Identification and Estimation of Binary Choice Models with Social Interactions and Unknown Group Structures[*]

Yonghong An[†] Wenzheng Gao[‡] Naibao Zhao[§]

December 6, 2025

Abstract

This paper studies identification and estimation of a game-theoretic binary choice model with social interactions where the group structure is latent. By exploiting two proxies for the group characteristics, we show that peer effects can be identified even though the group structure are *unobservable*. Based on the identification method, a semiparametric nonlinear least square estimator is proposed. Monte Carlo experiments demonstrate that the semiparametric estimator has good finite-sample performance. In the empirical application of our method, we find positive and significant peer effects among students in their choices regarding private schoolwork tutoring, and show that ignoring the latent group structures can lead to significant bias in the estimation of peer effects.

Keywords: Binary Choice; Peer Effect; Incomplete Information; Identification

JEL classification: C14; C31; C57

[*] This work was supported by the National Natural Science Foundation of China (Grant Numbers 72103101 and 72273017). We thank Toru Kitagawa, Xiaojun Song, Zhenting Sun, and all seminar participants for their constructive comments and suggestions. We also thank Yuqian Sun for her excellent research assistance. Any remaining errors are our own responsibility.

[†] Department of Economics, Texas A&M University. Email: yonghongan@tamu.edu.

[‡] School of Economics, Nankai University. Email: wenzhenggao@nankai.edu.cn.

[§] Department of Marketing, Texas A&M University. Email: nzhao@mays.tamu.edu.

1 Introduction

Social interaction models study how economic agents within well-defined reference groups (e.g., friends, colleagues, or neighbors) strategically interact with each other through their decision-making processes with respect to socioeconomic activities. Manski (1993) characterizes the impacts of such strategic interactions as the influence of individuals' decisions (peer effects), observable characteristics of group members (contextual effects) and unobserved group heterogeneity (correlated effects). Estimating peer effects is important for policy analyses because it can generate a "social multiplier" where aggregate relationships will overstate individual elasticities (Glaeser et al. 2003). Recent empirical studies have found evidence of peer effects on crime (Glaeser et al. 1996), adolescent behavior (Gaviria and Raphael 2001, Nakajima 2007), retirement savings (Duflo and Saez 2002), in-school achievements (Calvó-Armengol et al. 2009), firm financial policy (Leary and Roberts 2014), and product adoption (Bailey et al. 2022), among others. However, an empirical challenge that applied researchers often encounter when analyzing such models is the unknown nature of group structures. This problem arises because the data contains either no or inaccurate information about group memberships[1]. Without prior information specifying the composition of reference groups, it is impossible to conduct inference on peer effects (Manski 1993).

In this paper, we aim to address the empirical challenge by introducing an econometric method designed to uncover peer effects while incorporating *unobservable* group structures. We model peer effects as the influence of binary choices made by members within the group. By employing two proxies and a monotonicity condition for the unobserved group heterogeneity, we demonstrate that it is possible to identify and estimate peer effects even in the absence of known group structures. A noteworthy feature of our method, setting it apart from previous studies, is its capability to distinguish peer effects from contextual plus correlated effects in this type of models. This distinction has significant policy implications because contextual and correlated effects do not generate the social multiplier.

The social interaction model under consideration is an incomplete information game-theoretic model with binary choices, resembling the one presented in Brock and Durlauf (2007). In the model, each individual's payoff function consists of four components: direct effects from their own characteristics, peer effects from the subjective expectation of average choices made by group members, contextual/correlated effects from unobserved group heterogeneity[2], and a stochastic component representing payoff shocks, assumed to be private information with a commonly known distribution. Our goal is to recover the parameters associated with these components. Under the condition of mild peer effects, we establish the existence of a unique rational expectation

[1] The inaccuracy primarily results from measurement errors, mainly stemming from sources of group structure data, which are predominantly surveys and questionnaires soliciting self-reports (Marsden 1990).

[2] Due to latent group structures, the characteristics of group members become unobservable as well. Therefore, the unobserved group heterogeneity in our model incorporates both contextual and correlated effects.

equilibrium, which can help avoid the incompleteness problem discussed in Tamer (2003).

Our identification strategy proceeds in two steps. First, we identify the subjective expectations of average choices made by group members. In this step, we conduct the following: (1) Under the rational expectation equilibrium, these subjective expectations are equivalent to the conditional expectations of binary choices based on the unobserved group heterogeneity. (2) We nonparametrically identify the conditional expectations via the matrix decomposition method in the measurement error literature (e.g., Hu 2008, 2017). This method requires two proxies for the unobserved group heterogeneity[3] and a monotonicity condition. Second, we identify payoff parameters by exploring the one-to-one mapping between subjective expectations and model parameters implied by the model structure. This mapping allows us to pin down a linear relationship between structural parameters. Then, we are able to identify direct, peer and contextual/correlated effects by utilizing the variations of individual characteristics and a normalization condition for the support of unobserved group heterogeneity.

Based on this identification procedure, we propose a semiparametric nonlinear least squares (SNLS) estimator for model parameters. Note that under the setup of our model, the data exhibits a locally dependent structure, wherein observations within each group are interdependent. This dependence structure is unknown, as we lack information about the composition of groups[4]. Despite the unknown dependence structure, we establish that the SNLS estimator is still root-n consistent and asymptotically normal. The inference procedure for model parameters relies on the dependence-robust subsampling method introduced in Song (2016) and Leung (2022). Monte Carlo experiments demonstrate the good performance of our proposed estimator and inference method in finite samples.

In the empirical application of the method developed in this paper, we investigate peer effects in the decisions of secondary school students to participate in private tutoring, utilizing data from the China Education Panel Survey (CEPS). CEPS is a nationally representative longitudinal survey that contains rich information about secondary school students, their parents, and teachers. We choose class advisors' subjective evaluation of class performance as proxies for the unobserved group heterogeneity. Our estimates suggest that there are positive and significant peer effects among students in their choices related to private schoolwork tutoring. In comparison, we also estimate models that naïvely treat classrooms as reference groups, and we find that estimated peer effects become insignificant, which demonstrates the empirical importance of incorporating unknown group structures.

This paper contributes to three strands of the literature. First, it is naturally related to studies of social interaction models with discrete outcomes, which have been extensively investigated since the pioneering work of Brock and Durlauf (2001, 2007). They propose a novel equilibrium char-

[3] These proxies can be other outcome variables or contaminated measurements of group characteristics.

[4] This is analogous to clustered data with unknown cluster memberships.

acterization of discrete choice models with social interactions. The key feature of their model is that individuals will form homogeneous rational expectations regarding the behaviors of all other members in the same group. Soetevent and Kooreman (2007) adopts a complete-information game framework to analyze peer effects in discrete choice models and uses simulated maximum likelihood for estimation. Lee et al. (2014) extend the model of Brock and Durlauf (2001, 2007) to allow for heterogeneous rational expectations based on publicly known characteristics and propose a maximum likelihood method to estimate model parameters. Yang and Lee (2017) further allow the heterogeneous expectations to depend on asymmetric private information. Xu (2018) employs a simultaneous game of incomplete information to study large network-based social interactions. Lin et al. (2021) identify and estimate heterogeneous social effects in binary choices with unknown network structures.[5] The key assumption of their identification strategy is that the latent network structures are functions of observable characteristics of group members, essentially ruling out correlated effects. Our paper differs from the specifications in these papers in that it allows for both unknown group structures and correlated effects.

Second, the paper also contributes to the literature on empirical games with incomplete information. Aguirregabiria and Mira (2007) employ a nested pseudo likelihood method to estimate dynamic discrete games of incomplete information. Aradillas-Lopez (2010) estimate a model where players' private values are independent of public information in the game. Bajari et al. (2010) use exclusion restrictions to identify static games with multiple equilibria. De Paula and Tang (2012) propose a test for multiple Bayesian Nash equilibria in discrete simultaneous games with incomplete information. Wan and Xu (2014) study semiparametric identification of binary decision games with two players and correlated private values. Lewbel and Tang (2015) show that it is possible to nonparametrically identify binary games with incomplete information using excluded regressors. Menzel (2016) develops asymptotic theory for discrete games with a large number of players. Aguirregabiria and Mira (2019) identify games with both multiple equilibria and unobserved heterogeneity. In comparison, we investigate an incomplete-information game involving simultaneous binary choices and social interactions. Our identification strategy is unique, utilizing the eigenvalue–eigenvector decomposition technique to nonparametrically identify equilibrium beliefs and exploring their structural links with model parameters subsequently.

Third, the paper also enriches the literature on nonparametric identification of measurement error models[6] and its applications in microeconomic models with latent variables such as auctions (Li et al. 2000, An et al. 2010, Hu et al. 2013, An 2017), incomplete-information games with multiple equilibria (Xiao 2018, Luo et al. 2022), dynamic discrete choices (Hu and Shum 2012, An et al. 2021), production functions for cognitive and noncognitive skills (Cunha et al. 2010), and two-sided matching models (Diamond and Agarwal 2017). To the best of our knowledge, only

[5] Lewbel et al. (2023) study identification and estimation of linear social network models without observing any network links where outcomes are continuous.

[6] See Hu (2017) for an overview of the literature on the nonparametric identification of measurement error models.

two papers have also employed measurement error approaches to study network models with social interactions. Lin and Hu (2024) study a binary social interaction model with misclassification errors in outcome variables but require knowledge of group structures. On the other hand, Zhang (2020) nonparametrically identify the reduced-form spillover effects of treatment responses in social networks with missing links. In contrast, our paper focuses on the structural model of social interactions, which is crucial for researchers to determine the mechanism through which peer effects influence personal outcomes[7].

The rest of the paper is organized as follows. Section 2 introduces the setting and basic assumptions of our model. Section 3 presents a constructive identification procedure. In Section 4, we propose an SNLS estimator for model parameters and establish its asymptotic properties. A subsampling inference method is also discussed. Section 5 discusses two extensions of our identification method. The finite-sample performance of the SNLS estimator and the subsampling method is examined through Monte Carlo simulations in Section 6. Section 7 contains the empirical analysis of peer effects in private tutoring. Section 8 concludes. All proofs are provided in the Appendix.

