0)$ into different U-statistics. Note that

$$Q_{n,2}(\boldsymbol{\theta}) = 2\left(n(n-1)\right)^{-1} \sum\nolimits_{i\neq j} \tau_{ij} \left[\sum\nolimits_{\upsilon=1}^{8} \kappa_\upsilon \varphi_{\upsilon,ij}(\boldsymbol{\theta})\right] \left[\sum\nolimits_{\upsilon=1}^{8} \kappa_\upsilon \left(\hat{\varphi}_{\upsilon,ij}(\boldsymbol{\theta}) - \varphi_{\upsilon,ij}(\boldsymbol{\theta})\right)\right].$$

Following Newey and McFadden (1994); Cattaneo, Crump, and Jansson (2013), we observe that the quadratic expansion of $\hat{\varphi}_\upsilon(\boldsymbol{\theta})$ can be written as below[3]. We expand the denominator up to cubic terms, dropping the subscript i, j in $\hat{\varphi}_{\upsilon,ij}$ and $\varphi_{\upsilon,ij}$, for simplicity.

$$\begin{aligned}
\hat{\varphi}_\upsilon - \varphi_\upsilon &= \frac{\hat{g}_\upsilon - \varphi_\upsilon \hat{f}_\upsilon}{\hat{f}_\upsilon} = \left[\hat{g}_\upsilon - \varphi_\upsilon \hat{f}_\upsilon\right] f_\upsilon^{-1} \left[1 - f_\upsilon^{-1}\left(\hat{f}_\upsilon - f_\upsilon\right) + o_p\left(f^{-1}\left(\hat{f}_\upsilon - f_\upsilon\right)\right)^2\right] \\
&= f_\upsilon^{-1}\left(\hat{g}_\upsilon - \varphi_\upsilon \hat{f}_\upsilon\right) - f_\upsilon^{-2}\left(\hat{f}_\upsilon - f_\upsilon\right)\left(\hat{g}_\upsilon - \varphi_\upsilon \hat{f}_\upsilon\right)
\end{aligned}$$

where the linear component and quadratic component can be written as

$$f_\upsilon^{-1}\left(\hat{g}_\upsilon - \varphi_\upsilon \hat{f}_\upsilon\right) = (n-2)^{-1} \sum\nolimits_{k=1,k\neq j\neq i}^{n} (d_k - \varphi_\upsilon) K_{n,\upsilon}/f_\upsilon \quad \text{(A.6)}$$

$$\begin{aligned}
f_\upsilon^{-2}\left(\hat{g}_\upsilon - \varphi_\upsilon \hat{f}_\upsilon\right)\left(\hat{f}_\upsilon - f_\upsilon\right) &= (n-2)^{-2} \sum\nolimits_{l=1,l\neq j\neq i}^{n} (d_l - \varphi_\upsilon) K_{n,\upsilon k}^2/f_\upsilon^2 \quad \text{(A.7)} \\
&\quad + (n-2)^{-2} \sum\nolimits_{k\neq l} (d_l - \varphi_\upsilon) K_{n,\upsilon k} K_{n,\upsilon l}/f_\upsilon^2 \\
&\quad - (n-2)^{-2} \sum\nolimits_{k=1,k\neq j\neq i}^{n} (d_k - \varphi_\upsilon) K_{n,\upsilon k}/f_\upsilon
\end{aligned}$$

Note that the quadratic component includes three terms which we will explain in turn. First, the third term in (A.7) is identical to the linear component in (A.6) except for the additional scale $(n-2)^{-1}$, suggesting that the third term in (A.6) will converge to zero faster than the linear component (A.6). Second, the second term in (A.7) is the cross-product term. Due to the i.i.d. sample, we can show that the second term

[3]The quadratic expansion of $\hat{a}/\hat{b}$ around a/b can be written as:

$$\hat{a}/\hat{b} - a/b = b^{-1}[\hat{a} - a - (a/b)(\hat{b} - b)] - b^{-2}(\hat{b} - b)[\hat{a} - a - (a/b)(\hat{b} - b)$$

asymptotically has the same order as the third term. By the same argument, we also expect that it will converge to zero faster than the linear component in (A.6). By virtue of these two facts, in the following context, we focus on term (A.6) and the first component of term (A.7).

Introduce $q_{n,v}(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta})$ to denote term (A.6) in $Q_{n,2}(\boldsymbol{\theta})$,

$$q_{n,v}(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta}) = \tau_{ij}(\boldsymbol{\theta}) \left[\sum\nolimits_{v=1}^{8} \kappa_v \varphi_{v,ij}(\boldsymbol{\theta})\right] (d_k - \varphi_{v,ij}(\boldsymbol{\theta})) K_{n,vk}/f_v(\boldsymbol{\theta}),$$

where $\omega_i = (Z_{1i}, Z_{2i})$, $\omega_j = (Z_{1j}, Z_{2j})$, $\omega_k = (Z_{1k}, Z_{2k}, d_k)$. We write linear combinations of $q_{n,v}$ as $q_n(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta}) = \sum_{v=1}^{8} \kappa_v q_{n,v}(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta})$. In addition, we introduce $\rho_{n,v}(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta})$ to denote the leading component of term (A.7) in $Q_{n,2}(\boldsymbol{\theta})$,

$$\rho_{n,v}(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta}) = \tau_{ij}(\boldsymbol{\theta}) \left[\sum\nolimits_{v=1}^{8} \kappa_v \varphi_{v,ij}(\boldsymbol{\theta})\right] (d_k - \varphi_{v,ij}(\boldsymbol{\theta})) K_{n,vk}^2/f_{v,ij}(\boldsymbol{\theta});$$

and similarly, let ρ_n denote linear combinations of $\rho_{n,v}$.

Finally, let $\mathbb{U}_n^3$ be the random probability measure that puts mass $(n(n-1)(n-2))^{-1}$ on each 3-tuple observation, and let $\mathbb{U}_n^4$ be the random probability measure that puts mass $\left(n(n-1)(n-2)^2\right)^{-1}$ on each 3-tuple observation as well. With this notation in place, we can write $Q_{n,2}(\boldsymbol{\theta})$ as

$$\begin{aligned} Q_{n,2}(\boldsymbol{\theta}) &= 2\left[\mathbb{U}_n^3 q_n(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta}) + \mathbb{U}_n^4 \rho_n(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta})\right] + o_p(1) \\ &= 2\sum\nolimits_{v=1}^{8} \kappa_v \left[\mathbb{U}_n^3 q_{n,v}(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta}) + \mathbb{U}_n^4 \rho_{n,v}(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta})\right] + o_p(1); \end{aligned}$$

In addition, we note that $\mathbb{U}_n^3 q_{n,v}(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta})$ can be further decomposed by applying

Hoeffding decomposition, for $\upsilon = 1, \ldots, 8$

$$\begin{aligned}
\mathbb{U}_n^3 q_{n,\upsilon}(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta}) &= \mathbb{E}\left[q_{n,\upsilon}(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta})\right] \\
&\quad + \mathcal{L}_{n,\upsilon}(\omega_i, \boldsymbol{\theta}) + \mathcal{L}_{n,\upsilon}(\omega_i, \boldsymbol{\theta}) + \mathcal{L}_{n,\upsilon}(\omega_j, \boldsymbol{\theta}) \\
&\quad + \mathcal{W}_{n,\upsilon}(\omega_i, \omega_j, \boldsymbol{\theta}) + \mathcal{W}_{n,\upsilon}(\omega_i, \omega_j, \boldsymbol{\theta}) + \mathcal{W}_{n,\upsilon}(\omega_j, \omega_k, \boldsymbol{\theta}) \\
&\quad + \mathcal{T}_{n,\upsilon}(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta})
\end{aligned}$$

where $\mathcal{L}_{n,\upsilon}(\cdot, \boldsymbol{\theta})$, $\mathcal{W}_{n,\upsilon}(\cdot, \cdot, \boldsymbol{\theta})$ and $\mathcal{T}_{n,\upsilon}(\cdot, \cdot, \cdot, \boldsymbol{\theta})$, respectively, correspond to the linear terms, quadratic terms and cubic terms in Hoeffding decomposition. (More details can be found in Appendix S.C.2.) We can decompose $\mathbb{U}_n^4 \rho_{n,\upsilon}(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta})$ in the same fashion.

The following two lemmas respectively provide asymptotic approximations of U-statistics $\mathbb{U}_n^3 q_{n,\upsilon}(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta}) - \mathbb{U}_n^3 q_{n,\upsilon}\left(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta}^0\right)$ and $\mathbb{U}_n^4 \rho_{n,\upsilon}(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta}) - \mathbb{U}_n^4 \rho_{n,\upsilon}\left(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta}^0\right)$, which are the building blocks of asymptotic approximation for $Q_{n,2}(\boldsymbol{\theta}) - Q_{n,2}\left(\boldsymbol{\theta}^0\right)$. In particular, in the following decomposition, we use $\boldsymbol{L}^{(1)}, \boldsymbol{W}^{(1)}$ and $\boldsymbol{T}^{(1)}$ to denote the first derivative of $\mathcal{L}$, $\mathcal{W}$, and $\mathcal{T}$. Note that the boldface symbol emphasizes that we deal with a vector of the first derivatives.

To start, we show the property of $\mathbb{U}_n^3 q_{n,\upsilon}(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta}) - \mathbb{U}_n^3 q_{n,\upsilon}\left(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta}^0\right)$ in Lemma .

Lemma A.9. *Given Assumptions 2-3 hold, uniformly over $O_p(\varsigma_n)$ neighborhood of $\boldsymbol{\theta}^0$*

$$\begin{aligned}
&\mathbb{U}_n^3 q_{n,\upsilon}(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta}) - \mathbb{U}_n^3 q_{n,\upsilon}\left(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta}^0\right) \\
&= \left(\boldsymbol{\theta} - \boldsymbol{\theta}^0\right)' \mathbb{E}\left(q_{n,\upsilon}^{(1)}\left(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta}^0\right)\right) \\
&+ \left(\boldsymbol{\theta} - \boldsymbol{\theta}^0\right)' \left(\boldsymbol{L}_{n,\upsilon i}^{(1)} + \boldsymbol{L}_{n,\upsilon j}^{(1)} + \boldsymbol{L}_{n,\upsilon k}^{(1)}\right) \\
&+ \left(\boldsymbol{\theta} - \boldsymbol{\theta}^0\right)' \left(\boldsymbol{W}_{n,\upsilon ij}^{(1)} + \boldsymbol{W}_{n,\upsilon ik}^{(1)} + \boldsymbol{W}_{n,\upsilon jk}^{(1)}\right) \\
&+ \left(\boldsymbol{\theta} - \boldsymbol{\theta}^0\right)' \boldsymbol{T}_{n,\upsilon ijk}^{(1)} + o_p\left(\left\|\boldsymbol{\theta} - \boldsymbol{\theta}^0\right\|^2\right)
\end{aligned}$$

where the order of each term can be shown as follows

(i) $\mathbb{E}q_{n,\upsilon}^{(1)}\left(\omega_i,\omega_j,\omega_k,\boldsymbol{\theta}^0\right)=q_{\upsilon}^{(1)}\left(\boldsymbol{\theta}^0\right)=O\left(h^{\iota}\right)$ *;*

(ii) $\boldsymbol{L}_{n,\upsilon i}^{(1)}=O_p\left(n^{-1/2}h^{\iota}\right)$ *,* $\boldsymbol{L}_{n,\upsilon j}^{(1)}=O_p\left(n^{-1/2}h^{\iota}\right)$ *and* $\boldsymbol{L}_{n,\upsilon k}^{(1)}=O_p\left(n^{-1/2}\right)$ *;*

(iii) for $\upsilon=1,2,5,6$*,*

$\boldsymbol{W}_{n,\upsilon ij}^{(1)}=O_p\left(n^{-1}h^{\iota}\right)$ *,* $\boldsymbol{W}_{n,\upsilon ik}^{(1)}=O_p\left(n^{-1}h^{-1}\right)$ *and* $\boldsymbol{W}_{n,\upsilon jk}^{(1)}=O_p\left(n^{-1}\right)$ *;*