Notations: Throughout the paper, all random elements are defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. For a random variable X with support $\operatorname{supp}(X)$, we use $F_X(x)$ and $f_X(x)$ to represent its cumulative distribution function (CDF) and probability density function (PDF) evaluated at $x \in \operatorname{supp}(X)$. Let $\xrightarrow{p}$ and $\xrightarrow{d}$ denote convergence in probability and distribution, respectively.

2 The Model

We consider a binary choice social interaction model with rational expectations similar to Brock and Durlauf (2001, 2007). The sample consists of n individuals within social groups, such as friends, classmates and colleagues. There are G social groups in the population and each individual i belongs to one of these G groups. The size of group g is n_g, i.e., $|\mathcal{N}_g| = n_g$. Consequently, $\sum_{g=1}^{G} n_g = n$. In this paper, we assume that the group memberships are unknown to researchers, i.e., we do not know which group individual i belong to for $i = 1, 2, ..., n$. Thus, G is also unknown to researchers.

Individual choices are coded by $Y_i \in \{0, 1\}$. We assume that the payoffs for $Y = 1$, $u_i(1)$ is additive in the various factors, i.e.,

$$u_i(1) = \alpha + \beta m_{i,g}^e + \gamma X_i + Z_g^* - \epsilon_i. \tag{2.1}$$

Following the literature of binary choice model, we impose the normalization restriction that $u_i(0) = 0$. Equation (2.1) is determined by four factors: Observable individual-specific charac-

[7] See Manski (2013) for a detailed comparison of reduced-form and structural approaches to social interaction models.

teristics X_i, which is a $d_x \times 1$ vector of discrete random variable; Subjective expectation by agent i of $\bar{Y}_g$, the average choice in the group, described by the value $m_{i,g}^e$; this is known as the *peer effects* as it describes how the behaviors of others affect each individual. The unobserved group heterogeneity, measured by a discrete scalar Z_g^*. It relates to how characteristics of a group affect its members. Note that it represents both contextual and correlated effects, as characteristics of group members are unobservable due to unknown group structures and are therefore subsumed in Z_g^*. We assume that the distribution of the unobserved heterogeneity Z_g^* is the same across g. This is an alternative approach to deal with unobserved network. The last one is the unobservable individual characteristics summarized by a scalar ϵ_i, which is assumed to be independently drawn from a known distribution (e.g., standard normal or logistic) with strictly increasing CDF F_ϵ and PDF f_ϵ that is bounded above by $\sup_e f_\epsilon(e) < \infty$. Besides, the known distribution does not depend on X_i, Z_g.

We assume that subjective beliefs are rational, given information of Z_g^* and $F_{X|Z^*}$, the CDF of X_i conditional Z_g^*. Therefore, the subjective expectations $m_{i,g}^e$ coincide with m_g, the mathematical expectation of the average choice in group g given Z_g^*. Since

$$\mathbb{E}\left(Y_i | X_i, Z_g^*\right) = F_\epsilon\left(\alpha + \beta m_g + \gamma X_i + Z_g^*\right), \tag{2.2}$$

m_g is then defined by

$$m_g \equiv \mathbb{E}\left(Y_i | Z_g^*, i \in g\right) = \int F_\epsilon\left(\alpha + \beta m_g + \gamma X_i + Z_g^*\right) dF_{X|Z^*}. \tag{2.3}$$

In the equilibrium, m_g can be characterized as the fixed point of (2.3). It is possible for there to exist multiple values of m_g that fulfill (2.3). However, when multiple equilibria exist, an obvious obstacle for identification and inference is the incompleteness of the econometric model (Tamer 2003). Therefore, we impose the following assumption:

Assumption 2.1. *(Unique Equilibrium) The sample is generated from a single equilibrium for all n.*

Assumption 2.1 is widely imposed for identifying and estimating an incomplete information game, see e.g., Aradillas-Lopez (2010), Bajari et al. (2010) and Lewbel and Tang (2015). It ensures that the equilibrium expected average choice can be directly identified and estimated as the conditional expectation using information from the sample.

It is worth mentioning that Assumption 2.1 can be satisfied if we restrict the strength of peer effects β not to be "too large" (Horst and Scheinkman 2006):

Lemma 2.1. *If $|\beta| < \dfrac{1}{\sup_e f_\epsilon(e)}$, the equilibrium expected average choice characterized by (2.3) will be unique.*

Proof. See Appendix. □

The upper bound of β in Lemma 2.1 guarantees that (2.3) is a contraction mapping. Similar conditions has been used in Brock and Durlauf (2001), Lee et al. (2014), Xu (2018) and Lin and Hu (2024) to show the uniqueness of equilibrium in incomplete information games. If the underlying distribution of ϵ is standard normal, the upper bound will be $\sqrt{2\pi} \approx 2.507$. For Logit-type models, the upper bound should be changed to 4.

3 Identification

In this section, we present sufficient conditions for identifying the binary choice model of social interactions with unknown group structures. Our identification strategy proceeds in two steps. First, we identify the equilibrium subjective expectations of average choices made by group members. Second, we identify payoff parameters by exploring the one-to-one mapping between subjective expectations and model parameters implied by the model structure and the nolinearity of conditional expectations. For notational simplicity, we will suppress the subscripts i and g whenever there is no ambiguity.

3.1 Identification of equilibrium subjective expectations

First, we discuss how to nonparametrically identify the equilibrium subjective expectation m, which equals to $\mathbb{E}(Y|Z^*)$ by (2.3). We employ the eigenvalue-eigenvector decomposition method in Hu (2008), relying on two discrete proxies for Z^*, denoted by Z and Z', and a monotonicity condition to achieve identification. These two proxies can be obtained as other categorical outcome variables related to social interactions or contaminated measurement of group characteristics based on self-reported and administrative network data. The following set of assumptions will be imposed in order to provide a baseline for identification analysis.

Assumption 3.1. *(Random Assignment) Individuals are randomly assigned to groups, i.e.,* $F_{X|Z^*} = F_X$

Assumption 3.1 is also imposed in Brock and Durlauf (2007) and utilizes the idea of random assignment by equating it with the independence of the distribution of individual characteristics within a group from the unobserved group heterogeneity Z_g^*. While this assumption may be appropriate for examples such as school classrooms (Eble and Hu 2022), it usually does not hold for groups such as friends or neighborhood. In Section 5 we discuss how to extend the identification results when Assumption 3.1 is not satisfied.

Under Assumption 3.1, we have

$$\mathbb{E}\left(Y|X, Z^*\right) = F_\epsilon\left(\alpha + \beta m + \gamma X + Z^*\right) \tag{3.1}$$

and

$$m = \int F_{\epsilon}\left(\alpha + \beta m + \gamma X + Z^{*}\right) dF_{X}. \tag{3.2}$$

Assumption 3.2. *(Conditional Independence)* $Z \perp Z' \perp Y | Z^{*}, X$

Assumption 3.2 requires the two proxies Z, Z' and the binary outcome Y to be independent with each other when conditioning on the latent variable Z^{*} and the observables X. This assumption can be understood as follows. Suppose $Z = h_1(X^{*}, X, e_z), Z' = h_2(X^{*}, X, e_{z'})$, and $Y = h_3(Z^{*}, X, e_y)$, where $h_j, j = 1, 2, 3$ are functions, $e_z, e_{z'}$, and e_y are random errors. Then the assumption states that the errors $e_z, e_{z'}$, and e_y are mutually independent. This assumption is commonly imposed in the measurement error literature, see, e.g., Hu (2008), Hu and Schennach (2008) and Hu (2017), among others.

Under Assumptions 3.1 and 3.2, we can represent (conditional) joint distributions of observables as mixtures of unobserved group heterogeneity Z^{*}.

Lemma 3.1. *Suppose Assumptions 3.1 and 3.2 are satisfied. Then,*
(i) $f_{Z,Z',Y|X}(z, z', y|x) = \sum_{z^{*} \in \mathrm{supp}(Z^{*})} f_{Z|Z^{*},X}(z|z^{*}, x) f_{Z'|Z^{*},X}(z'|z^{*}, x) f_{Y|Z^{*},X}(y|z^{*}, x) f_{Z^{*}}(z^{*})$;
(ii) $f_{Z,Z'|X}(z, z'|x) = \sum_{z^{*} \in \mathrm{supp}(Z^{*})} f_{Z|Z^{*},X}(z|z^{*}, x) f_{Z'|Z^{*},X}(z'|z^{*}, x) f_{Z^{*}}(z^{*})$.

Proof. See Appendix. □

Lemma 3.1 is a direct consequence of the random assignment and conditional independence assumptions. It implies that the conditional joint distributions of two proxies Z, Z' and binary outcome Y on covariates X are multiplicatively separable given Z^{*}. To identify m, we need to first recover these latent mixture components.

Assumption 3.3. *(Equal Support)* $|\mathrm{supp}(Z)| = |\mathrm{supp}(Z')| = |\mathrm{supp}(Z^{*})| = K$.

Assumption 3.3 (i) resembles the one in Hu (2017). It implies that the supports and Z, Z' and Z^{*} share the same carnality K. Otherwise, the proxies would lack sufficient information to identify the distribution of the latent variables. This assumption can be relaxed to allow $|\mathrm{supp}(Z)|$ and $|\mathrm{supp}(Z')|$ to be larger than $|\mathrm{supp}(Z^{*})|$.[8] In Section 5 we demonstrate that m can still be nonparametrically identified in this scenario.