(iv) for $\upsilon=3,4,7,8$*,*

$\boldsymbol{W}_{n,\upsilon ij}^{(1)}=O_p\left(n^{-1}h^{\iota}\right)$ *,* $\boldsymbol{W}_{n,\upsilon ik}^{(1)}=O_p\left(n^{-1}h^{-1/2}\right)$ *and* $\boldsymbol{W}_{n,\upsilon jk}^{(1)}=O_p\left(n^{-1}h^{-1/2}\right)$ *;*

(v) $\boldsymbol{T}_{n,\upsilon ijk}^{(1)}=O_p\left(n^{-3/2}h^{-1}\right)$ *.*

The proof of Lemma A.9 is shown in Appendix S.C.2. In this lemma, we first use the Hoeffding decomposition to decompose $\mathbb{U}_n^3 q_{n,\upsilon}\left(\omega_i,\omega_j,\omega_k,\boldsymbol{\theta}\right)$ and $\mathbb{U}_n^3 q_{n,\upsilon}\left(\omega_i,\omega_j,\omega_k,\boldsymbol{\theta}^0\right)$ into the means, linear terms, quadratic terms and cubic terms, respectively. Then, we expand each term around $\boldsymbol{\theta}^0$, which gives the results as shown above. The results in Lemma A.9 play an important role in proving Proposition A.11 below. First, we observe that the leading terms are the linear terms $\boldsymbol{L}_{n,\upsilon k}^{(1)}$ in the Hoeffding decomposition with order of $O_p\left(n^{-1/2}\right)$, for $\upsilon=1,\cdots,8$. These terms will be used to derive the asymptotic linear representation and asymptotic normality. Second, we notice that the first-order bias is carried in the terms $\mathbb{E}q_{n,\upsilon}^{(1)}\left(\omega_i,\omega_j,\omega_k,\theta^0\right)$ with order of $O_p\left(h^{\iota}\right)$, for $\upsilon=1,\cdots,8$, which will be addressed by the higher-order mean squared error approximation. Finally, the rest of terms will vanish in limit, because they decay faster than the leading term $\boldsymbol{L}_{n,\upsilon k}^{(1)}$ for $\upsilon=1,\cdots,8$.

Next, we show the property of $\mathbb{U}_n^4\rho_{n,\upsilon}\left(\omega_i,\omega_j,\omega_k,\boldsymbol{\theta}\right)-\mathbb{U}_n^4\rho_{n,\upsilon}\left(\omega_i,\omega_j,\omega_k,\boldsymbol{\theta}^0\right)$.

Lemma A.10. *Given that Assumptions 2-3 hold, uniformly over* $O_p\left(\varsigma_n\right)$*neighborhood*

of $\boldsymbol{\theta}^0$

$$\begin{aligned}
&\mathbb{U}_n^4 \rho_{n,\upsilon}\left(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta}\right) - \mathbb{U}_n^4 \rho_{n,\upsilon}\left(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta}^0\right) \\
&= \left(\boldsymbol{\theta} - \boldsymbol{\theta}^0\right)' \mathbb{E}\left(\rho_{n,\upsilon}^{(1)}\left(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta}^0\right)\right) \\
&+ \left(\boldsymbol{\theta} - \boldsymbol{\theta}^0\right)' \left(\boldsymbol{L}_{\rho n,\upsilon i}^{(1)} + \boldsymbol{L}_{\rho n,\upsilon j}^{(1)} + \boldsymbol{L}_{\rho n,\upsilon k}^{(1)}\right) \\
&+ \left(\boldsymbol{\theta} - \boldsymbol{\theta}^0\right)' \left(\boldsymbol{W}_{\rho n,\upsilon ij}^{(1)} + \boldsymbol{W}_{n,\upsilon ik}^{(1)} + \boldsymbol{W}_{\rho n,\upsilon jk}^{(1)}\right) \\
&+ \left(\boldsymbol{\theta} - \boldsymbol{\theta}^0\right)' \boldsymbol{T}_{\rho n,\upsilon ijk}^{(1)} + o_p\left(\left\|\boldsymbol{\theta} - \boldsymbol{\theta}^0\right\|^2\right)
\end{aligned}$$

where the order of each term is shown as follows

(i) $\mathbb{E}\rho_{n,\upsilon}^{(1)}\left(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta}^0\right) = \rho_{\upsilon}^{(1)}\left(\boldsymbol{\theta}^0\right) = O\left(n^{-1}h^{\iota-2}\right)$;

(ii) $\boldsymbol{L}_{\rho n,\upsilon i}^{(1)} = O_p\left(n^{-3/2}h^{\iota-2}\right)$, $\boldsymbol{L}_{\rho n,\upsilon j}^{(1)} = O_p\left(n^{-3/2}h^{\iota-2}\right)$ *and* $\boldsymbol{L}_{\rho n,\upsilon k}^{(1)} = O_p\left(n^{-3/2}h^{-2}\right)$;

(iii) *for* $\upsilon = 1, 2, 5, 6$,
$\boldsymbol{W}_{\rho n,\upsilon ij}^{(1)} = O_p\left(n^{-2}h^{\iota-2}\right)$, $\boldsymbol{W}_{\rho n,\upsilon ik}^{(1)} = O_p\left(n^{-2}h^{-3}\right)$ *and* $\boldsymbol{W}_{\rho n,\upsilon jk}^{(1)} = o_p\left(n^{-2}h^{-2}\right)$;

(iv) *for* $\upsilon = 3, 4, 7, 8$,
$\boldsymbol{W}_{\rho n,\upsilon ij}^{(1)} = O_p\left(n^{-2}h^{\iota-2}\right)$, $\boldsymbol{W}_{\rho n,\upsilon ik}^{(1)} = O_p\left(n^{-2}h^{-5/2}\right)$ *and* $\boldsymbol{W}_{\rho n,\upsilon jk}^{(1)} = O_p\left(n^{-2}h^{-5/2}\right)$;

(v) $\boldsymbol{T}_{\rho n,\upsilon ijk}^{(1)} = O_p\left(n^{-5/2}h^{-3}\right)$.

The proof of Lemma A.10 is given in Appendix S.C.2 using the same procedure as in the proof of Lemma A.9. We find that $\mathbb{U}_n^4 \rho_{n,\upsilon}\left(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta}\right)$ has the extra scale $(n-2)^{-1}$ and all the terms converge to zero faster than the leading terms in Lemma A.9. In addition, we note that the higher-order bias is $\mathbb{E}\rho_{n,\upsilon}^{(1)}\left(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta}^0\right)$ with order of $O\left(n^{-1}h^{\iota-2}\right)$. These terms will be dominated by the higher-order bias in $\mathbb{E}\left(\gamma_{n,\upsilon}^{(1)}\left(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta}^0\right)\right)$ as shown in Lemma A.9 below. This suggests that these bias terms will not contribute to the higher-order MSE later.

Lemmas A.9 and A.10 show the order of each element in the decomposition in $Q_{n,2}\left(\boldsymbol{\theta}\right) - Q_{n,2}\left(\boldsymbol{\theta}^0\right)$. Given these results, we give an approximation of $Q_{n,2}\left(\boldsymbol{\theta}\right) - Q_{n,2}\left(\boldsymbol{\theta}^0\right)$ in Proposition A.11.

Proposition A.11. *Suppose Lemma A.9 and Lemma A.10 hold. Then, uniformly*

over $O_p(\varsigma_n)$ *neighborhood of* $\boldsymbol{\theta}^0$

$$\begin{aligned} Q_{n,2}(\boldsymbol{\theta}) - Q_{n,2}(\boldsymbol{\theta}^0) &= (\boldsymbol{\theta}-\boldsymbol{\theta}^0)' \left(n^{-1}\sum\nolimits_{k=1}^{n}\psi_k\right) + o_p\left(\left\|\boldsymbol{\theta}-\boldsymbol{\theta}^0\right\|^2\right) \\ &\quad + O_p\left(\left\|\boldsymbol{\theta}-\boldsymbol{\theta}^0\right\| h^{\iota}\right) + o_p(h^{\iota}) \end{aligned}$$

where $\psi_k = 2\sum_{\upsilon=1}^{8}\kappa_\upsilon\left(\nabla_\theta\xi_{n,\upsilon k} - \mathbb{E}\left(\nabla_\theta\xi_{n,\upsilon k}\right)\right)$, *where*

$$\nabla_\theta\xi_{n,\upsilon k} = \begin{cases} (d_k - \varphi(z_{1k}, z_{2k}))\int\int \nabla_\theta\zeta\left(\boldsymbol{z}_k, \mathbf{s}, \boldsymbol{\theta}^0\right) f(s_1,s_2)\, d(s_1,s_2), & \upsilon = 1 \\ (d_k - \varphi(z_{1k}, z_{2k}))\int\int \nabla_\theta\zeta\left(z_{1k}, r_2, s_1, z_{2k}, \boldsymbol{\theta}^0\right) \frac{f(z_{1k},r_2)f(s_1,z_{2k})}{f(z_{1k},z_{2k})} d(r_2,s_1) & \upsilon = 3 \\ (d_k - \varphi(z_{1k}, z_{2k}))\int\int \nabla_\theta\zeta\left(2\boldsymbol{\theta}^0 - \boldsymbol{z}_k, \boldsymbol{s}, \boldsymbol{\theta}^0\right) f(s_1,s_2)\, d(s_1,s_2) & \upsilon = 5 \\ (d_k - \varphi(z_{1k}, z_{2k}))\int\int \nabla_\theta\zeta\left((2\theta_1^0 - z_{1k}), r_2, s_1, (2\theta_2^0 - z_{2k}), \boldsymbol{\theta}^0\right) & \\ \times \frac{f\left(\left(2\theta_1^0 - z_{1k}\right), r_2\right) f\left(s_1, \left(2\theta_2^0 - z_{2k}\right)\right)}{f(z_{1k},z_{2k})} d(r_2,s_1) & \upsilon = 7 \end{cases}$$

In addition $n^{-1/2}\sum_{k=1}^{n}\psi_k \to^d N(0,\Sigma)$, *where* $\Sigma = \mathbb{E}\left[\psi_k\psi_k'\right]$.

Proof of Proposition A.11: Provided results in Lemma A.9 and Lemma A.10, we observe that the leading terms are $\boldsymbol{L}_{n,\upsilon k}^{(1)}$ with order of $n^{-1/2}$, for $\upsilon = 1, \cdots, 8$. More specifically, we note that

$$\boldsymbol{L}_{n,\upsilon k}^{(1)} = n^{-1}\sum\nolimits_{k=1}^{n}\left(\nabla_\theta\xi_{n,\upsilon k} - \mathbb{E}\nabla_\theta\xi_{n,\upsilon k}\right) + n^{-1}\sum\nolimits_{k=1}^{n}\left(\nabla_\theta\tau_{n,\upsilon k} - \mathbb{E}\nabla_\theta\tau_{n,\upsilon k}\right),$$

and the asymptotic normality of $\boldsymbol{L}_{n,\upsilon k}^{(1)}$ will depend on the $\nabla_\theta\xi_{n,\upsilon k} - \mathbb{E}\nabla_\theta\xi_{n,\upsilon k}$, when the remainder converges to zero in probability. It has been shown that the remainder term $n^{-1}\sum_{k=1}^{n}\left(\nabla_\theta\tau_{n,\upsilon k} - \mathbb{E}\nabla_\theta\tau_{n,\upsilon k}\right) = o_p(h^{\iota})$. Hence, we can write $\boldsymbol{L}_{n,k}^{(1)} = \sum_{\upsilon=1}^{8}\kappa_\upsilon\boldsymbol{L}_{n,\upsilon k}^{(1)} = n^{-1}\sum_{\upsilon=1}^{8}\sum_{k=1}^{n}\kappa_\upsilon\left(\nabla_\theta\xi_{n,\upsilon k} - \mathbb{E}\nabla_\theta\xi_{n,\upsilon k}\right) + o_p(h^{\iota})$ and denote

$$\psi_k = 2\sum\nolimits_{\upsilon=1}^{8}\kappa_\upsilon\left(\nabla_\theta\xi_{n,\upsilon k} - \mathbb{E}\nabla_\theta\xi_{n,\upsilon k}\right).$$

In addition, we show that $\mathbb{V}\left(\nabla_\theta\xi_{n,\upsilon k}\right) = \mathcal{V}_{\upsilon\upsilon,k}$ with the bounded $\mathcal{V}_{\upsilon\upsilon,k}$ and we also calculate the covariance $cov\left(\nabla_\theta\xi_{n,\upsilon k}, \nabla_\theta\xi_{n,\upsilon' k}\right) = \mathcal{V}_{\upsilon\upsilon',k}$ in Appendix S.D.2. Now,

from Appendix S.C.2, it implies that $n^{-1/2}\sum_{k=1}^{n}\psi_k \to^d N(0,\Sigma)$, where $\Sigma = \mathbb{E}[\psi_k\psi_k']$, with $\Sigma_{\upsilon\upsilon} = \mathcal{V}_{\upsilon\upsilon k}$ and $\Sigma_{\upsilon\upsilon'} = \mathcal{V}_{\upsilon\upsilon',k}$ for $\upsilon = 1,\cdots,8$, and $\upsilon \neq \upsilon'$.