By imposing Assumption 3.3, we can define the following $K \times K$ matrices for each $Y = y$ and $X = x$:

[8] In such cases, we can combine the possible values Z and Z' take such that Assumption 3.3 holds, see, e.g., Xiao (2018).

$$M_{Z,Z',Y|X} = \begin{bmatrix} f_{Z,Z',Y|X}(z_1, z_1', y|x) & f_{Z,Z',Y|X}(z_1, z_2', y|x) & \cdots & f_{Z,Z',Y|X}(z_1, z_K', y|x) \\ f_{Z,Z',Y|X}(z_2, z_1', y|x) & f_{Z,Z',Y|X}(z_2, z_2', y|x) & \cdots & f_{Z,Z',Y|X}(z_2, z_K', y|x) \\ \vdots & \vdots & \ddots & \vdots \\ f_{Z,Z',Y|X}(z_K, z_1', y|x) & f_{Z,Z',Y|X}(z_K, z_2', y|x) & \cdots & f_{Z,Z',Y|X}(z_K, z_K', y|x) \end{bmatrix},$$

$$M_{Z,Z'|X} = \begin{bmatrix} f_{Z,Z'|X}(z_1, z_1'|x) & f_{Z,Z'|X}(z_1, z_2'|x) & \cdots & f_{Z,Z'|X}(z_1, z_K'|x) \\ f_{Z,Z'|X}(z_2, z_1'|x) & f_{Z,Z'|X}(z_2, z_2'|x) & \cdots & f_{Z,Z'|X}(z_2, z_K'|x) \\ \vdots & \vdots & \ddots & \vdots \\ f_{Z,Z'|X}(z_K, z_1'|x) & f_{Z,Z'|X}(z_K, z_2'|x) & \cdots & f_{Z,Z'|X}(z_K, z_K'|x) \end{bmatrix},$$

$$M_{Z|Z^*,X} = \begin{bmatrix} f_{Z|Z^*,X}(z_1|z_1^*, x) & f_{Z|Z^*,X}(z_1|z_2^*, x) & \cdots & f_{Z|Z^*,X}(z_1|z_K^*, x) \\ f_{Z|Z^*,X}(z_2|z_1^*, x) & f_{Z|Z^*,X}(z_2|z_2^*, x) & \cdots & f_{Z|Z^*,X}(z_2|z_K^*, x) \\ \vdots & \vdots & \ddots & \vdots \\ f_{Z'|Z^*,X}(z_K'|z_1^*, x) & f_{Z'|Z^*,X}(z_K'|z_2^*, x) & \cdots & f_{Z'|Z^*,X}(z_K'|z_K^*, x) \end{bmatrix},$$

$$M_{Z'|Z^*,X} = \begin{bmatrix} f_{Z'|Z^*,X}(z_1'|z_1^*, x) & f_{Z'|Z^*,X}(z_1'|z_2^*, x) & \cdots & f_{Z'|Z^*,X}(z_1'|z_K^*, x) \\ f_{Z'|Z^*,X}(z_2'|z_1^*, x) & f_{Z'|Z^*,X}(z_2'|z_2^*, x) & \cdots & f_{Z'|Z^*,X}(z_2'|z_K^*, x) \\ \vdots & \vdots & \ddots & \vdots \\ f_{Z'|Z^*,X}(z_K'|z_1^*, x) & f_{Z'|Z^*,X}(z_K'|z_2^*, x) & \cdots & f_{Z'|Z^*,X}(z_K'|z_K^*, x) \end{bmatrix},$$

$$D_{Y|Z^*,X} = \begin{bmatrix} f_{Y|Z^*,X}(y|z_1^*, x) & & & \\ & f_{Y|Z^*,X}(y|z_2^*, x) & & \\ & & \ddots & \\ & & & f_{Y|Z^*,X}(y|z_K^*, x) \end{bmatrix},$$

$$D_{Z^*} = \begin{bmatrix} f_{Z^*}(z_1^*) & & & \\ & f_{Z^*}(z_2^*) & & \\ & & \ddots & \\ & & & f_{Z^*}(z_K^*) \end{bmatrix}.$$

Then, the equations in Lemma 3.1 can be rewritten into matrix expression

$$M_{Z,Z',Y|X} = M_{Z|Z^*,X} D_{Y|Z^*,X} D_{Z^*} M_{Z'|Z^*,X}^T \tag{3.3}$$

and

$$M_{Z,Z'|X} = M_{Z|Z^*,X} D_{Z^*} M_{Z'|Z^*,X}^T. \tag{3.4}$$

To use the eigen-decomposition technique, we need to ensure that the $K \times K$ matrices $M_{Z|Z^*,X}$ and $M_{Z'|Z^*,X}$ are nonsingular, which is equivalent to the following rank condition.

Assumption 3.4. *(Full Rank) $M_{Z|Z^*,X}$ and $M_{Z'|Z^*,X}$ have rank K.*

Assumption 3.4 has been adopted in prior works such as Hu (2008, 2017) and Xiao (2018). However, when the measurements Z and Z' encompass fewer values than Z^*, Assumption 3.4 fails to hold. In the subsequent lemma, we demonstrate its equivalence to assuming the invertibility of the matrix $M_{Z,Z'|X}$, which comprises observed probabilities. As a result, Assumption 3.5 can be verified by testing $H_0 : \text{rank}(M_{Z,Z'|X}) = K$. This verification process can be implemented using methods detailed in Robin and Smith (2000), Kleibergen and Paap (2006), and Chen and Fang (2019).

Lemma 3.2. *Under Assumptions 3.3, Assumption 3.4 holds if and only if the rank of the matrix $M_{Z,Z'|X}$ is K.*

Proof. See Appendix. □

Lemma 3.2 also implies that the cardinality of $\text{supp}(Z^*)$ can be identified as the rank of $M_{Z,Z'|X}$. Then, algebraic manipulations of the matrix equations (3.3) and (3.4), summarized in the proof of Proposition 3.1 below, implies that

$$M_{Z,Z',Y|X}M^{-1}_{Z,Z'|X} = M_{Z|Z^*,X}D_{Y|Z^*,X}M^{-1}_{Z|Z^*,X}. \tag{3.5}$$

This equation indicates that the observed matrix on the left hand side of (3.5) has an eigenvalue-eigenvector decomposition. Then, the conditional density matrix $D_{Y|Z^*,X}$ can be identified up to the permutation of its diagonal entries. Then, the following lemma ensures that the identification is unique.

Lemma 3.3. *Under Assumption 2.1, $f_{Y|Z^*,X}$ is strictly increasing in Z^*.*

Proof. See Appendix. □

Lemma 3.3 ensures that the eigenvalues $\{f_{Y|Z^*,X}(y|z^*, x)\}_{k=1,2,\ldots,K}$ are distinctive and rules out the case of duplicate eigenvalues. Furthermore, it also fixes the ordering of the eigenvalues and eigenvectors. This condition guarantees that the decomposition in (3.5) is unique and thus we can identify $f_{Y|Z^*,X}$.

Once $f_{Y|Z^*,X}$ is identified, the conditional expectation m^* can be identified accordingly because of the equation

$$f_{Y|Z^*} = \sum_{x\in\text{supp}(X)} f_{Y|Z^*,X}(\cdot|\cdot, x)f_{X|Z^*}(x|z^*) = \sum_{x\in\text{supp}(X)} f_{Y|Z^*,X}(\cdot|\cdot, x)f_X(x), \tag{3.6}$$

where the last equality is by Assumption 3.1.

Proposition 3.1. *Under Assumptions 3.1-3.4, the conditional density $f_{Y|Z^*,X}$ and the conditional expectation m are nonparametrically identified.*

Proof. See Appendix. □

A byproduct of the identification procedure above is the eigenvector matrix $M_{Z|Z,X}$. The identification of the conditional density $f_{Z|Z',X}$, which is necessary for the estimation method in the next section, relies on the availability of this matrix. We summarize the procedure in Lemma 3.4 below.

Lemma 3.4. *Under Assumptions 3.1-3.4, the conditional density $f_{Z^*|Z',X}$ can be nonparametrically identified as the elements of $M_{Z|Z^*,X}^{-1}M_{Z|Z',X}$, where*

$$M_{Z|Z',X} = \begin{bmatrix} f_{Z|Z',X}(z_1|z'_1,x) & f_{Z|Z',X}(z_1|z'_2,x) & \cdots & f_{Z|Z',X}(z_1|z'_K,x) \\ f_{Z|Z',X}(z_2|z'_1,x) & f_{Z|Z',X}(z_2|z'_2,x) & \cdots & f_{Z|Z',X}(z_2|z'_K,x) \\ \vdots & \vdots & \ddots & \vdots \\ f_{Z|Z',X}(z_K|z'_1,x) & f_{Z|Z',X}(z_K|z'_2,x) & \cdots & f_{Z|Z',X}(z_K|z'_K,x) \end{bmatrix}.$$

Proof. See Appendix. □

3.2 Identification of model parameters

Next, we identify the structural parameters $\theta \equiv (\alpha, \beta, \gamma)_{1\times d_\theta}$. Besides the observable variables (Y, X, Z, Z') specified in the previous step, we can treat m as known because it has been identified. Note that $f_{Y|Z^*,X}$ has also been identified in the first step, and it equals $F_\epsilon(\alpha + \beta m + \gamma X + Z^*)$ when $Y = 1$, and $1 - F_\epsilon(\alpha + \beta m + \gamma X + Z^*)$ when $Y = 0$. Thus, we expect that θ is identifiable if different values of θ leads to different values of the distribution function F_ϵ. Formally, we establish the following observationally equivalence concept.

Definition 3.1. *The set of parameters $\theta = (\alpha, \beta, \gamma)$ is observationally equivalent to the alternative set of parameters $\bar{\theta} = (\bar{\alpha}, \bar{\beta}, \bar{\gamma})$ if*

$$F_\epsilon(\alpha + \beta m + \gamma X + Z^*) = F_\epsilon\left(\bar{\alpha} + \bar{\beta} m + \bar{\gamma} X + Z^*\right)$$

for all elements of $\text{supp}(X)$ *and* $\text{supp}(Z^*)$.