Therefore, here we will complete the proof by collecting all other remainder terms in $\mathbb{U}_n^3 q_{n,\upsilon}(\omega_i,\omega_j,\omega_k,\boldsymbol{\theta}) - \mathbb{U}_n^3 q_{n,\upsilon}\left(\omega_i,\omega_j,\omega_k,\boldsymbol{\theta}^0\right)$ and $\mathbb{U}_n^4 \rho_{n,\upsilon}(\omega_i,\omega_j,\omega_k,\boldsymbol{\theta}) - \mathbb{U}_n^4 \rho_{n,\upsilon}\left(\omega_i,\omega_j,\omega_k,\boldsymbol{\theta}^0\right)$. *Q.E.D.*

So far, we have already discussed the property of $Q_{n,1}(\boldsymbol{\theta}) - Q_{n,1}\left(\boldsymbol{\theta}^0\right)$ and $Q_{n,2}(\boldsymbol{\theta}) - Q_{n,2}\left(\boldsymbol{\theta}^0\right)$, respectively. Finally, we proceed to show that $Q_{n,3}(\boldsymbol{\theta}) - Q_{n,3}\left(\boldsymbol{\theta}^0\right)$. In particular, we note that we can decompose $Q_{n,3}(\boldsymbol{\theta})$ as follows,

$$\begin{aligned} Q_{n,3}(\boldsymbol{\theta}) &= (n(n-1))^{-1}\sum\nolimits_{i\neq j}\tau_{ij}\left[\sum\nolimits_{\upsilon=1}^{8}\kappa_\upsilon\left(\hat{\varphi}_{\upsilon,ij}(\boldsymbol{\theta}) - \varphi_{\upsilon,ij}(\boldsymbol{\theta})\right)\right]^2 \\ &= (n(n-1))^{-1}\sum\nolimits_{i\neq j}\tau_{ij}\sum\nolimits_{\upsilon=1}^{8}\left(\hat{\varphi}_{\upsilon,ij}(\boldsymbol{\theta}) - \varphi_{\upsilon,ij}(\boldsymbol{\theta})\right)^2 \\ &\quad +2(n(n-1))^{-1}\sum\nolimits_{i\neq j}\tau_{ij}\sum\nolimits_{\upsilon\neq\upsilon'}\kappa_\upsilon\kappa_{\upsilon'}\left(\hat{\varphi}_{\upsilon,ij}(\boldsymbol{\theta}) - \varphi_{\upsilon,ij}(\boldsymbol{\theta})\right)\left(\hat{\varphi}_{\upsilon',ij}(\boldsymbol{\theta}) - \varphi_{\upsilon',ij}(\boldsymbol{\theta})\right). \end{aligned}$$

Note we decompose $[\hat{\varphi}_\upsilon - \varphi_\upsilon]^2$ (omitting ij indices) in $Q_{n,3}(\boldsymbol{\theta})$ as the same way we did for the decomposition in $Q_{n,2}(\boldsymbol{\theta})$.

$$\begin{aligned} & [\hat{\varphi}_\upsilon - \varphi_\upsilon]^2 \\ = & \left[\hat{g}_\upsilon - \varphi_\upsilon\hat{f}_\upsilon\right]^2 f_\upsilon^{-2}\left(1 - f^{-1}\left(\hat{f}_\upsilon - f_\upsilon\right) + o_p\left(f^{-1}\left(\hat{f}_\upsilon - f_\upsilon\right)\right)\right)^2 \qquad \text{(A.8)} \\ = & f_\upsilon^{-2}\left[\hat{g}_\upsilon - \varphi_\upsilon\hat{f}_\upsilon\right]^2 + o_p(1) = f_\upsilon^{-2}\left[\frac{1}{n-2}\sum\nolimits_{k=1}^{n-2}(d_k - \varphi_\upsilon)K_{n,\upsilon}\right]^2 \\ = & f_\upsilon^{-2}(n-2)^{-2}\sum\nolimits_{k=1}^{n-2}\sum\nolimits_{l=1}^{n-2}(d_k - \varphi_\upsilon)(d_l - \varphi_\upsilon)K_{n,\upsilon k}K_{n,\upsilon l} + o_p(1) \\ = & (n-2)^{-2}\sum\nolimits_{k=1}^{n-2}f_\upsilon^{-2}(d_k - \varphi_\upsilon)^2 K_{n,\upsilon k}^2 \\ & + (n-2)^{-2}\sum\nolimits_{k\neq l}f_\upsilon^{-2}(d_k - \varphi_\upsilon)(d_l - \varphi_\upsilon)K_{n,\upsilon k}K_{n,\upsilon l} + o_p(1). \end{aligned}$$

It is easy to verify that in term A.8, the cross-product term in the last expression will be dominated by the quadratic term, by the fact that the extra averaging removes the scale factor $1/h^2$ due to the i.i.d. sample. It implies that the cross-product term

converges to zero faster than the quadratic term. Analogously, we consider a similar decomposition for the cross-product terms with different values υ and υ',

$$\begin{aligned}
& \left(\hat{\varphi}_{\upsilon}-\varphi_{\upsilon}\right)\left(\hat{\varphi}_{\upsilon'}-\varphi_{\upsilon'}\right) \\
= & \left[\hat{g}_{\upsilon}-\varphi_{\upsilon}\hat{f}_{\upsilon}\right] f_{\upsilon}^{-1}\left(1-f^{-1}\left(\hat{f}_{\upsilon}-f_{\upsilon}\right)+o_p\left(f^{-1}\left(\hat{f}_{\upsilon}-f_{\upsilon}\right)\right)\right) \\
& \times\left[\hat{g}_{\upsilon'}-\varphi_{\upsilon'}\hat{f}_{\upsilon'}\right] f_{\upsilon'}^{-1}\left(1-f^{-1}\left(\hat{f}_{\upsilon'}-f_{\upsilon'}\right)+o_p\left(f^{-1}\left(\hat{f}_{\upsilon'}-f_{\upsilon'}\right)\right)\right) \\
= & \left[\hat{g}_{\upsilon}-\varphi_{\upsilon}\hat{f}_{\upsilon}\right]\left[\hat{g}_{\upsilon'}-\varphi_{\upsilon'}\hat{f}_{\upsilon'}\right]\left(f_{\upsilon}f_{\upsilon'}\right)^{-1}+o_p\left(1\right) \\
= & \left(n-2\right)^{-2}\sum\nolimits_{k=1}^{n-2}\sum\nolimits_{l=1}^{n-2}\left(f_{\upsilon}f_{\upsilon'}\right)^{-1}\left(d_k-\varphi_{\upsilon}\right)\left(d_l-\varphi_{\upsilon'}\right)K_{n,\upsilon k}K_{n,\upsilon l}+o_p\left(1\right) \\
= & \left(n-2\right)^{-2}\sum\nolimits_{k=1}^{n-2}\left(f_{\upsilon}f_{\upsilon'}\right)^{-1}\left(d_k-\varphi_{\upsilon}\right)\left(d_k-\varphi_{\upsilon'}\right)K_{n,\upsilon_1 k}K_{n,\upsilon_2 k} \\
& +\left(n-2\right)^{-2}\sum\nolimits_{k\neq l}\left(f_{\upsilon}f_{\upsilon'}\right)^{-1}\left(d_k-\varphi_{\upsilon}\right)\left(d_l-\varphi_{\upsilon'}\right)K_{n,\upsilon k}K_{n,\upsilon' l}+o_p\left(1\right).
\end{aligned} \tag{A.9}$$

Following a similar argument as above, in (A.9) we observe that the cross-product term in the last expression is dominated by the quadratic term due to the extra averaging. In the following, we will focus on the quadratic terms in (A.8) and (A.9).

Next, we introduce $\gamma_{n,\upsilon}\left(\omega_i,\omega_j,\omega_k,\boldsymbol{\theta}\right)$ to denote the quadratic term in (A.8),

$$\gamma_{n,\upsilon}\left(\omega_i,\omega_j,\omega_k,\boldsymbol{\theta}\right)=\left(d_k-\varphi_{\upsilon}\right)^2K_{n,\upsilon k}^2/f_{\upsilon}^2,$$

where $\omega_i=(Z_{1i},Z_{2i})$, $\omega_j=(Z_{1j},Z_{2j})$ and $\omega_k=(d_k,Z_{1k},Z_{2k})$. Similarly, we write the linear combination of $\gamma_{n,\upsilon}$ as $\gamma_n\left(\omega_i,\omega_j,\omega_k,\theta\right)=\sum_{\upsilon=1}^{8}\left(d_k-\varphi_{\upsilon}\right)^2K_{n,\upsilon k}^2/f_{\upsilon}^2$. Then similarly, we introduce $\gamma_{n,\upsilon\upsilon'}\left(\omega_i,\omega_j,\omega_k,\boldsymbol{\theta}\right)$ to denote the quadratic term in (A.9),

$$\gamma_{n,\upsilon\upsilon'}\left(\omega_i,\omega_j,\omega_k,\boldsymbol{\theta}\right)=\left(d_k-\varphi_{\upsilon}\right)\left(d_k-\varphi_{\upsilon'}\right)K_{n,\upsilon k}K_{n,\upsilon' k}/f_{\upsilon}f_{\upsilon'}.$$

Finally, let $\mathbb{U}_n^4$ be the random probability measure that put mass $\left(n\left(n-1\right)\left(n-2\right)^2\right)^{-1}$ on each order 3-tuple observation (the extra n component due to the product of the

choice probability). Following these notations, we can write $Q_{n,3}(\boldsymbol{\theta})$ as

$$\begin{aligned} Q_{n,3}(\boldsymbol{\theta}) &= \mathbb{U}_n^4 \gamma_n(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta}) + \sum\nolimits_{\upsilon \neq \upsilon'} \mathbb{U}_n^4 \gamma_{n,\upsilon\upsilon'}(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta}) + o_p(1) \\ &= \sum\nolimits_{\upsilon=1}^{8} \kappa_\upsilon \left[\mathbb{U}_n^4 \gamma_{n,\upsilon}(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta}) + \sum\nolimits_{\upsilon \neq \upsilon'} \mathbb{U}_n^4 \gamma_{n,\upsilon\upsilon'}(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta}) \right] + o_p(1). \end{aligned}$$

The following two lemmas provide the order of each term in $\mathbb{U}_n^4 \gamma_{n,\upsilon}(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta}) - \mathbb{U}_n^4 \gamma_{n,\upsilon}\left(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta}^0\right)$ and $\mathbb{U}_n^4 \gamma_{n,\upsilon\upsilon'}(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta}) - \mathbb{U}_n^4 \gamma_{n,\upsilon\upsilon'}\left(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta}^0\right)$, respectively, which are the key elements for showing the asymptotic property of $Q_{n,3}(\boldsymbol{\theta}) - Q_{n,3}\left(\boldsymbol{\theta}^0\right)$.