Definition 3.1 indicates that the identification of θ holds if observational equivalence between model parameters and an alternative implies they are identical, i.e., $\theta = \bar{\theta}$. Let $\kappa = \alpha + \beta m + Z^*$ and $\bar{\kappa} = \bar{\alpha} + \bar{\beta} m + Z^*$. It is obvious that F_ϵ is monotone increasing in $\kappa + \gamma X$. Hence, the observational equivalence requirement holds if and only if

$$\kappa + \gamma X = \bar{\kappa} + \bar{\gamma} X,$$

or

$$(\kappa - \bar{\kappa}) + (\gamma - \bar{\gamma})X = 0. \tag{3.7}$$

Define the $d_x \times n$ matrix $\boldsymbol{X}_n = (X_i)_{1 \le i \le n}$ and the $n \times 1$ vector of ones $\boldsymbol{\iota}_n$. The matrix form of (3.7) is

$$(\kappa - \bar{\kappa})\boldsymbol{\iota}_n + (\gamma - \bar{\gamma})\boldsymbol{X} = \mathbf{0}. \tag{3.8}$$

To pin down the parameters α, β, and γ, we impose the following assumptions.

Assumption 3.5. *(i) ι_n and $\boldsymbol{X}_n$ are linearly independent; (ii) The values of z_1^* and z_K^* are known.*

If $\boldsymbol{\iota}_n$ and $\boldsymbol{X}_n$ are linearly independent, (3.8) implies that $\kappa = \bar{\kappa}$ and $\gamma = \bar{\gamma}$, i.e., κ and γ are identifiable. Hence, the identification sources for κ and γ stem from the intra-group variations in individual characteristics X. Sufficient condition for such linear independence may be (but is not limited to) that there exists a group g_0 such that the support of $(1, X_{i,g_0}^T)$ is not contained in a proper linear subspace of $\mathbb{R}^{d_x+1}$, where X_{i,g_0} represents the characteristics of agent i in group g_0. Part (ii) imposes that the lower and upper bounds of the unobserved heterogeneity Z^* are known. It can be understood as normalization of boundary of supp(Z^*). Such normalization restriction has been used in the literature for identifying structural models, e.g., Perrigne and Vuong (2011, 2012) and An et al. (2023) normalize the boundary values of unobserved effort.

To separately recover α and β from κ, we must rely on inter-group variations of m from agents in different groups. Specifically, for groups that have smallest and largest unobserved group heterogeneity, i.e., z_1^* and z_K^*, we have the following system of equations:

$$\alpha \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \beta \begin{bmatrix} m(z_1^*) \\ m(z_K^*) \end{bmatrix} + \begin{bmatrix} z_1^* \\ z_K^* \end{bmatrix} = \begin{bmatrix} \kappa(z_1^*) \\ \kappa(z_K^*) \end{bmatrix}, \tag{3.9}$$

where $m(z_1^*)$ and $m(z_K^*)$ are known because we have identified m in the first step and it is strictly monotonic in Z^* by (3.2) and Lemma 3.3. Note that (3.9) represents a linear system with two equations but four unknowns (α, β, z_1^*, z_K^*). Then, it is clear that we need to impose additional restrictions for joint identification of α and β and that is why Part (ii) of Assumption 3.5 is imposed. Under this assumption, (3.9) admits a unique solution

$$\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} 1 & m(z_1^*) \\ 1 & m(z_K^*) \end{bmatrix}^{-1} \begin{bmatrix} \kappa(z_1^*) - z_1^* \\ \kappa(z_K^*) - z_K^* \end{bmatrix},$$

where the matrix inverse exists because of the strict monotonicity of m.

We summarize the identification results in the following proposition, whose proof is already presented in the text above.

Proposition 3.2. *Under Assumptions 2.1 and 3.1-3.5, the structural parameters θ are identified.*

4 Semiparametric estimation

In this section, we discuss the estimation method of the baseline model and its asymptotic properties. Let $\theta_0 = (\alpha_0, \beta_0, \gamma_0)$ denote the true value of the structural parameters. We can estimate θ_0 via the semiparametric nonlinear least squares (SNLS) method. Specifically, define

$$m^c = \begin{bmatrix} m(z_1^*) \\ m(z_2^*) \\ \vdots \\ m(z_K^*) \end{bmatrix}, \quad f^c_{Z^*|Z}(z^*|\cdot) = \begin{bmatrix} f(z_1^*|\cdot) \\ f(z_2^*|\cdot) \\ \vdots \\ f(z_K^*|\cdot) \end{bmatrix},$$

where $(z_1^*, z_2^*, ..., z_K^*)$ represents different values of $Z^* \in \text{supp}(Z^*)$. In the sample, we observe the variables $W_i \equiv (Y_i, X_i^T, Z_i, Z_i')^T,\ i = 1, 2, ..., n$, which lead to an observed conditional moment

$$\begin{aligned} \mathbb{E}(Y|Z' = z', X = x) &= \sum_{z^* \in \text{supp}(Z^*)} \mathbb{E}(Y|Z^* = z^*, Z' = z', X = x) f_{Z^*|Z',X}(z^*|z', x) \\ &= \sum_{z^* \in \text{supp}(Z^*)} \mathbb{E}(Y|Z^* = z^*, X = x) f_{Z^*|Z',X}(z^*|z', x) \\ &= \sum_{z^* \in \text{supp}(Z^*)} F_\epsilon\left(\alpha + \beta m + \gamma x + z^*\right) f_{Z^*|Z',X}(z^*|z', x) \\ &\equiv g(z', x; \theta_0, \sigma_0), \end{aligned} \tag{4.1}$$

where the second equality is by Assumption 3.2, and the third equality is by (3.1). The observed moment function above depends on the nuisance functions $\sigma_0 \equiv [(m^c)^T, (f^c_{Z^*|Z',X})^T]^T$.

In the following, we discuss the estimation procedure for the nuisance functions σ_0. Then, we can estimate the joint distributions of Z, Z', Y and X using a simple frequency estimator,

$$\widehat{f}_{Z,Z',Y,X}(z, z', y, x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(Z_i = z, Z_i' = z', Y_i = y, X_i = x),$$

where $\mathbb{1}(\cdot)$ is the indicator function. Similarly, we can estimate $f_{Z,Z',X}$ and f_X using the frequency estimator. Then, the conditional distribution matrices $M_{Z,Z',Y|X}$ and $M_{Z,Z'|X}$ can be estimated by stacking the estimate of $f_{Z,Z',Y,X}$, $f_{Z,Z',X}$ and f_X as follows:

$$\widehat{M}_{Z,Z',Y|X} = \left[\frac{\widehat{f}_{Z,Z',Y,X}(z_l, z_k', y, x)}{\widehat{f}_X(x)} \right]_{l,k}$$

and

$$\widehat{M}_{Z,Z'|X} = \left[\frac{\widehat{f}_{Z,Z',X}(z_l, z_k', x)}{\widehat{f}_X(x)} \right]_{l,k}.$$

Next, following the identification procedure, there exists known functions ψ and ϕ such that the conditional densities $f_{Y|Z^*,X}$ and $f_{Z|Z^*,X}$ are estimated as

$$\widehat{f}_{Y|Z^*,X} = \psi\left(\widehat{M}_{Z,Z',Y|X}\widehat{M}^{-1}_{Z,Z'|X}\right)$$

and

$$\widehat{f}_{Z|Z^*,X} = \phi\left(\widehat{M}_{Z,Z',Y|X}\widehat{M}^{-1}_{Z,Z'|X}\right),$$

respectively. Specifically, ψ and ϕ compute the eigenvalues and eigenvectors of the matrix. Although the expressions of ψ and ϕ are complicated, Andrew et al. (1993) shows that they are well-behaved analytic functions. Then, the conditional expectation m is estimated as

$$\widehat{m} = \sum_{x\in\text{supp}(X)} \widehat{f}_{Y|Z^*,X}(1|z^*,x)\widehat{f}_X(x)$$

by (3.6). Finally, the conditional density of Z^* can be estimated as the elements of

$$\widehat{M}_{Z^*|Z',X} = \widehat{M}^{-1}_{Z|Z^*,X}\widehat{M}_{Z|Z',X} \tag{4.2}$$

by Lemma 3.4, where $\widehat{M}_{Z^*|Z',X} = [\widehat{f}_{Z^*|Z',X}(z^*_l|z'_k,x)]_{l,k}$. Note that $\widehat{M}_{Z|Z',X} = [\widehat{f}_{Z|Z',X}(z_l|z'_k,x)]_{l,k}$ is directly obtained from data using the frequency estimator.

Let $\widehat{\sigma}$ denote the estimate of σ_0 with the nuisance functions m^c and $f^c_{Z^*|Z}$ replacing by their estimates. By (4.1), the SNLS estimator $\widehat{\theta}$ is defined as follows:

$$\widehat{\theta} = \underset{\theta\in\Theta}{\text{argmin}} \sum_{i=1}^{n}\left[Y_i - g(Z'_i, X_i;\theta,\widehat{\sigma})\right]^2, \tag{4.3}$$

where Θ is the parameter space.

4.1 Consistency

[We choose union of multiple unobserved groups such that Z and Z' can be fully covered by the union. Assume that the size of the union is $\tilde{n}_g$ and the number of such union is $\tilde{G}$ such that $\sum \tilde{n}_g\tilde{G} = n$, $\tilde{G} \leq G$, $\tilde{n}_g \geq n_g$.]

In this section, we establish the consistency of our two-step semiparametric estimator. First, we show that the estimator of the nuisance functions σ is uniformly consistent. Note that we can not assume the dependent variable $\{Y_i\}_{i=1,2,...,n}$ is i.i.d. across individuals because for i and j within the same group, their outcomes Y_i and Y_j are not independent because of the same Z^*_g. Therefore, we impose the following assumption:

Assumption 4.1. *(i) The group size n_g is fixed for $g = 1, 2, ..., G$; (ii) The observables W_i are* *independent across different groups and identically distributed; (iii) There exists a constant $c > 0$ such that $f_X \geq c$ and $f_{Z',X} \geq c$.*

Assumption 4.1 (i) implies that the number of groups G goes to infinity as $n \to \infty$. Therefore, we will focus on the asymptotics with $G \to \infty$ and n_g fixed. This data structure is analogous to the panel data with $n \to \infty$ and time period t fixed. Hence, condition (ii) ensures that the law of large numbers and central limit theorem still work for our data. Condition (iii) is a technical condition that guarantees the densities are bounded away from zero. Let $\|\cdot\|_\infty$ denote the sup norm and σ_0 the true value of σ. we have

Proposition 4.1. *Suppose the assumptions in Proposition 3.1 and Assumption 4.1 hold. Then, $\|\widehat{\sigma} - \sigma_0\|_\infty = O_p(n^{-1/2})$.*

Proof. See Appendix. □

Proposition 4.1 means that the estimator of the nuisance functions σ is uniformly consistent at the rate $n^{-1/2}$. The uniform convergence rate is consistent with that of the conventional frequency estimation under i.i.d. data setting.