Lemma A.12. *Given that Assumptions 2-3 hold, uniformly over $O_p(\varsigma_n)$ neighborhood of $\boldsymbol{\theta}^0$, when $\upsilon = 1, 2, 5, 6$,*

$$\begin{aligned} & \mathbb{U}_n^4 \gamma_{n,\upsilon}(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta}) - \mathbb{U}_n^4 \gamma_{n,\upsilon}\left(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta}^0\right) \\ = & \left(\boldsymbol{\theta} - \boldsymbol{\theta}^0\right)' \mathbb{E}\left(\gamma_{n,\upsilon}^{(1)}\left(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta}^0\right)\right) \\ & + \left(\boldsymbol{\theta} - \boldsymbol{\theta}^0\right)' \left(\boldsymbol{L}_{\gamma n,\upsilon i}^{(1)} + \boldsymbol{L}_{\gamma n,\upsilon k}^{(1)}\right) \\ & + \left(\boldsymbol{\theta} - \boldsymbol{\theta}^0\right)' \boldsymbol{W}_{\gamma n,\upsilon ik}^{(1)} + o_p\left(\left\|\boldsymbol{\theta} - \boldsymbol{\theta}^0\right\|^2\right), \end{aligned}$$

where the order of each term is as follows

(i) $\mathbb{E}\gamma_{n,\upsilon}^{(1)}(\omega_i, \omega_j, \omega_k, \theta^0) = \gamma_\upsilon^{(1)}(\theta^0) = O(n^{-1}h^{-2})$;

(ii) $\boldsymbol{L}_{\gamma n,\upsilon i}^{(1)} = O_p\left(n^{-3/2}h^{-2}\right)$ *and* $\boldsymbol{L}_{\gamma n,\upsilon k}^{(1)} = O_p\left(n^{-3/2}h^{-2}\right)$;

(iii) $\boldsymbol{W}_{\gamma n,\upsilon ik}^{(1)} = O_p(n^{-2}h^{-2})$.

When $\upsilon = 3, 4, 7, 8$,

$$\begin{aligned} & \mathbb{U}_n^4 \gamma_{n,\upsilon}(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta}) - \mathbb{U}_n^4 \gamma_{n,\upsilon}\left(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta}^0\right) \\ = & \left(\boldsymbol{\theta} - \boldsymbol{\theta}^0\right)' \mathbb{E}\left(\gamma_{n,\upsilon}^{(1)}\left(\omega_i, \omega_j, \omega_k, \boldsymbol{\theta}^0\right)\right) \\ & + \left(\boldsymbol{\theta} - \boldsymbol{\theta}^0\right)' \left(\boldsymbol{L}_{\gamma n,\upsilon i}^{(1)} + \boldsymbol{L}_{\gamma n,\upsilon j}^{(1)} + \boldsymbol{L}_{\gamma n,\upsilon k}^{(1)}\right) \\ & + \left(\boldsymbol{\theta} - \boldsymbol{\theta}^0\right)' \left(\boldsymbol{W}_{\gamma n,\upsilon ij}^{(1)} + \boldsymbol{W}_{\gamma n,\upsilon ik}^{(1)} + \boldsymbol{W}_{\gamma n,\upsilon jk}^{(1)}\right) \\ & + \left(\boldsymbol{\theta} - \boldsymbol{\theta}^0\right)' \boldsymbol{T}_{\gamma n,\upsilon ijk}^{(1)} + o_p\left(\left\|\boldsymbol{\theta} - \boldsymbol{\theta}^0\right\|^2\right), \end{aligned}$$

where the order of each term is shown as follows

$$
\begin{aligned}
&(i) \quad \mathbb{E}\gamma_{n,\upsilon}^{(1)}\left(\omega_i,\omega_j,\omega_k,\boldsymbol{\theta}^0\right)=\gamma_{\upsilon}^{(1)}\left(\boldsymbol{\theta}^0\right)=O\left(n^{-1}h^{-2}\right);\\
&(ii) \quad \boldsymbol{L}_{\gamma n,\upsilon i}^{(1)}=O_p\left(n^{-3/2}h^{-2}\right),\boldsymbol{L}_{\gamma n,\upsilon j}^{(1)}=O_p\left(n^{-3/2}h^{-2}\right) \text{ and } \boldsymbol{L}_{\gamma n,\upsilon j}^{(1)}=O_p\left(n^{-3/2}h^{-2}\right);\\
&(iii) \quad \boldsymbol{W}_{\gamma n,\upsilon ij}^{(1)}=O_p\left(n^{-2}h^{-2}\right),\boldsymbol{W}_{\gamma n,\upsilon ik}^{(1)}=O_p\left(n^{-2}h^{-5/2}\right) \text{ and } \boldsymbol{W}_{\gamma n,\upsilon jk}^{(1)}=O_p\left(n^{-2}h^{-5/2}\right);\\
&(iv) \quad \boldsymbol{T}_{\gamma n,\upsilon ijk}^{(1)}=O_p\left(n^{-5/2}h^{-3}\right).
\end{aligned}
$$

The proof of Lemma A.12 is given in Appendix S.C.2. Similar to $\mathbb{U}_n^4\rho_{n,\upsilon}$ in Lemma A.10, $\mathbb{U}_n^4\gamma_{n,\upsilon}$ has the extra scale $(n-2)^{-1}$ and is also attributed to the higher-order expansion, meaning that all the decomposition terms will decline to zero faster than the leading terms in Lemma A.9. In addition, we show that the higher-order bias $\mathbb{E}\gamma_{n,\upsilon}^{(1)}$ is order of $O\left(n^{-1}h^{-2}\right)$, for $\upsilon=1,\cdots,8$. This order is the same as that of the variance of the estimated choice probability. It is not coincidental by the fact that $\mathbb{E}\gamma_{n,\upsilon}^{(1)}$ takes a similar form as the variance of the estimated choice probability and they thus have the same order. Finally, as mentioned before, since $\mathbb{E}\gamma_{n,\upsilon}^{(1)}$ dominates $\mathbb{E}\rho_{n,\upsilon}^{(1)}$, the former will contribute to the higher-order MSE expansion.

Next, we show the orders of cross-product terms. Due to the similarity of the terms, we will only discuss the case when $\upsilon=1$, $\upsilon'=3$ and $\upsilon=1$, $\upsilon'=5$.

Lemma A.13. *Given that Assumptions 2-3 hold, uniformly over $O_p\left(\varsigma_n\right)$ neighborhood of $\boldsymbol{\theta}^0$, when $\upsilon=1$ and $\upsilon'=3$,*

$$
\begin{aligned}
&\mathbb{U}_n^4\gamma_{n,13}\left(\omega_i,\omega_j,\omega_k,\boldsymbol{\theta}\right)-\mathbb{U}_n^4\gamma_{n,13}\left(\omega_i,\omega_j,\omega_k,\boldsymbol{\theta}^0\right)\\
=\;&\left(\boldsymbol{\theta}-\boldsymbol{\theta}^0\right)'\mathbb{E}\left(\gamma_{n,13}^{(1)}\left(\omega_i,\omega_j,\omega_k,\boldsymbol{\theta}^0\right)\right)\\
&+\left(\boldsymbol{\theta}-\boldsymbol{\theta}^0\right)'\left(\boldsymbol{L}_{\gamma n,13i}^{(1)}+\boldsymbol{L}_{\gamma n,13j}^{(1)}+\boldsymbol{L}_{\gamma n,13k}^{(1)}\right)\\
&+\left(\boldsymbol{\theta}-\boldsymbol{\theta}^0\right)'\left(\boldsymbol{W}_{\gamma n,13ij}^{(1)}+\boldsymbol{W}_{\gamma n,13ik}^{(1)}+\boldsymbol{W}_{\gamma n,13jk}^{(1)}\right)\\
&+\left(\boldsymbol{\theta}-\boldsymbol{\theta}^0\right)'\boldsymbol{T}_{\gamma n,13ijk}^{(1)}+o_p\left(\left\|\boldsymbol{\theta}-\boldsymbol{\theta}^0\right\|^2\right),
\end{aligned}
$$

where the order of each term can be shown as follows

$$
\begin{aligned}
&(i) \quad \mathbb{E}\gamma_{n,13}^{(1)}\left(\omega_i,\omega_j,\omega_k,\boldsymbol{\theta}^0\right)=\gamma_{n,13}^{(1)}\left(\boldsymbol{\theta}^0\right)=O\left(n^{-1}h^{-1}\right);\\
&(ii) \quad \boldsymbol{L}_{\gamma n,13i}^{(1)}=O_p\left(n^{-3/2}h^{-1}\right),\boldsymbol{L}_{\gamma n,13j}^{(1)}=O_p\left(n^{-3/2}h^{-1}\right) \text{ and } \boldsymbol{L}_{\gamma n,13k}^{(1)}=O_p\left(n^{-3/2}h^{-1}\right);\\
&(iii) \quad \boldsymbol{W}_{\gamma n,13ij}^{(1)}=O_p\left(n^{-2}h^{-1}\right),\boldsymbol{W}_{\gamma n,13ik}^{(1)}=O_p\left(n^{-2}h^{-2}\right) \text{ and } \boldsymbol{W}_{\gamma n,13jk}^{(1)}=O_p\left(n^{-2}h^{-3/2}\right);\\
&(iv) \quad \boldsymbol{T}_{\gamma n,vijk}^{(1)}=O_p\left(n^{-5/2}h^{-3}\right).
\end{aligned}
$$

In addition, when $\upsilon=1$ *and* $\upsilon'=5$,

$$
\begin{aligned}
&\mathbb{U}_n^4\gamma_{n,15}\left(\omega_i,\omega_j,\omega_k,\boldsymbol{\theta}\right)-\mathbb{U}_n^4\gamma_{n,15}\left(\omega_i,\omega_j,\omega_k,\boldsymbol{\theta}^0\right)\\
=\;&\left(\boldsymbol{\theta}-\boldsymbol{\theta}^0\right)'\mathbb{E}\left(\gamma_{n,15}^{(1)}\left(\omega_i,\omega_j,\omega_k,\boldsymbol{\theta}^0\right)\right)\\
&+\left(\boldsymbol{\theta}-\boldsymbol{\theta}^0\right)'\left(\boldsymbol{L}_{\gamma n,15i}^{(1)}+\boldsymbol{L}_{\gamma n,15k}^{(1)}\right)\\
&+\left(\boldsymbol{\theta}-\boldsymbol{\theta}^0\right)'\boldsymbol{W}_{\gamma n,15ik}^{(1)}+o_p\left(\left\|\boldsymbol{\theta}-\boldsymbol{\theta}^0\right\|^2\right),
\end{aligned}
$$

where the order of each term can be shown as follows

$$
\begin{aligned}
&(i) \quad \mathbb{E}\gamma_{n,15}^{(1)}\left(\omega_i,\omega_j,\omega_k,\boldsymbol{\theta}^0\right)=\gamma_{n,15}^{(1)}\left(\boldsymbol{\theta}^0\right)=O\left(n^{-1}\right);\\
&(ii) \quad \boldsymbol{L}_{\gamma n,15i}^{(1)}=O_p\left(n^{-3/2}h^{-2}\right) \text{ and } \boldsymbol{L}_{\gamma n,15k}^{(1)}=O_p\left(n^{-3/2}h^{-2}\right);\\
&(iii) \quad \boldsymbol{W}_{\gamma n,15ik}^{(1)}=O_p\left(n^{-2}h^{-3}\right).
\end{aligned}
$$

The proof of Lemma A.13 also appears in Appendix S.C.4. Lemma A.13 suggests that all the terms will converge to zero faster than the leading terms and also faster than the first-order bias and higher-order bias. This suggests that the terms will not contribute to the asymptotic linear representation, the asymptotic normality, nor the higher-order MSE.