Next, we show that the SNLS estimator $\widehat{\theta}$ is a consistent estimator for θ_0. We impose the following assumptions:

Assumption 4.2. *(i) Θ is compact; (ii) For all $(\theta, \sigma) \in \Theta \times \Sigma$, $g(z', x, \theta, \sigma)$ is measurable of z' and x and is continuously differentiable in θ up to order 3 for all z' and x; (iii) There exists a function $h(w)$ with $\mathbb{E}[h(w)] < \infty$ such that $g(z', x; \theta, \sigma)^2 \leq h(w)$ and $\|\nabla_{\theta^T} g(z', x; \theta, \sigma)\|^2 \leq h(w)$ for all $w \in \text{supp}(W)$.*

Assumptions 4.2 (i)-(ii) are standard in the M-estimation literature. See, e.g., Newey and McFadden (1994) and Wooldridge (1994). Condition (iii) is a technical condition for the law of large numbers. The consistency of the estimator $\widehat{\theta}$ is summarized in the following theorem:

Theorem 4.1. *Suppose the assumptions in Proposition 3.2 and Assumptions 4.1-4.2 are satisfied. Then,*

$$\widehat{\theta} \xrightarrow{p} \theta_0.$$

Proof. See Appendix. □

4.2 Asymptotic normality

We now show the asymptotic distribution of the estimator $\widehat{\theta}$. Note that we need to account for the presence of the nuisance functions σ. From the first order condition of the optimization problem (4.3), $\widehat{\theta}$ solves

$$\frac{1}{n}\sum_{i=1}^{n}[Y_i - g(Z_i', X_i; \widehat{\theta}, \widehat{\sigma})]\nabla_{\theta^T} g(Z_i', X_i; \widehat{\theta}, \widehat{\sigma}) = 0.$$

Define $s(W_i;\theta,\sigma) = [Y_i - g(Z_i', X_i;\theta,\sigma)]\nabla_{\theta^T} g(Z_i', X_i;\theta,\sigma)$. Then, by the mean value theorem we can obtain

$$\frac{1}{n}\sum_{i=1}^{n} s(W_i;\theta_0,\widehat{\sigma}) + \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta s(W_i;\widetilde{\theta},\widehat{\sigma})(\widehat{\theta}-\theta_0) = 0, \tag{4.4}$$

where $\widetilde{\theta}$ is between $\widehat{\theta}$ and θ_0. If $1/n\sum_{i=1}^{n}\nabla_\theta s(W_i;\widetilde{\theta},\widehat{\sigma})$ is invertible, rearranging (4.4) leads to

$$\sqrt{n}(\widehat{\theta}-\theta_0) = \left[\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta s(W_i;\widetilde{\theta},\widehat{\sigma})\right]^{-1}\left[-\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s(W_i;\theta_0,\widehat{\sigma})\right].$$

Define $v_0 = (f_{Z,Z',Y,X}, f_{Z,Z',X}, f_{Z',X}, f_X)^T$ and $\widehat{v}$ contains the frequency estimators of all densities in v_0. Note that by the identification result, we can express the nuisance functions σ_0 and their estimates $\widehat{\sigma}$ as known functions of v_0 and $\widehat{v}$, respectively. Besides, we also impose the following notations:

$$\rho(w) \equiv \widetilde{l}(w) - \mathbb{E}[\widetilde{l}(w)],$$

$$\widetilde{l}(w) \equiv \mathbb{E}\left[\nabla_{v^T}\{[y-g(z',x;\theta_0,\sigma_0)]\nabla_{\theta^T}g(z',x;\theta_0,\sigma_0)\}\boldsymbol{\iota}_{d_v}\middle| W=w\right],$$

$$H \equiv \mathbb{E}[\nabla_\theta s(W_i;\theta_0,\sigma_0)]$$

and

$$D \equiv \lim_{n\to\infty}\frac{1}{n}\operatorname{Var}\left\{\sum_{i=1}^{n}[s(W_i;\theta_0,\sigma_0)+\rho(W_i)]\right\},$$

where $\boldsymbol{\iota}_{d_v}$ is a $d_v \times 1$ vector of ones and d_v is the dimension of v.

To obtain the asymptotic distribution of $\widehat{\theta}$, the key step is to show that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s(W_i;\theta_0,\widehat{\sigma}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[s(W_i;\theta_0,\sigma_0)+\rho(W_i)] + o_p(1),$$

where $\rho(W_i)$ is the correction term that accounts for the nonparametric estimation of the nuisance functions. The expression of $\rho(W_i)$ is obtained from the linearization of $g(z,x;\theta_0,\sigma)$ with respect to v. Then, we impose the following assumptions:

Assumption 4.3. *(i) $\theta_0 \in \operatorname{int}(\Theta)$; (ii) $g(z',x;\theta,\sigma)$ and $\nabla_{\theta^T}g(z',x;\theta,\sigma)$ are continuously differentiable in v up to order 2 with uniformly bounded derivatives; (iii) $\mathbb{E}[\|\rho(W_i)\|^2] < \infty$; (iv) There exists a function $h(w)$ with $\mathbb{E}[h(w)] < \infty$ such that $\|\nabla_\theta s(w;\theta,\sigma)\|^2 \le h(w)$ for all $w \in \operatorname{supp}(W)$; (v) H exists and is nonsingular; (vi) f_ϵ and $\nabla_{\theta^T}f_\epsilon$ are uniformly continuous in m.*

Assumption 4.3 (i)-(v) are standard in the semiparametric M-estimation literature, see, e.g., Andrews (1994), Newey (1994a) and Newey and McFadden (1994). Condition (vi) guarantees the uniform negligibility of the remainder terms in the score function and the Hessian matrix when we approximating the nuisance functions by their estimates (Kasy 2019).

Theorem 4.2. *Suppose the assumptions in Theorem 4.1 and Assumption 4.3 hold. Then,*

$$\sqrt{n}(\widehat{\theta}-\theta_0) \xrightarrow{d} N(0, H^{-1}DH^{-1}).$$

Proof. See Appendix. □

Since the knowledge of group memberships is not available, we can not directly estimate the asymptotic variance $H^{-1}DH^{-1}$ by the analogy principle because D includes unknown covariance terms across individuals within each group. Nevertheless, the $\sqrt{n}$-consistency of the semiparametric NLS estimator $\widehat{\theta}$ enables us to apply the resampling method in Leung (2022) for inference of the parameters of interest. We discuss this inference procedure in the next section.

4.3 Inference by subsampling

We consider testing $H_0 : \theta_0 = \theta$ for some $\theta \in \Theta$. Let $R_n \geq 2$ be an integer and Π the set of permutations on $\{1, 2, ..., n\}$. Let $\{\pi_r\}_{r=1}^{R_n}$ be a set of R_n i.i.d. uniform draws from Π and $\pi \equiv (\pi_1, \pi_2, ..., \pi_{R_n})$.

The results in the previous section implies that $\widehat{\theta}$ is asymptotically linear in the sense that

$$\sqrt{n}(\widehat{\theta}-\theta_0) = \left[\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta s(W_i;\widehat{\theta},\widehat{\sigma})\right]^{-1}\left\{-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[s(W_i;\theta_0,\sigma_0)+\rho(W_i)\right]\right\}+o_p(1).$$

Then, following Leung (2022), we define the test statistic as

$$T_M(\theta;\pi) = \frac{1}{\sqrt{R_n}}\sum_{r=1}^{R_n}\widehat{V}^{-1/2}\chi_{\pi_r(1)}, \tag{4.5}$$

where

$$\chi_i = -\left[\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta s(W_i;\widehat{\theta},\widehat{\sigma})\right]^{-1}\left[s(W_i;\theta,\widehat{\sigma})+\widehat{\rho}(W_i)\right],$$

$$\widehat{V} = \frac{1}{n}\sum_{i=1}^{n}(\chi_i-\overline{\chi})(\chi_i-\overline{\chi})^T,$$

$\overline{\chi}$ is the sample average of $\{\chi_i\}_{i=1}^{n}$. Following Newey (1994b, equation 7), $\widehat{\rho}(W_i)$ is obtained by computing via numerical differentiation the (demeaned) first-order effect of the ith observation in each component of $\widehat{v}$ on $1/n\sum_{i=1}^{n}s(W_i;\widehat{\theta},\widehat{\sigma})$.

The mean-type statistic in (4.5) is computed by drawing R_n observations with replacement from $\{\widehat{V}^{-1/2}\chi_i\}_{i=1}^{n}$, and then taking the average and scaling up by $\sqrt{R_n}$. Note that we compute $\widehat{V}$ using the full sample. Since $\widehat{\sigma}$ is $\sqrt{n}$-consistent, Theorem A.1 of Leung (2022) implies the following result.

Proposition 4.2. *Suppose the following conditions hold. (i)* $n^{-1}\sum_{i=1}^{n}\|\widetilde{\chi}_i\|^{2+\lambda}=O_p(1)$ *for some* $\lambda>0$*, where*

$$\widetilde{\chi}_i=-H^{-1}\left[s(W_i;\theta_0,\sigma_0)+\rho(W_i)\right].$$

(ii) $V\equiv\operatorname{Var}(\widetilde{\chi}_i)$ *is positive definite. (iii)* $R_n\to\infty$ *and* $R_n/n=o(1)$*. Then, under the null hypothesis* $H_0:\theta_0=\theta$*,* $T_M(\theta;\pi)\xrightarrow{d}N(0,I_{d_\theta})$ *conditional on* $\boldsymbol{W}\equiv(W_i)_{1\le i\le n}$*, where* I_{d_θ} *is a* $d_\theta\times d_\theta$ *identity matrix.*

Proof. See Appendix. □

Proposition 4.2 enables us to use standard normal critical values for testing $H_0:\theta_0=\theta$. The intuition behind this proposition can be explained as follows: the test statistic $T_M(\theta;\pi)$ can be decomposed into two parts

$$T_M(\theta_0;\pi)=\underbrace{T_M(\theta_0;\pi)-\mathbb{E}\left[T_M(\theta_0;\pi)|\boldsymbol{W}\right]}_{\text{Part I}}+\underbrace{\mathbb{E}\left[T_M(\theta_0;\pi)|\boldsymbol{W}\right]}_{\text{Part II}}.$$

While the first part converges in distribution to a standard normal random variable as $R_n\to\infty$ because random permutations are i.i.d. conditional on the data, the second part is a bias term that is asymptotically negligible if we make R_n diverge at a sufficiently slow rate. Following Leung (2022), the tuning parameter R_n is chosen by trading-off the power of the test and the bias, which yields the optimal R_n as

$$R_n^*=\sqrt{n}.$$

5 Extensions

The identification method described in Section 3 can be extended in several ways. We shall formally discuss two cases: nonrandom assignment and unequal support.

5.1 Nonrandom assignment

Consider the case where individuals are nonrandomly assigned, which means that Assumption 3.1 no longer holds, i.e., X and Z^* are dependent with each other.

For identification, we maintain Assumptions 3.2–3.5. First, the conditional densities $f_{Y|Z^*,X}$ and $f_{Z|Z^*,X}$ can still be identified using the eigen-decomposition technique in Section 3. Specifically, the matrix equations (3.3) and (3.4) will be modified to

$$M_{Z,Z',Y|X}=M_{Z|Z^*,X}D_{Y|Z^*,X}D_{Z^*|X}M_{Z'|Z^*,X}^{T}\tag{5.1}$$

and

$$M_{Z,Z'|X}=M_{Z|Z^*,X}D_{Z^*|X}M_{Z'|Z^*,X}^{T}\tag{5.2}$$

without Assumption 3.1, where

$$D_{Z^*|X} = \begin{bmatrix} f_{Z^*|X}(z_1^*|x) & & & \\ & f_{Z^*|X}(z_2^*|x) & & \\ & & \ddots & \\ & & & f_{Z^*|X}(z_K^*|x) \end{bmatrix}.$$

for each $X = x$. By arguments similar to the proof of Proposition 3.1, equations (5.1) and (5.2) still lead to the eigen-decomposition equation (3.5).

Second, the conditional density $f_{Z^*|Z',X}$ can be identified analogously to Lemma 3.4. Then, by the law of total probability we have

$$f_{Z^*,X}(z^*, x) = \sum_{z' \in \text{supp}(Z')} f_{Z^*|Z',X}(z^*|z', x) f_{Z',X}(z', x) \tag{5.3}$$

and

$$f_{Z^*}(z^*) = \sum_{x \in \text{supp}(X)} f_{Z^*,X}(z^*, x) \tag{5.4}$$

for each $X = x$. The matrix version of equation (5.3) can be written as

$$M_{Z^*,X} = M_{Z^*|Z',X} M_{Z',X}, \tag{5.5}$$

where

$$M_{Z^*,X} = \begin{bmatrix} f_{Z^*,X}(z_1^*, x) \\ f_{Z^*,X}(z_2^*, x) \\ \vdots \\ f_{Z^*,X}(z_K^*, x) \end{bmatrix}, \quad M_{Z',X} = \begin{bmatrix} f_{Z',X}(z_1', x) \\ f_{Z',X}(z_2', x) \\ \vdots \\ f_{Z',X}(z_K', x) \end{bmatrix},$$

$$M_{Z^*|Z',X} = \begin{bmatrix} f_{Z^*|Z',X}(z_1^*|z_1', x) & f_{Z^*|Z',X}(z_1^*|z_2', x) & \cdots & f_{Z^*|Z',X}(z_1^*|z_K', x) \\ f_{Z^*|Z',X}(z_2^*|z_1', x) & f_{Z^*|Z',X}(z_2^*|z_2', x) & \cdots & f_{Z^*|Z',X}(z_2^*|z_K', x) \\ \vdots & \vdots & \ddots & \vdots \\ f_{Z^*|Z',X}(z_K^*|z_1', x) & f_{Z^*|Z',X}(z_K^*|z_2', x) & \cdots & f_{Z^*|Z',X}(z_K^*|z_K', x) \end{bmatrix}.$$

Equation (5.5) implies that the joint density $f_{Z^*,X}$ can be identified as elements of $M_{Z^*|Z',X} M_{Z',X}$, and the density f_{Z^*} is identified accordingly by (5.4). Consequently, the conditional density $f(X|Z^*)$ can be identified as elements of the matrix

$$D_{Z^*}^{-1} M_{Z^*,X}$$

by Bayes' theorem. Hence, the conditional expectation m^* is nonparametrically identified by the first equality of (3.6). Finally, we can identify the structural parameters θ by Assumption 3.5. The estimation of θ can be conducted similarly as described in Section 4.

5.2 Unequal support

In this section, we consider the identification of model primitives while relaxing the equal support Assumption 3.3. Specifically, we will consider two cases: first, only one proxy (Z') has larger support. Second, both proxies have larger supports.

5.2.1 One proxy has larger support

Assumption 3.3′. *$|\operatorname{supp}(Z)| = |\operatorname{supp}(Z^*)| = K$, $|\operatorname{supp}(Z')| = K' > K$.*

Assumption 3.3′ allows the cardinality of $\operatorname{supp}(Z')$ to be larger than K. Under this assumption, the matrices $M_{Z,Z'|X}$, $M_{Z'|Z^*,X}$ are rectangular and hence we cannot directly invert them to obtain the eigen-decomposition equation (3.5).

To address this technical challenge, we utilize the concept of the generalized inverse of a matrix. Let $M^+_{Z,Z'|X}$ and $M^+_{Z'|Z^*,X}$ be the Moore–Penrose inverse of $M_{Z,Z'|X}$ and $M_{Z'|Z^*,X}$[9]. Then, by (3.4) and Assumption 3.4 we have

$$M^+_{Z,Z'|X} = \left(M^T_{Z'|Z^*,X}\right)^+ \left(M_{Z|Z^*,X} D_{Z^*}\right)^+ = M_{Z'|Z^*,X}\left(M^T_{Z'|Z^*,X} M_{Z'|Z^*,X}\right)^{-1} D^{-1}_{Z^*} M^{-1}_{Z|Z^*,X} \quad (5.6)$$

where the first equality is by the product property of the Moore–Penrose inverse. Post-multiplying (3.3) by (5.6) leads to

$$M_{Z,Z',Y|X} M^+_{Z,Z'|X} = M_{Z|Z^*,X} D_{Y|Z^*,X} M^{-1}_{Z|Z^*,X},$$

which implies that the matrix $M_{Z,Z',Y|X} M^+_{Z,Z'|X}$ still has an eigenvalue-eigenvector decomposition. Therefore, we can follow the rest of the procedure in Section 3 to identify model primitives m and θ.

5.2.2 Both proxies have larger supports

Assumption 3.3″. *$|\operatorname{supp}(Z^*)| = K$, $|\operatorname{supp}(Z)| = |\operatorname{supp}(Z')| = K' > K$.*

Under Assumption 3.3″, both proxies Z and Z' can have larger supports than Z^*. Consequently, the matrix $M_{Z|Z^*,X}$ is also rectangular, and thus we cannot even employ the eigen-decomposition

[9] For a real matrix A, its Moore–Penrose inverse is defined as the unique matrix A^+ that satisfies the following conditions: $AA^+A = A$, $A^+AA^+ = A^+$, $(AA^+)^T = AA^+$, $(A^+A)^T = A^+A$. In general, A^+ can be obtained by a singular value decomposition. However, if A is full column rank, $A^+ = (A^TA)^{-1}A^T$. On the other hand, $A^+ = A^T(AA^T)^{-1}$ if A is full row rank. Moreover, $A^+ = A^{-1}$ if A is nonsingular (Golub and Van Loan 2013).

method for identification, as the eigenvector matrix should be square. In this scenario, we can generate a new proxy variable $\widetilde{Z}$ with $|\operatorname{supp}(\widetilde{Z})| = K$ by combining some values in the support of Z. Formally, this process is conducted through a subjective function $q : \operatorname{supp}(Z) \mapsto \operatorname{supp}(\widetilde{Z})$ and $\widetilde{Z} = q(Z)$. This is equivalent to grouping rows of $M_{Z|Z^*,X}$ and $M_{Z,Z'|X}$ so that the new matrices $M_{\widetilde{Z}|Z^*,X}$ and $M_{\widetilde{Z},Z'|X}$ have K rows. The following lemma ensures these new matrices are full (row) rank.