Lemmas A.12 and A.13 gives the order of each element in the decomposition of $Q_{n,3}\left(\boldsymbol{\theta}\right)-Q_{n,3}\left(\boldsymbol{\theta}^0\right)$. Given these results, we can directly derive the quadratice approximation of $Q_{n,3}\left(\boldsymbol{\theta}\right)-Q_{n,3}\left(\boldsymbol{\theta}^0\right)$.

Proposition A.14. *Suppose that Lemma C.3 and Lemma C.4 hold. Then uniformly*

over $O_p(\varsigma_n)$ *neighborhood of* $\boldsymbol{\theta}^0$

$$Q_{n,3}(\boldsymbol{\theta}) - Q_{n,3}\left(\boldsymbol{\theta}^0\right) = o_p\left(\left\|\boldsymbol{\theta} - \boldsymbol{\theta}^0\right\|^2\right) + O_p\left(\left\|\boldsymbol{\theta} - \boldsymbol{\theta}^0\right\| / nh^2\right) + o_p\left(n^{-1}h^{-2}\right)$$

Proof of Proposition A.14: The desired result can be obtained by directly collecting the terms in Lemmas A.12 and A.13. *Q.E.D*

Until now, we have shown the quadratic approximation of $Q_{n,1}(\boldsymbol{\theta}) - Q_{n,1}\left(\boldsymbol{\theta}^0\right)$, $Q_{n,2}(\boldsymbol{\theta}) - Q_{n,2}\left(\boldsymbol{\theta}^0\right)$ and $Q_{n,3}(\boldsymbol{\theta}) - Q_{n,3}\left(\boldsymbol{\theta}^0\right)$, respectively. We will use these results above to prove the $\sqrt{n}$-consistency and asymptotic normality of the estimator.

A.3.3.2 Proofs of Theorems

Having examined the asymptotic property of the sample objective function, we follow to show $\sqrt{n}$-consistency and asymptotic normality of the estimator. Recall that $Q(\boldsymbol{\theta})$ is the population objective function that restricts the observables in the interior of the support.

Proof of Theorem I.14: To show $\sqrt{n}$-consistency of $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}$, we will use Theorem 1 in Sherman (1994), which requires that the following conditions hold: (A1) $\left\|\boldsymbol{\theta} - \boldsymbol{\theta}^0\right\| = o_p(1)$ (A2) there exists a neighborhood $\mathcal{N}$ of $\boldsymbol{\theta}^0$ and a constant $c > 0$ for which $-Q(\boldsymbol{\theta}) - \left(-Q\left(\boldsymbol{\theta}^0\right)\right) \leq -c\left\|\boldsymbol{\theta} - \boldsymbol{\theta}^0\right\|^2$ for all $\boldsymbol{\theta}$ in $\mathcal{N}$ (Given the minimization problem in our context, we add the negative sign to transform it into a maximization problem to fit the theorem); (A3) uniformly over $o_p(1)$ neighborhood of $\boldsymbol{\theta}^0$,

$$Q_n(\boldsymbol{\theta}) - Q_n\left(\boldsymbol{\theta}^0\right) = Q(\boldsymbol{\theta}) - Q^{tr}\left(\boldsymbol{\theta}^0\right) + O_p\left(\left\|\boldsymbol{\theta} - \boldsymbol{\theta}^0\right\| / \sqrt{n}\right) + o_p\left(\left\|\boldsymbol{\theta} - \boldsymbol{\theta}^0\right\|^2\right) + O_p(\epsilon_n).$$

Then the rate of convergence of the estimator $\boldsymbol{\theta}$ is $\left\|\boldsymbol{\theta} - \boldsymbol{\theta}^0\right\| = O_p(\max(\epsilon_n^{1/2}, 1/\sqrt{n}))$. Now, in the following, we show our estimator $\hat{\boldsymbol{\theta}}_n$ is $\sqrt{n}$-consistent.

Condition (A1) holds by virtue of the consistency of the estimator $\hat{\boldsymbol{\theta}}_n$ shown in

Theorem 3.1. Condition (A2) is satisfied by the fact that

$$\begin{aligned} & -Q\left(\hat{\boldsymbol{\theta}}_n\right)-\left(-Q\left(\boldsymbol{\theta}^0\right)\right) \\ = & -\left(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^0\right)^{'}\nabla_\theta Q\left(\boldsymbol{\theta}^0\right)-\frac{1}{2}\left(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^0\right)^{'}\nabla_{\theta\theta}Q\left(\boldsymbol{\theta}^0\right)\left(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^0\right)+o_p\left(\left\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^0\right\|^2\right), \end{aligned}$$

where

$$\begin{aligned} \nabla_\theta Q\left(\boldsymbol{\theta}^0\right) &= \mathbb{E}\left[\tau_{ij}\left(\boldsymbol{\theta}^0\right)\left(\sum\nolimits_{\upsilon=1}^{8}\kappa_\upsilon\varphi_{\upsilon,ij}\left(\boldsymbol{\theta}^0\right)\right)\nabla_\theta\left[\tau_{ij}\left(\boldsymbol{\theta}^0\right)\sum\nolimits_{\upsilon=1}^{8}\kappa_\upsilon\varphi_{\upsilon,ij}\left(\boldsymbol{\theta}^0\right)\right]\right]; \\ \nabla_{\theta\theta}Q\left(\boldsymbol{\theta}^0\right) &= \mathbb{E}\left[\nabla_\theta\left[\tau_{ij}\left(\boldsymbol{\theta}^0\right)\sum\nolimits_{\upsilon=1}^{8}\kappa_\upsilon\varphi_{\upsilon,ij}\left(\boldsymbol{\theta}^0\right)\right]\nabla_\theta\left[\tau_{ij}\left(\boldsymbol{\theta}^0\right)\sum\nolimits_{\upsilon=1}^{8}\kappa_\upsilon\varphi_{\upsilon,ij}\left(\boldsymbol{\theta}^0\right)\right]\right]^{'}. \end{aligned}$$

From the property of the function that contains $\sum_{\upsilon=1}^{8}\tau_{ij}\left(\boldsymbol{\theta}^0\right)\varphi_{\upsilon,ij}\left(\boldsymbol{\theta}^0\right)$, we can directly show that $\nabla_\theta Q\left(\boldsymbol{\theta}^0\right)=0$. In addition, due to the quadratic form and nonzero difference on the density, we can show that $\nabla_{\theta\theta}Q\left(\boldsymbol{\theta}^0\right)$ is positive definite; and moreover from Assumption 3 with the bounded derivative of the function φ, $\left\|\nabla_{\theta\theta}Q\left(\boldsymbol{\theta}^0\right)\right\|\leq c$ for some constant $c>0$. Then the desired result follows from combining these results.

For Condition (A3), choose ς_n such that uniformly over the neighborhood $o_p\left(1\right)$ of $\boldsymbol{\theta}^0$,

$$Q_n\left(\hat{\boldsymbol{\theta}}_n\right)-Q_n\left(\boldsymbol{\theta}^0\right)=Q_{n,1}\left(\hat{\boldsymbol{\theta}}_n\right)-Q_{n,1}\left(\boldsymbol{\theta}^0\right)+Q_{n,2}\left(\hat{\boldsymbol{\theta}}_n\right)-Q_{n,2}\left(\boldsymbol{\theta}^0\right)+Q_{n,3}\left(\hat{\boldsymbol{\theta}}_n\right)-Q_{n,3}\left(\boldsymbol{\theta}^0\right),$$

which follows to verify

$$\begin{aligned} &\text{(i)} \quad Q_{n,1}\left(\hat{\boldsymbol{\theta}}_n\right)=Q_1\left(\hat{\boldsymbol{\theta}}_n\right)+o_p\left(1\right), and Q_{n,1}\left(\boldsymbol{\theta}^0\right)=Q_1\left(\boldsymbol{\theta}^0\right)+o_p\left(1\right); \\ &\text{(ii)} \quad Q_{n,2}\left(\hat{\boldsymbol{\theta}}_n\right)-Q_{n,2}\left(\boldsymbol{\theta}^0\right)=O_p\left(\left\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^0\right\|/\sqrt{n}\right)+o_p\left(\left\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^0\right\|^2\right) \\ &\qquad +O_p\left(\left\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^0\right\|h^\iota\right)+o_p\left(h^\iota\right); \\ &\text{(iii)} \quad Q_{n,3}\left(\hat{\boldsymbol{\theta}}_n\right)-Q_{n,3}\left(\boldsymbol{\theta}^0\right)=O_p\left(\left\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^0\right\|/nh^2\right)+o_p\left(n^{-1}h^{-2}\right). \end{aligned}$$

It is straightforward to check that (i), (ii) and (iii) hold from Proposition A.8, A.11, A.14 respectively.

Now, let $O_p(\varsigma_n) = O_p(h^\iota) + O_p(n^{-1}h^{-2})$. Then $O_p\left(\left\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0\right\| h^\iota\right) + O_p\left(\left\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0\right\| / nh^2\right) = O_p(\epsilon_n)$, with $\epsilon_n = h^{2\iota} + n^{-2}h^{-4}$. Let the bandwidth satisfy $nh^{2\iota} \to 0$ and $nh^4 \to \infty$, which gives the desired result.

Proof of Theorem I.16: To show asymptotic linearity and asymptotic normality, we will follow Theorem 2 in Sherman (1994), which requires that (B1) $\hat{\boldsymbol{\theta}}_n$ is $\sqrt{n}$-consistent for $\boldsymbol{\theta}^0$, an interior point of Θ; (B2) uniformly over $O_p(1/\sqrt{n})$ neighborhoods of $\boldsymbol{\theta}^0$

$$Q_n\left(\hat{\boldsymbol{\theta}}_n\right) - Q_n\left(\boldsymbol{\theta}^0\right) = \frac{1}{2}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0\right)' \Gamma \left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0\right) + \frac{1}{\sqrt{n}}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0\right)' M_n + o_p\left(\frac{1}{n}\right),$$

where Γ is a positive definite matrix, and M_n is normally distributed.

The first part of Condition (B1) follows from Theorem I.15 and the second part is satisfied by Assumption 1. For Condition (B2), let $\varsigma_n = 1/\sqrt{n}$. We proceed to check that uniformly over $O_p(1/\sqrt{n})$ neighborhood of $\boldsymbol{\theta}^0$ (i) $Q_{n,1}\left(\hat{\boldsymbol{\theta}}_n\right) - Q_{n,2}\left(\boldsymbol{\theta}^0\right) = \frac{1}{2}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0\right)' \Gamma \left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0\right) + o_p(1)$ by the same argument as in Proposition C.1; (ii) $Q_{n,2}\left(\hat{\boldsymbol{\theta}}_n\right) - Q_{n,2}\left(\boldsymbol{\theta}^0\right) = \frac{1}{\sqrt{n}}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0\right)' M_n + o_p\left(\left\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0\right\|^2\right) + o_p\left(\frac{1}{n}\right)$, with $M_n = n^{-1/2}\sum_{k=1}^n \psi_k$ where $\psi_k = \sum_{\upsilon=1}^8 \kappa_\upsilon\left(\nabla_\theta \xi_{n,\upsilon k} - \mathbb{E}\nabla_\theta \xi_{n,\upsilon k}\right)$ and $M_n \to^d N\left(0, \mathbb{E}\left[\psi_k \psi_k'\right]\right)$ by the same argument in Proposition C.2 and the bandwidth satisfies $nh^{2\iota} \to 0$ and $nh^4 \to \infty$; (iii) Finally, $Q_{n,3}\left(\hat{\boldsymbol{\theta}}_n\right) - Q_{n,3}\left(\boldsymbol{\theta}^0\right) = o_p\left(\frac{1}{n}\right)$ by the same argument as in Proposition A.14 and the bandwidth satisfies $nh^{2\iota} \to 0$ and $nh^4 \to \infty$. After verifying these conditions, the asymptotic linearity follows from the fact that