Lemma 5.1. *Under Assumptions 3.3″ and 3.4, the row vectors of $M_{Z|Z^*,X}$, and $M_{Z,Z'|X}$ can be grouped such that the new matrices $M_{\widetilde{Z}|Z^*,X}$ and $M_{\widetilde{Z},Z'|X}$ have rank K.*

Proof. The proof is similar to Xiao (2018, Lemma 2) or Luo et al. (2022, Lemma 3) and hence is omitted. □

We can then follow the idea in the previous section to obtain the eigen-decomposition. Specifically, equations (3.3) and (3.4) can be modified as

$$M_{\widetilde{Z},Z',Y|X} = M_{\widetilde{Z}|Z^*,X} D_{Y|Z^*,X} D_{Z^*} M^T_{Z'|Z^*,X} \tag{5.7}$$

and

$$M_{\widetilde{Z},Z'|X} = M_{\widetilde{Z}|Z^*,X} D_{Z^*} M^T_{Z'|Z^*,X}. \tag{5.8}$$

The Moore-Penrose inverse of (5.8) is

$$M^+_{\widetilde{Z},Z'|X} = M_{Z'|Z^*,X} \left(M^T_{Z'|Z^*,X} M_{Z'|Z^*,X} \right)^{-1} D^{-1}_{Z^*} M^{-1}_{\widetilde{Z}|Z^*,X}. \tag{5.9}$$

Post-multiplying (5.7) by (5.9) yields again the eigen-decomposition equation

$$M_{\widetilde{Z},Z',Y|X} M^+_{\widetilde{Z},Z'|X} = M_{\widetilde{Z}|Z^*,X} D_{Y|Z^*,X} M^{-1}_{\widetilde{Z}|Z^*,X}.$$

The identification of model primitives m and θ follows accordingly by imposing Assumption 3.5.

6 Simulation studies

In this section, we use Monte Carlo experiments to demonstrates the finite sample performance of the estimator in Section 4. In each iteration of the simulations, we generate data using sample size $n = 1000$. Then, the whole sample is divided into equally sized groups with each having 4 individuals. Therefore, the number of groups is 250. Consistent with the model setup in Section 2, we consider the following data generating process of the outcome variable:

$$Y = \mathbb{1}(\alpha + \beta m + \gamma X + Z^* \geq \epsilon), \tag{6.1}$$

where the covariate $X \sim Bernoulli(0.5)$ is i.i.d. across individuals. The true values of the parameters are $\alpha = -1$, $\beta = 0.5$ and $\gamma = -1$. We consider five distributions for the error ϵ as follows:

1. **UN**: $\epsilon \sim \text{Unif}(-\sqrt{3}, \sqrt{3})$
2. **NR**: $\epsilon \sim N(0, 1)$
3. **T3**: $\epsilon \sim t_3$ (Student's t distribution with 3 degrees of freedom)
4. **LG**: $\epsilon \sim \text{Logistic}(0, 1)$ (standard logistic distribution)

The discrete latent variable Z^* has the support $\{1, 2, 3, 4\}$ with probability mass functions $P_{Z^*} \equiv (Pr(Z^* = 1), Pr(Z^* = 2), Pr(Z^* = 3), Pr(Z^* = 4)) = (0.2, 0.3, 0.2, 0.3)$. The variable Z^* is generated as follows:

$$Z^* = \begin{cases} 1 & \text{if } e_{Z^*} \leq Pr(Z^* = 1), \\ 2 & \text{if } Pr(Z^* = 1) < e_{Z^*} \leq Pr(Z^* \leq 2), \\ 3 & \text{if } Pr(Z^* \leq 2) < e_{Z^*} \leq Pr(Z^* \leq 3), \\ 4 & \text{if } Pr(Z^* \leq 3) < e_{Z^*} \leq Pr(Z^* \leq 4), \end{cases}$$

where e_{Z^*} is uniformly distributed on $[0, 1]$ and is independent of all other variables. In the experiments, we also generate two indicators Z and Z' for Z^*. Z and Z' share the same support $\{1, 2, 3, 4\}$ and probability mass functions as Z^*, i.e., $P_Z = P_{Z'} = (0.2, 0.3, 0.2, 0.3)$. The indicator Z (Z') is generated similarly as Z^*, with e_{Z^*} replaced by $0.5e_{Z^*} + 0.5e_Z$ ($0.5e_{Z^*} + 0.5e_{Z'}$), where e_Z ($e_{Z'}$) is another independent random variable with a uniform distribution on $[0, 1]$. Hence, the correlation between Z^* and Z (Z') is caused by the common random variable e_Z ($e_{Z'}$). The variables Z^*, Z and Z' are i.i.d. across different groups. The equilibrium conditional expectation m is generated by solving the fixed point of (3.2).

In the estimation, we compare two NLS estimators: (i) SNLS: the proposed semiparametric NLS estimator in (4.3); (ii) Naïve: the estimator that neglects unkown group structures, i.e., treating the indicator Z as Z^*, estimating m via frequency estimator

$$\widehat{m}(z) = \frac{\sum_{i=1}^{n} Y_i \cdot \mathbb{1}(Z_i = z)}{\sum_{i=1}^{n} \mathbb{1}(Z_i = z)}$$

and then obtaining the estimator $\widehat{\theta}_n$ as

$$\widehat{\theta}_n = \underset{\theta \in \Theta}{\text{argmin}} \sum_{i=1}^{n} \left[Y_i - F_\epsilon \left(\alpha + \beta\widehat{m} + \gamma X + Z\right)\right]^2.$$

The simulation results are provided in Table 1, where the SNLS and naïve estimates are presented in panels A and B, respectively. For each estimator, we report the bias, the standard deviation

(Std.dev) and the root mean squared error (RMSE) over 1000 replications. The results show that our estimator performs well in finite samples. Overall, the SNLS estimators of all four parameters has much smaller biases than the naïve estimators that ignore the unknown group structure. For example, consider the estimation of peer effects β. When the error distribution is standard normal, the bias of SNLS estimator is 0.1728, which is about 8.26% of the bias (2.0913) of the naïve estimator. When the error distribution changes to standard logistic, the bias of the SNLS estimator (0.0272) changes to 1.63% of the bias of the naïve estimator (1.6673). Furthermore, the SNLS estimator achieves considerably reductions in the RMSEs for the estimated social parameters, relative to the naïve estimator.

Table 1: Simulation Results

	Panel A: SNLS								
	$\alpha = -1$			$\beta = 0.5$			$\gamma = -1$		
Model	Bias	Std.dev	RMSE	Bias	Std.dev	RMSE	Bias	Std.dev	RMSE
UN	0.0181	0.5491	0.5491	0.1132	0.7681	0.7761	-0.0570	0.2263	0.2333
NR	-0.0389	0.5538	0.5549	0.1728	0.7988	0.8168	-0.0261	0.2289	0.2302
T3	-0.0281	0.5488	0.5492	0.1100	0.7924	0.7996	-0.0326	0.2282	0.2304
LG	0.0292	0.5885	0.5890	0.0272	0.7984	0.7985	-0.0362	0.2039	0.2070
	Panel B: Naïve								
	$\alpha = -1$			$\beta = 0.5$			$\gamma = -1$		
UN	-1.1831	0.7013	1.3752	1.9313	0.9795	2.1653	-0.1901	0.1555	0.2455
NR	-1.1922	0.6250	1.3459	2.0913	0.8407	2.2538	-0.2323	0.1195	0.2612
T3	-1.1743	0.6918	1.3627	2.0457	0.9466	2.2539	-0.1873	0.1472	0.2381
LG	-0.9936	0.7018	1.2162	1.6673	0.9791	1.9333	-0.1215	0.1736	0.2118

To examine the finite sample performance of the subsampling method in Secition 4.3, we also run simulations to show the size control of the test for $H_0 : \theta_0 = (-1, 0.5, -1)$. The tuning parameter R_n is chosen to be $\sqrt{1000} \approx 32$. The results of the experiments shown in Table 2 indicates that the empirical sizes obtained from using the subsampling method are well controlled under various nominal levels.

Table 2: Simulated Test Size

	Null Hypothesis								
	$\alpha = -1$			$\beta = 0.5$			$\gamma = -1$		
Model	0.1	0.05	0.01	0.1	0.05	0.01	0.1	0.05	0.01
UN	0.099	0.050	0.016	0.114	0.063	0.018	0.091	0.057	0.017
NR	0.085	0.054	0.014	0.119	0.053	0.015	0.109	0.055	0.014
T3	0.095	0.046	0.012	0.084	0.041	0.013	0.104	0.050	0.009
LG	0.103	0.053	0.018	0.106	0.057	0.015	0.082	0.044	0.008

7 Empirical application: peer effects in private tutoring

In this section, we apply our method to investigate competitive peer effects, i.e., the peer effects in competition group, on the decisions of middle school students to participate in private tutoring in Chinese middle school.

7.1 Background and data

In China, compulsory education comprises 6 years of elementary school (Grades 1-6), typically beginning at the age of 6 and finishing at the age of 12, followed by 3 years of middle school (Grades 7-9), typically aged 13-15. The 9th graders need to take the High School Entrance Examination (a.k.a., Zhongkao), which is usually held in July every year, to compete for general high school admissions, especially for selective general high schools.[10] Unlike the admission quotas of colleges, which are distributed at the provincial level, high schools primarily recruit students at the county level. More specifically, most high schools allocate a fixed quota to each local middle school in their respective counties, i.e., students primarily compete at a local level.

It is very common for students to participate in private tutoring in order to obtain a higher scores in Zhongkao. A student's decision may be influenced by peers in the same class, particularly those with similar academic performance. This is because in middle schools, all the students in a class share the same set of teachers and the same school schedule, and students mainly interact with their classmates. Unfortunately, examining competitive peer effects is challenging without knowledge of the specific competition group to which a student belongs. With only each student's class information available, the typical method of estimating the potential peer effects in schools is to assume a linear-in-mean specification. Each class is considered as a group in which each student in the class is linked to all others in the same class. The peers' expected participation decision in private tutoring is simply measured by the average participation decisions of other students in the class.[11]. However, such an approach may not be appropriate because students in a classroom have different levels of competitiveness and in practice students only compete with those whose academic performance is similar to them.

The dataset we use to estimate the competitive peer effects is the China Education Panel Survey (CEPS). CEPS is the first large-scale, nationally representative, longitudinal survey for middle school students and contains rich information from students, parents, teachers, and school administrator questionnaires. In its baseline survey completed in the 2013-2014 academic year,

[10] There are two types of high schools in China: general high schools and vocational high schools. General high schools are academically oriented and are the primary pathway for students aiming to enter university, while vocational high schools focus on providing students with specific technical skills and vocational training for direct entry into the job market. Currently, the majority of middle school students prefer and aim for the general high school and university path.