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0\right) = n^{-1/2}\sum\nolimits_{k=1}^n -\Gamma^{-1}\psi_k + O_p(1/\sqrt{n}),$$

and the asymptotic normality follows from the fact that $M_n = n^{-1/2}\sum_{k=1}^n \psi_k \to^d$

$N\left(0, \mathbb{E}\left[\psi_k \psi_k'\right]\right)$ and Delta method, that is,

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0\right) \to^d N\left(0, \Gamma^{-1}\Sigma\Gamma^{-1\prime}\right),$$

and we denote $\Sigma = \mathbb{E}\left[\psi_k \psi_k'\right]$, which completes the proof. *Q.E.D.*

A.3.4 Higher-Order Mean Squared Error Approximation

Proof of Theorem I.17: We observe that

$$\bar{\Gamma}_n\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0\right)\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0\right)'\bar{\Gamma}_n' = G_n + r_n,$$

where $G_n = \left(\boldsymbol{L}_n^{(1)} + q^{(1)}\left(\boldsymbol{\theta}^0\right) + \gamma^{(1)}\left(\boldsymbol{\theta}^0\right)\right)\left(\boldsymbol{L}_n^{(1)} + q^{(1)}\left(\boldsymbol{\theta}^0\right) + \gamma^{(1)}\left(\boldsymbol{\theta}^0\right)\right)'$ and r_n summarizes all the remainder terms other than those in G_n. In addition, we also note that

$$\begin{aligned} G_n &= \left(\boldsymbol{L}_n^{(1)} + q^{(1)}\left(\boldsymbol{\theta}^0\right) + \gamma^{(1)}\left(\boldsymbol{\theta}^0\right)\right)\left(\boldsymbol{L}_n^{(1)} + q^{(1)}\left(\boldsymbol{\theta}^0\right) + \gamma^{(1)}\left(\boldsymbol{\theta}^0\right)\right)' \\ &= A_{G_n} + R_{G_n} \end{aligned}$$

where $A_{G_n} = \frac{1}{n}\left(\sum_{k=1}^n \psi_k\right)\left(\sum_{k=1}^n \psi_k\right)' + h^2\mathcal{B}\mathcal{B}' + \frac{1}{n^2h^4}\mathcal{B}^h\mathcal{B}^{h\prime}$ and R_{G_n} collects all the remainder terms. Then, let $\rho_{h,n} = tr\left(S\left(h\right)\right)$

$$\begin{aligned} \mathbb{E}\left[A_{G_n}\right] &= \frac{1}{n}\Sigma + S\left(h\right) + o(\rho_{h,n}) \\ &= \frac{1}{n}\Sigma + h^2\mathcal{B}\mathcal{B}' + \frac{1}{n^2h^4}\mathcal{B}^h\mathcal{B}^{h\prime} + o(\rho_{h,n}) \end{aligned}$$

It is straightforward to verify that the cross-product term is dominated by the squared product using CSI. In addition, the other higher-order terms are dominated by bias terms using Lemmas A.9, A.10, A.12 and A.13. Now, note that the second term in the last equality is the first order bias, where $\mathcal{B}$ is from the expansion of

$\mathbb{E}\left(q_n\left(\omega_i,\omega_j,\omega_k,\boldsymbol{\theta}\right)\right)-\mathbb{E}\left(q_n\left(\omega_i,\omega_j,\omega_k,\boldsymbol{\theta}^0\right)\right)$ in $Q_{n,2}$; the third term is the higher-order bias, $\mathcal{B}^h$ is from the expansion of $\mathbb{E}\left(\gamma_n\left(\omega_i,\omega_j,\omega_k,\boldsymbol{\theta}\right)\right)-\mathbb{E}\left(\gamma_n\left(\omega_i,\omega_j,\omega_k,\boldsymbol{\theta}^0\right)\right)$ in $Q_{n,3}$, which goes to zero slower than the expansion of $\mathbb{E}\left(\rho_n\left(\omega_i,\omega_j,\omega_k,\boldsymbol{\theta}\right)\right)-\mathbb{E}\left(\rho_n\left(\omega_i,\omega_j,\omega_k,\boldsymbol{\theta}^0\right)\right)$ in $Q_{n,2}$. Following Appendixes S.D, S.E and S.F, we can write $\mathcal{B}=\sum_{\upsilon=1}^{8}\mathcal{B}_\upsilon$ and $\mathcal{B}^h=\sum_{\upsilon=1}^{8}\mathcal{B}_\upsilon^h$, where

$$\mathcal{B}_\upsilon=\begin{cases}\mathbb{E}_{[i,j]}\left[\nabla_\theta\zeta_{ij}\left(\boldsymbol{\theta}^0\right)\sum\limits_{\iota_1+\iota_2=\iota,0<\iota_1,\iota_2\leq\iota}\left[\int\int u_1^{\iota_1}u_2^{\iota_2}K\left(\mathbf{u}\right)d\mathbf{u}\right]\vartheta_{\iota_1,\iota_2}\left(Z_{1i},Z_{2i}\right)\right], & \upsilon=1;\\ \mathbb{E}_{[i,j]}\left[\nabla_\theta\zeta_{ij}\left(\boldsymbol{\theta}^0\right)\sum\limits_{\iota_1+\iota_2=\iota,0<\iota_1,\iota_2\leq\iota}\left[\int\int u_1^{\iota_1}u_2^{\iota_2}K\left(\mathbf{u}\right)d\mathbf{u}\right]\vartheta_{\iota_1,\iota_2}\left(Z_{1i},Z_{2j}\right)\right], & \upsilon=3;\\ \mathbb{E}_{[i,j]}\left[\nabla_\theta\zeta_{ij}\left(\boldsymbol{\theta}^0\right)\sum\limits_{\iota_1+\iota_2=\iota,0<\iota_1,\iota_2\leq\iota}\left[\int\int u_1^{\iota_1}u_2^{\iota_2}K\left(\mathbf{u}\right)d\mathbf{u}\right]\vartheta_{\iota_1,\iota_2}\left(2\theta_1-Z_{1i},2\theta_2-Z_{2i}\right)\right], & \upsilon=5;\\ \mathbb{E}_{[i,j]}\left[\nabla_\theta\zeta_{ij}\left(\boldsymbol{\theta}^0\right)\sum\limits_{\iota_1+\iota_2=\iota,0<\iota_1,\iota_2\leq\iota}\left[\int\int u_1^{\iota_1}u_2^{\iota_2}K\left(\mathbf{u}\right)d\mathbf{u}\right]\vartheta_{\iota_1,\iota_2}\left(2\theta_1-Z_{1i},2\theta_2-Z_{2j}\right)\right], & \upsilon=7;\end{cases}$$

and

$$\mathcal{B}_\upsilon^h=\begin{cases}\sigma^2\mathbb{E}_{[i]}\left[-\nabla_\theta f_5\left(2\theta_1-z_{1i},2\theta_2-z_{2i}\right)/f_5^{-2}\left(2\theta_1-z_{1i},2\theta_2-z_{2i}\right)\right]\int\int K^2\left(\mathbf{u}\right)d\mathbf{u} & \upsilon=5;\\ \sigma^2\mathbb{E}_{[i,j]}\left[-\nabla_\theta f_7\left(2\theta_1-z_{1i},2\theta_2-z_{2j}\right)/f_7^{-2}\left(2\theta_1-z_{1i},2\theta_2-z_{2j}\right)\right]\int\int K^2\left(\mathbf{u}\right)d\mathbf{u} & \upsilon=7.\end{cases}$$

Finally, from Lemmas I.12, I.13 and I.14, we can show that $r_n+R_{G_n}/tr\left(S\left(h\right)\right)=o_p\left(1\right)$, which gives the desired result. *Q.E.D.*

A.3.5 Trimming

Until now, our proof has been based on the sample objective function without the second trimming component. The following theorem provides the justification for the exercise above; that is, we will show that the difference between the sample objective function with/without trimming components is asymptotically negligible and will not affect the asymptotic property of the estimator.

Following Linton and Xiao (2001), we consider the following smoothed trimming

(also see Andrews (1995)). We use $g(\cdot)$ to denote a density function that has support $[0,1]$ which satisfies that $g(0) = g(1) = 0$. In addition, we write $g(r)$ as a function with the support on $[b, 2b]$,

$$g(r) = \frac{1}{b} g\left(\frac{r}{b} - 1\right),$$

where b is the trimming parameter. Then we can define our trimming function as a function of $g_b(t)$, that is

$$G(s) = \begin{cases} 0 & s < b \\ \int_{-\infty}^{s} g(r)\, dr & b \leq s \leq 2b \\ 1 & s > 2b \end{cases}.$$

Here, we take $g(\cdot)$ to be the Beta density function as in Linton and Xiao (2001); $g(\cdot)$ can be written as,

$$g(s) = B(k+1)^{-1} t^k (1-t)^k, 0 \leq t \leq 1;$$

for some integer k, where $B(k)$ is the Beta function. When $b \leq s \leq 2b, G(s)$ can be expressed as

$$G(s) = B(k+1)^{-1} \left\{ \frac{(k!)^2}{(2k+1)!} - \sum_{l=0}^{k} \frac{(k!)^2}{(k-l)!\,(k+l-1)!} \left(\frac{s-b}{b}\right)^{k-l} \left(1 - \frac{s-b}{b}\right)^{k+l+1} \right\}.$$

Recall the sample objective function with trimming can be written as

$$\tilde{Q}_n = \frac{1}{n(n-1)} \sum\nolimits_{i \neq j} \left[\sum\nolimits_{v=1}^{8} \hat{\varphi}_{v,ij}(\theta)\right]^2 \hat{\tau}_{ij} G\left(\min_{v} \hat{\varphi}_{v,ij}\right),$$

and the sample objective function without trimming can be written as

$$Q_n = \frac{1}{n(n-1)} \sum\nolimits_{i \neq j} \left[\sum\nolimits_{v=1}^{8} \kappa_v \hat{\varphi}_{v,ij}(\theta)\right]^2 \hat{\tau}_{ij}.$$

Then we will show that

$$Q_n(\boldsymbol{\theta}) - \tilde{Q}_n(\boldsymbol{\theta}) = \frac{1}{n(n-1)} \sum\nolimits_{i \neq j} \left[\sum\nolimits_{v=1}^{8} \hat{\varphi}_{v,ij}(\boldsymbol{\theta}) \right]^2 \hat{\tau}_{ij}(\boldsymbol{\theta}) \left(1 - G\left(\min_{v} \hat{\varphi}_{v,ij}(\boldsymbol{\theta}) \right) \right).$$

Proposition A.15. *Suppose that Assumption 4 holds. Then, we have* $Q_n(\boldsymbol{\theta}) - \tilde{Q}_n(\boldsymbol{\theta}) = O_p\left(n^{-1}b^{1/2}\right)$

Proof of Proposition A.15: Here we use ψ_n to denote $Q_n(\boldsymbol{\theta}) - \tilde{Q}_n(\boldsymbol{\theta})$. Due to the i.i.d. assumption, we need only calculate the second moment of ψ_n. Define $\hat{\varphi}_{v^*,ij} = \min_v \hat{\varphi}_{v,ij}$. For simplicity, we omit $\boldsymbol{\theta}$ in the expression. Note that

$$\psi_n = \left(\frac{1}{n(n-1)} \right) \sum\nolimits_{i \neq j} \left[\sum\nolimits_{v=1}^{8} \hat{\varphi}_{v,ij} \right]^2 \left[1 - G\left(\hat{\varphi}_{v^*,ij} \right) \right].$$

Note that from Assumption 6, we have

$$G\left(\hat{\varphi}_{v^*,ij}\right) - G\left(\varphi_{v^*,ij}\right) = \sum_{l=0}^{L-1} \frac{1}{l!} g^{(l)}\left(\varphi_{v^*,ij}\right) \left(\hat{\varphi}_{v^*,ij} - \varphi_{v^*,ij}\right)^l + \frac{1}{L!} g^{(L)}\left(\varphi_{v^*,ij}\right) \left(\hat{\varphi}_{v^*,ij} - \varphi_{v^*,ij}\right)^L.$$

Then it follows that

$$\begin{aligned}
\psi_n &= \left(\frac{1}{n(n-1)} \right) \sum\nolimits_{i \neq j} \left[\sum\nolimits_{v=1}^{8} \hat{\varphi}_{v,ij}(\theta) \right]^2 \left[1 - G\left(\hat{\varphi}_{v^*,ij} \right) \right] \\
&= \left(\frac{1}{n(n-1)} \right) \sum\nolimits_{i \neq j} \left[\sum\nolimits_{v=1}^{8} \hat{\varphi}_{v,ij}(\theta) \right]^2 \left[1 - G\left(\varphi_{v^*,ij} \right) \right] \\
&\quad - \sum_{l=0}^{L-1} \frac{1}{l!} \left(\frac{1}{n(n-1)} \right) \sum\nolimits_{i \neq j} \left[\sum\nolimits_{v=1}^{8} \hat{\varphi}_{v,ij}(\theta) \right]^2 g^{(l)}\left(\varphi_{v^*,ij}\right) \left(\hat{\varphi}_{v^*,ij} - \varphi_{v^*,ij}\right)^l \\
&\quad - \frac{1}{L!} \left(\frac{1}{n(n-1)} \right) \sum\nolimits_{i \neq j} \left[\sum\nolimits_{v=1}^{8} \hat{\varphi}_{v,ij}(\theta) \right]^2 g^{(L)}\left(\varphi_{v^*,ij}\right) \left(\hat{\varphi}_{v^*,ij} - \varphi_{v^*,ij}\right)^L \\
&= \psi_{n,1} + \psi_{n,2} + \psi_{n,3}.
\end{aligned}$$

We will show that $\psi_{n,1} = O_p\left(n^{-1}b^{1/2}\right)$, $\psi_{n,2} = o_p\left(n^{-1}b^{1/2}\right)$ and $\psi_{n,3} = o_p\left(n^{-1}b^{1/2}\right)$.

Note that

$$\begin{aligned}\mathbb{E}\left[\psi_{n,1}^2\right] &= \mathbb{E}\left[\left(\frac{1}{n(n-1)}\right)^2 \sum\nolimits_{i\neq j}\left[\sum\nolimits_{\upsilon=1}^{8}\hat{\varphi}_{\upsilon,ij}(\theta)\right]^4 \left[1-G\left(\varphi_{\upsilon^*,ij}\right)\right]^2\right] \\ &\leq \frac{C}{n(n-1)}\mathbb{E}\left[1-G\left(\varphi_{\upsilon^*,ij}\right)\right]^2.\end{aligned}$$

by the fact $\sum_{\upsilon=1}^{8}\hat{\varphi}_{\upsilon,ij}(\theta)$ are bounded and the i.i.d assumption. Then, we can show that

$$\begin{aligned}\mathbb{E}\left[1-G\left(\varphi_{\upsilon^*,ij}\right)\right]^2 &= \int_0^b \left[1-G_b(s)\right]^2 ds + \int_b^{2b}\left[1-G_b(s)\right]^2 ds + \int_b^1 \left[1-G_b(s)\right]^2 ds \\ &\approx b + \int_b^{2b}\left[c(b)\sum_{l=0}^{2k+1}(s-b)^l b^{-l}\right] ds \\ &\approx \tilde{c}(b)b;\end{aligned}$$

for the constant $\tilde{c}(b)$, where the first equation follows from writing out the expectation; the second approximate equality follows from the fact that $G(s)=1$, for $s>2b$ and express $G(s)$ as $(2k+1)$th order polynomial in $(s-b)/b$; the last equality follows from by exchanging the integral and calculating the integral; that is,

$$\begin{aligned}\int_b^{2b}\left[\sum_{l=0}^{2k+1} c_l(b)(s-b)^l b^{-l}\right] ds &= \sum_{l=0}^{2k+1}\int_b^{2b} c_l(b)(s-b)^l b^{-l} ds \\ &= \sum_{l=0}^{2k+1}\frac{1}{l+1}c_l(b)(s-b)^{l+1}b^{-l}\Big|_{s=b}^{2b} \\ &= b\sum_{l=0}^{2k+1} c_l(b);\end{aligned}$$

where $c_l(b)$ is some function of b in the approximation. Then the desired result follows $\psi_{n,1}=O_p\left(n^{-1}b^{1/2}\right)$. Following the same steps, we can show that $\psi_{n,2}=o_p\left(n^{-1}b^{1/2}\right)$ and $\psi_{n,3}=o_p\left(n^{-1}b^{1/2}\right)$. For brevity, we omit the proofs here. *Q.E.D.*

BIBLIOGRAPHY

BIBLIOGRAPHY

Abrevaya, J. (1999): "Leapfrog Estimation of a Fixed-Effects Model with Unknown Transformation of the Dependent Variable," *Journal of Econometrics*, 93(2), 203–228.

——— (2000): "Rank Estimation of a Generalized Fixed-Effects Regression Model," *Journal of Econometrics*, 95(1), 1–23.

——— (2003): "Pairwise-Difference Rank Estimation of the Transformation Model," *Journal of Business and Economic Statistics*, 21(3), 437–47.

Abrevaya, J., J. A. Hausman, and S. Khan (2010): "Testing for Causal Effects in a Generalized Regression Model with Endogenous Regressors," *Econometrica*, 78(6), 2043–2061.

Ackerberg, D. A., and G. Gowrisankaran (2006): "Quantifying Equilibrium Network Externalities in the ACH Banking Industry," *RAND Journal of Economics*, 37(3), 738–761.

Ahn, H., and J. L. Powell (1993): "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism," *Journal of Econometrics*, 58(1), 3–29.

Andrews, D. W. (1994): "Asymptotics for Semiparametric Econometric Models via Stochastic Equicontinuity," *Econometrica*, 62(1), 43–72.

——— (1995): "Nonparametric Kernel Estimation for Semiparametric Models," *Econometric Theory*, 11(3), 560–586.

Andrews, D. W., and P. J. Barwick (2012): "Inference for Parameters Defined by Moment Inequalities: A Recommended Moment Selection Procedure," *Econometrica*, 80(6), 2805–2826.

Andrews, D. W., and M. M. Schafgans (1998): "Semiparametric Estimation of the Intercept of a Sample Selection Model," *Review of Economic Studies*, 65(3), 497–517.

Aradillas-Lopez, A. (2012): "Pairwise-Difference Estimation of Incomplete Information Games," *Journal of Econometrics*, 168(1), 120–140.

Aradillas-Lopez, A., B. E. Honoré, and J. L. Powell (2007): "Pairwise Difference Estimation with Nonparametric Contral Variables," *International Economic Review*, 48(4), 1119–1158.

Bai, J., and S. Ng (2001): "A Consistent Test for Conditional Symmetry in Time Series Models," *Journal of Econometrics*, 103(1), 225–258.

Bajari, P., H. Hong, J. Krainer, and D. Nekipelov (2010): "Estimating Static Models of Strategic Interactions," *Journal of Business and Economic Statistics*, 28(4), 469–482.

Bajari, P., H. Hong, and S. P. Ryan (2010): "Identification and Estimation of a Discrete Game of Complete Information," *Econometrica*, 78(5), 1529–1568.

Berry, S., and E. Tamer (2006): "Identification in Models of Oligopoly Entry," in *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress*, ed. by R. Blundell, W. K. Newey, and T. Persson. Econometric Society Monographs, No. 42. Cambridge and New York: Cambridge University Press.

Berry, S. T. (1992): "Estimation of a Model of Entry in the Airline Industry," *Econometrica*, 60(4), 889–917.

Bhattacharya, J., A. M. Shaikh, and E. Vytlacil (2008): "Treatment Effect Bounds under Monotonicity Assumptions: An Application to Swan-Ganz Catheterization," *American Economic Review*, 98(2), 351–56.

Bjorn, P. A., and Q. H. Vuong (1984): "Simultaneous Equations Models for Dummy Endogenous Variables: A Game Theoretic Formulation with an Application to Labor Force Participation," Discussion paper, Division of the Humanities and Social Sciences, Caltech.

Blundell, R., and J. L. Powell (2003): "Endogeneity in Nonparametric and Semiparametric Regression Models," in *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, ed. by D. M. Kreps, and K. F. Wallis. Econometric Society Monographs, No. 36. Cambridge and New York: Cambridge University Press.

Bresnahan, T. F., and P. C. Reiss (1990): "Entry in Monopoly Market," *Review of Economic Studies*, 57(4), 531–553.

——— (1991a): "Empirical Models of Discrete Games," *Journal of Econometrics*, 48(1), 57–81.

——— (1991b): "Entry and Competition in Concentrated Markets," *Journal of Political Economy*, 99(5), 977–1009.

Brock, W. A., and S. N. Durlauf (2001): "Discrete Choice with Social Interactions," *Review of Economic Studies*, 68(2), 235–260.

Buchinsky, M., and J. Hahn (1998): "An Alternative Estimator for the Censored Quantile Regression Model," *Econometrica*, 66(3), 653–671.

Card, D., and L. Giuliano (2013): "Peer Effects and Multiple Equilibria in the Risky Behavior of Friends," *Review of Economics and Statistics*, 95(4), 1130–1149.

Carrell, S. E., B. I. Sacerdote, and J. E. West (2011): "From Natural Variation to Optimal Policy? the Lucas Critique Meets Peer Effects," Discussion paper, National Bureau of Economic Research.

Cattaneo, M. D., R. K. Crump, and M. Jansson (2013): "Generalized Jackknife Estimators of Weighted Average Derivatives," *Journal of the American Statistical Association*, 108(504), 1243–1256.

Cavanagh, C., and R. P. Sherman (1998): "Rank Estimators for Monotonic Index Models," *Journal of Econometrics*, 84(2), 351–381.

Chamberlain, G. (1986): "Asymptotic Efficiency in Semi-parametric Models with Censoring," *Journal of Econometrics*, 32(2), 189–218.

Chen, S. (1999a): "Distribution-Free Estimation of the Random Coefficient Dummy Endogenous Variable Model," *Journal of Econometrics*, 91(1), 171–199.

——— (1999b): "Semiparametric Estimation of a Location Parameter in the Binary Choice Model," *Econometric Theory*, 15(01), 79–98.

——— (2000): "Rank Estimation of a Location Parameter in the Binary Choice Model," *Journal of Econometrics*, 98(2), 317–334.

Chen, S., and S. Khan (2003): "Semiparametric Estimation of a Heteroskedastic Sample Selection Model," *Econometric Theory*, 19(06), 1040–1064.

Chen, S., S. Khan, and X. Tang (2013): "Informational Content of Special Regressors in Heteroskedastic Binary Response Models," Discussion paper, Economic Research Initiatives at Duke (ERID), Duke University.

Chen, S., and Y. Zhou (2010): "Semiparametric and Nonparametric Estimation of Sample Selection Models under Symmetry," *Journal of Econometrics*, 157(1), 143–150.

Chernozhukov, V., H. Hong, and E. Tamer (2007): "Estimation and Confidence Regions for Parameter Sets in Econometric Models," *Econometrica*, 75(5), 1243–1284.

Chesher, A. (2005): "Nonparametric Identification under Discrete Variation," *Econometrica*, 73(5), 1525–1550.

——— (2010): "Instrumental Variable Models for Discrete Outcomes," *Econometrica*, 78(2), 575–601.

Ciliberto, F., and E. Tamer (2009): "Market Structure and Multiple Equilibria in Airline Markets," *Econometrica*, 77(6), 1791–1828.

Cosslett, S. R. (1997): "Nonparametric Maximum Likelihood Methods," in *Robust Inference*, ed. by G. S. Maddala, and C. R. Rao. Handbook of Statistics series, vol. 15. Amsterdam; New York and Oxford: Elsevier Science, North-Holland.

De Paula, A., and X. Tang (2012): "Inference of Signs of Interaction Effects in Simultaneous Games with Incomplete Information," *Econometrica*, 80(1), 143–172.

Dunker, F., S. Hoderlein, and H. Kaido (2013): "Random Coefficients in Static Games of Complete Information," Discussion paper, Centre for Microdata Methods and Practice (CEMMAP), University College London.

Ellickson, P. B., S. Houghton, and C. Timmins (2013): "Estimating Network Economies in Retail Chains: A Revealed Preference Approach," *RAND Journal of Economics*, 44(2), 169–193.

Ellickson, P. B., and S. Misra (2011): "Estimating Discrete Games," *Marketing Science*, 30(6), 997–1010.

Fang, K.-T., S. Kotz, and K. W. Ng (1990): *Symmetric Multivariate and Related Distributions*. London: Chapman and Hall.

Fox, J. T., and N. Lazzati (2013): "Identification of Discrete Choice Models for Bundles and Binary Games," Discussion paper, University of Michigan, Ann Arbor.

Frühwirth-Schnatter, S. (2006): *Finite Mixture and Markov Switching Models: Modeling and Applications to Random Processes*. New York : Springer.

Han, A. K. (1987): "Non-parametric Analysis of a Generalized Regression Model: The Maximum Rank Correlation Estimator," *Journal of Econometrics*, 35(2), 303–316.

Hansen, B. E. (2008): "Uniform Convergence Rates for Kernel Estimation with Dependent Data," *Econometric Theory*, 24(3), 726–748.

——— (2014): *Econometrics*. mimeo, University of Wisconsin, Madison.

Heckman, J. J. (1978): "Dummy Endogenous Variables in a Simultaneous Equation System," *Econometrica*, 46(4), 931–959.

——— (1990): "Varieties of Selection Bias," *American Economic Review*, 80(2), 313–18.

Holmes, T. J. (2011): "The Diffusion of Wal-Mart and Economies of Density," *Econometrica*, 79(1), 253–302.

Hong, H., and M. Shum (2010): "Pairwise-Difference Estimation of a Dynamic Optimization Model," *Review of Economic Studies*, 77(1), 273–304.

Hong, H., and E. Tamer (2003): "Inference in Censored Models with Endogenous Regressors," *Econometrica*, 71(3), 905–932.

Honoré, B. E. (1992): "Trimmed LAD and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects," *Econometrica*, 60(3), 533–565.

Honoré, B. E., and L. Hu (2004): "Estimation of Cross Sectional and Panel Data Censored Regression Models with Endogeneity," *Journal of Econometrics*, 122(2), 293–316.

Honoré, B. E., and E. Kyriazidou (2000): "Panel Data Discrete Choice Models with Lagged Dependent Variables," *Econometrica*, 68(4), 839–874.

Honoré, B. E., E. Kyriazidou, and C. Udry (1997): "Estimation of Type 3 Tobit Models Using Symmetric Trimming and Pairwise Comparisons," *Journal of Econometrics*, 76(1), 107–128.

Honoré, B. E., and J. Powell (1997): "Pairwise Difference Estimators for Nonlinear Models," Discussion paper, University of California, Berkeley.

Honoré, B. E., and J. L. Powell (1994): "Pairwise Difference Estimators of Censored and Truncated Regression Models," *Journal of Econometrics*, 64(1), 241–278.

Horowitz, J. L. (1992): "A Smoothed Maximum Score Estimator for the Binary Response Model," *Econometrica*, 60(3), 505–531.

Horowitz, J. L., and W. Härdle (1996): "Direct Semiparametric Estimation of Single-Index Models with Discrete Covariates," *Journal of the American Statistical Association*, 91(436), 1632–1640.

Hu, L. (2002): "Estimation of a Censored Dynamic Panel Data Model," *Econometrica*, 70(6), 2499–2517.

Huang, C.-I. (2013): "Intra-household Effects on Demand for Telephone Service: Empirical Evidence," *Quantitative Marketing and Economics*, 11(2), 231–261.

Hult, H., and F. Lindskog (2002): "Multivariate Extremes, Aggregation and Dependence in Elliptical Distributions," *Advances in Applied Probability*, 34(3), 587–608.

Ichimura, H. (1993): "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models," *Journal of Econometrics*, 58(1), 71–120.

Ichimura, H., and P. E. Todd (2007): "Implementing Nonparametric and Semiparametric Estimators," in *Handbook of Econometrics*, ed. by J. J. Heckman, and E. E. Leamer. Amsterdam; London and New York: Elsevier, North-Holland.

Imbens, G. W., and W. K. Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations Models without Additivity," *Econometrica*, 77(5), 1481–1512.

Imbens, G. W., and G. Ridder (2009): "Estimation and Inference for Generalized Full and Partial Means and Average Derivatives," Discussion paper, Harvard University.

Jia, P. (2008): "What Happens When Wal-Mart Comes to Town: An Empirical Analysis of the Discount Retailing Industry," *Econometrica*, 76(6), 1263–1316.

Jochmans, K. (2011): "Identification in Bivariate Binary-Choice Models with Elliptical Innovations," Discussion paper, Sciences Po.

Kendall, M. (1938): "A New Measure of Rank Correlation," *Biometrika*, 30, 81–93.

Khan, S., and D. Nekipelov (2010): "Information Bounds and Impossibility Theorems for Simultaneous Discrete Response Models," Discussion paper, Duke University.

——— (2012): "Information Structure and Statistical Information in Discrete Response Models," Discussion paper, Economic Research Initiatives at Duke (ERID), Duke University.

Khan, S., and E. Tamer (2007): "Irregular Identification, Support Conditions and Inverse Weight Estimation," Discussion paper, Duke University.

——— (2010): "Irregular Identification, Support Conditions, and Inverse Weight Estimation," *Econometrica*, 78(6), 2021–2042.

Kline, B. (2012): "Identification of Complete Information Games," Discussion paper, University of Texas, Austin.

Kyriazidou, E. (1997): "Estimation of a Panel Data Sample Selection Model," *Econometrica*, 65(6), 1335–1364.

Lee, M.-j. (1996): "Nonparametric Two-stage Estimation of Simultaneous Equations with Limited Endogenous Regressors," *Econometric Theory*, 12(2), 305–330.

Lewbel, A. (1997): "Semiparametric Estimation of Location and Other Discrete Choice Moments," *Econometric Theory*, 13(1), 32–51.

——— (1998): "Semiparametric Latent Variable Model Estimation with Endogenous or Mismeasured Regressors," *Econometrica*, 66(1), 105–121.

——— (2000): "Semiparametric Qualitative Response Model Estimation with Unknown Heteroscedasticity or Instrumental Variables," *Journal of Econometrics*, 97(1), 145–177.

Lewbel, A., and X. Tang (2013): "Identification and Estimation of Games with Incomplete Information Using Excluded Regressors," Discussion paper, University of Pennsylvania.

Linton, O., and Z. Xiao (2001): "Second-order Approximation for Adaptive Regression Estimators," *Econometric Theory*, 17(5), 984–1024.

Manski, C. F. (1985): "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator," *Journal of Econometrics*, 27(3), 313–333.

——— (1993): "Identification of Endogenous Social Effects: The Reflection Problem," *Review of Economic Studies*, 60(3), 531–542.

Manuszak, M. D., and A. Cohen (2004): "Endogenous Market Structure with Discrete Product Differentiation and Multiple Equilibria: An Empirical Analysis of Competition Between Banks and Thrifts," Discussion paper, Carnegie Mellon University.

Moffitt, R. A., et al. (2001): "Policy Interventions, Low-level Equilibria, and Social Interactions," *Social Dynamics*, 4, 45–82.

Müller, H.-G. (1988): *Nonparametric Regression Analysis of Longitudinal Data*. Berlin; New York: Springer-Verlag.

Nelsen, R. B. (1993): "Some Concepts of Bivariate Symmetry," *Journal of Nonparametric Statistics*, 3(1), 95–101.

Newey, W. K. (1991): "Efficient Estimation of Tobit Models under Conditional Symmetry," in *Nonparametric and Semiparametric Methods in Econometrics and Statistics: Proceedings of the Fifth International Symposium in Economic Theory and Econometrics*, ed. by W. A. Barnett, J. Powell, and G. E. Tauchen. Cambridge; New York and Melbourne: Cambridge University Press.

——— (1994): "Kernel Estimation of Partial Means and a General Variance Estimator," *Econometric Theory*, 10(2), 1–21.

Newey, W. K., and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, ed. by Z. Griliches, R. F. Engle, M. D. Intriligator, and D. McFadden. Amsterdam; London and New York: Elsevier, North-Holland.

Newey, W. K., and J. L. Powell (2003): "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71(5), 1565–1578.

Orhun, A. Y. (2013): "Spatial Differentiation in the Supermarket Industry: The Role of Common Information," *Quantitative Marketing and Economics*, 11(1), 3–37.

Pakes, A., J. Porter, K. Ho, and J. Ishii (2011): "Moment Inequalities and Their Application," Discussion paper, Harvard University.

Powell, J. L. (1986): "Symmetrically Trimmed Least Squares Estimation for Tobit Models," *Econometrica*, 54(6), 1435–1460.

Powell, J. L., J. H. Stock, and T. M. Stoker (1989): "Semiparametric Estimation of Index Coefficients," *Econometrica*, 57(6), 1403–1430.

Powell, J. L., and T. M. Stoker (1996): "Optimal Bandwidth Choice for Density-Weighted Averages," *Journal of Econometrics*, 75(2), 291–316.

Romano, J. P., and A. M. Shaikh (2010): "Inference for the Identified Set in Partially Identified Econometric Models," *Econometrica*, 78(1), 169–211.

Ryan, S. P., and C. Tucker (2012): "Heterogeneity and the Dynamics of Technology Adoption," *Quantitative Marketing and Economics*, 10(1), 63–109.

Serfling, R. J. (1980): *Approximation Theorems of Mathematical Statistics*. New York: John Wiley and Sons.

——— (2006): "Multivariate Symmetry and Asymmetry," *Encyclopedia of Statistical Sciences*.

Shaikh, A. M., and E. J. Vytlacil (2011): "Partial Identification in Triangular Systems of Equations with Binary Dependent Variables," *Econometrica*, 79(3), 949–955.

Sherman, R. P. (1994): "U-Processes in the Analysis of a Generalized Semiparametric Regression Estimator," *Econometric Theory*, 10(2), 372–395.

Silverman, B. W. (1986): *Density Estimation for Statistics and Data Analysis*. London; New York: Chapman and Hall.

Soetevent, A. R., and P. Kooreman (2007): "A Discrete-Choice Model with Social Interactions: With an Application to High School Teen Behavior," *Journal of Applied Econometrics*, 22(3), 599–624.

Sweeting, A. (2006): "Coordination, Differentiation, and the Timing of Radio Commercials," *Journal of Economics and Management Strategy*, 15(4), 909–942.

Tamer, E. (2003): "Incomplete Simultaneous Discrete Response Model with Multiple Equilibria," *Review of Economic Studies*, 70(1), 147–165.

Vytlacil, E., and N. Yildiz (2007): "Dummy Endogenous Variables in Weakly Separable Models," *Econometrica*, 75(3), 757–779.

Wan, Y., and H. Xu (2012): "Semiparametric Estimation of Binary Decision Games of Incomplete Information with Correlated Private Signals," Discussion paper, Pennsylvania State University.

Wand, M. P., and M. C. Jones (1994): *Kernel Smoothing*. London; New York: Chapman and Hall/CRC.

Zhou, Y. (2014a): "Identification and Estimation of Entry Games with Symmetry of Unobservables: A Theoretical Perspective," Discussion paper, University of Michigan, Ann Arbor.

——— (2014b): "Identification and Estimation of Entry Games with Symmetry of Unobservables: A Simulation Design," Discussion paper, University of Michigan, Ann Arbor.

Zhu, T., and V. Singh (2009): "Spatial Competition with Endogenous Location Choices: An Application to Discount Retailing," *Quantitative Marketing and Economics*, 7(1), 1–35.