[11] For example, Duflo and Saez (2002) use this approach to estimate how colleagues' participation in a retirement plan influences an individual's decision.

CEPS interviewed 19,487 students, including 10,279 7th graders and 9,208 9th graders, in 438 classrooms from 112 middle schools across 28 counties in mainland China. 7th graders were followed up and were in 8th grade in the 2014-2015 academic year, while 9th graders in the baseline survey were not followed up since they had already graduated. In this paper, we use the data from the baseline survey.

A notable feature of the CEPS data is that it asks school administrators how schools assign their students to classes upon entry into middle schools. Among 112 schools sampled, 85.3 percent report random assignment. In addition, both homeroom teachers and main subject teachers are randomly assigned to each class after students' random assignment.

Table 3 presents summary statistics for all the variables in our data. Grade9, Male, OnlyChild, Family income, and Parents' education are covariates X. Grade9 equals 1 when a student is in 9th grade and zero otherwise; male is equal to 1 for male students and 0 for female students; OnlyChild is set to 1 when the student is the only child in his/her family and 0 otherwise; family income takes the values low, medium, and high, which are numerically coded as $1, 2$, and 3, respectively; and parents' education takes a value of one when any of the parents of a student has a high school or higher educational level, and takes a value of zero otherwise. The measurement Z is the rank of a classroom relative to other classes perceived by a homeroom teacher when he/she starts his/her role as the homeroom teacher of the class. It takes value $1, 2, 3, 4, 5$ with 5 being the best. Another measurement Z' is the rank of a classroom relative to other classes perceived by a homeroom teacher when the survey was conducted. Its support is the same as Z. The outcomes are classified into two categories: schoolwork and non-schoolwork. Schoolwork includes Math, English, Chinese, and Physical Education, subjects required by Zhongkao. Non-schoolwork mainly includes fine arts, i.e., music, dance, calligraphy, and painting, etc. that are not required by Zhongkao. In this dataset, the latent group g is unobserved competitiveness in Zhongkao within a classroom, i.e., a latent group g contains those students who are at the same competitive level in the same classroom. This is a reasonable assumption. Recall that the students in a class are randomly assigned, and the

Table 3: Descriptive Statistics (18,334 observations)

Variable	Mean	Std. Dev.	Min	Max
Grade9	0.4692	0.4991	0	1
Male	0.5110	0.4999	0	1
OnlyChild	0.4375	0.4961	0	1
Income	1.8472	0.4953	1	3
Parents' education	0.3903	0.4878	0	1
Z (relative rank, initial)	2.7586	1.0553	1	5
Z' (relative rank, current)	3.3975	0.9604	1	5
Private Tutoring (schoolwork)	0.3793	0.4852	0	1
Private Tutoring (non-schoolwork)	0.2492	0.4325	0	1

Note: This table shows the summary statistics for 18,334 students within 438 classes.

distribution of academic performance in a class should be close to that of the school as a whole. Therefore, the latent group structure in each classroom should be close to identical and students only need to compete with the same latent group of students in his/her class. Under this assumption, Z^* is the unobserved competitiveness in Zhongkao that characterizes group g. Homeroom teachers' subjective evaluations Z and Z' can be understood as $Z = g(Z^*, X, \varepsilon); Z' = h(Z^*, X, \eta)$, where ε and η are random shocks to the teachers' subjective evaluation, e.g., personal judgment, which are independent across the two evaluations. Because middle school teachers' performance is evaluated based on the rate of students who are admitted to high school. It is a natural assumption that Z, Z', and Z^* have the same support. Under the assumption that ε and η are independent, $Z \perp Z'|Z^*, X$ holds. The outcome variable Y is determined by X, Z^*, peer effects, and a random shock through equation (2.1). Once we condition on Z^* and X, the random shock that affects the tutoring decision, i.e., a student's desirability, is independent from random shocks to teachers' subjective evaluation ε and η. This results in $Y \perp Z \perp Z'|Z^*, X$.

7.2 Estimation of peer effects

We apply the semiparametric estimation method to estimate the peer effects in private tutoring. In the estimation, we assume the distribution of ϵ is standard normal.

We present our estimation results (coefficients) in Tables 4 and 5 for schoolwork and non-schoolwork tutoring, respectively. For comparison, we also provide in columns (b) and (c) the "Naïve" estimates which assume away the latent group structure and treat Z and Z' as Z^*, respectively. Table 4 shows that using our method, gender and family income significantly affect participation in tutoring for schoolwork. Specifically, male students and those whose family income is higher are more likely to pursue tutoring for schoolwork. These findings are consistent with the existing literature. For example, Sun et al. (2020) show that the benefits of tutoring (educational and especially psychological) tend to be larger for male students than for female students. Zhang and Xie (2016) document that higher parental education and higher family income significantly increase both the probability of receiving private tutoring and the level of spending on it.

We also find significant peer effects in private tutoring for schoolwork. Based on the estimated coefficients, we estimate the marginal peer effects at the median. That is, for each possible value of $Z^* \in \{1, 2, 3, 4, 5\}$ we estimate the partial effects of m on Y at the median values of X. The corresponding estimates for different values of Z^* are $0.0005, 0.0183, 0.1735, 0.5242$, and 0.4623, respectively, with an average 0.24. That is, at the median values of X, the peer effect increases the probability of schoolwork tutoring by 24%. Our estimate of peer effects is consistent with the fact that students in a classroom compete with each other for Zhongkao. As a consequence, a student is more likely to participate in private tutoring on schoolwork if others in the same latent competitiveness group get private tutoring.

By contrast, ignoring the latent group leads to completely different results. First of all, there are no estimated peer effects in private tutoring for schoolwork. When the latent group structure is neglected and Z or Z' are used for Z^*, the peer effects at the median of X are -6.0% and -23%, respectively. Although we find that treating Z or Z' as Z^* underestimates the peer effects, in general the direction of the bias is unknown *ex ante*. Moreover, when Z is used for Z^*, 9th graders and males are less likely to pursue tutoring while those students who are only children are more likely to do so. When Z' is used for Z^*, none of the covariates affects tutoring. These results are less reliable than those estimated using our method. A comparison of the results indicates that neglecting the latent group structure could lead to significant bias.

Table 4: Empirical Results: Schoolwork

		Naïve estimation	
Variable	SNLS	Z	Z'
	(a)	(b)	(c)
Grade9	-0.1487	-0.8110***	-0.2127
	(-1.2602)	(-2.6536)	(-0.2117)
Male	0.1080*	-1.0852*	0.0192
	(1.8514)	(-1.6481)	(0.0193)
OnlyChild	0.6017	1.6886***	0.9422
	(0.1704)	(5.5904)	(1.5139)
Income	1.0580**	-0.5407	0.6321
	(2.0326)	(-1.1619)	(1.0869)
Parents' Education	0.9027	0.6931	0.5591
	(1.3110)	(1.3829)	(0.5801)
Peer Effects	1.4793**	-0.4090	-1.3496
	(2.1110)	(-0.0724)	(-0.5543)
Constant	-7.8635	-2.6058	-5.3204***
	(-1.2198)	(-1.1404)	(-4.5740)

Note: t-statistics in parentheses.
* 10% significant, ** 5% significant, *** 1% significant.

We present the results for non-schoolwork tutoring in Table 5. Different from the results of schoolwork tutoring, we estimate that male students are less likely, and students with higher family income are more likely, to participate in non-schoolwork tutoring. Our findings are supported by the existing literature. For example, Deer et al. (2023) reports that girls are more interested in and more engaged with visual art activities than boys. In estimating the peer effects, both our method and the naïve one find no peer effects for non-schoolwork tutoring. The absence of peer effects for non-schoolwork tutoring is consistent with the fact that non-schoolwork tutoring is primarily driven by individual interests and there is no competition involved.

Table 5: Empirical Results: Non-school work

Variable	SNLS	Naïve estimation	
		Z	Z'
	(a)	(b)	(c)
Grade9	-0.2170	-0.4444**	-0.3901
	(-0.6647)	(-2.5351)	(-0.2600)
Male	-1.1548*	-1.0078***	-0.7996
	(-1.8978)	(-10.3968)	(-0.3105)
OnlyChild	1.0279	0.5907*	0.4178
	(0.7061)	(1.7382)	(0.2578)
Income	0.8008***	0.5932***	0.5183
	(2.7800)	(6.3579)	(0.1034)
Parents' Education	1.3071	1.3883	0.8999*
	(0.1736)	(0.4186)	(1.8390)
Peer Effects	0.0004	-0.7503	0.0260
	(0.1271)	(-0.1258)	(0.0045)
Constant	-6.5199*	-5.7336**	-5.9649
	(-1.6938)	(-2.1046)	(-1.1006)

Note: t-statistics in parentheses.
* 10% significant, ** 5% significant, *** 1% significant.

8 Conclusion

In the context of binary choice models involving social interactions and unknown group structures, this paper presents an identification method for the underlying model primitives. This is achieved by employing the eigen-decomposition technique with two proxies and a monotonicity condition. Additionally, we introduce a two-stage SNLS method for estimating model parameters and derive its asymptotic properties. For inferential purposes, we also offer a dependent-robust subsampling method. As an application of the proposed method, we investigate social interactions among secondary school students and find positive and significant peer effects in their choices towards private schoolwork tutoring. However, there are no peer effects in non-school tutoring. We also find that ignoring the latent group can lead to significant bias in estimating peer effects.

References

Aguirregabiria, V. and P. Mira (2007). Sequential estimation of dynamic discrete games. *Econometrica 75*(1), 1–53.

Aguirregabiria, V. and P. Mira (2019). Identification of games of incomplete information with multiple equilibria and unobserved heterogeneity. *Quantitative Economics 10*(4), 1659–1701.

An, Y. (2017). Identification of first-price auctions with non-equilibrium beliefs: A measurement error approach. *Journal of Econometrics 200*(2), 326–343.

An, Y., S. Hong, and D. Zhang (2023). A structural analysis of simple contracts. *Journal of Econometrics 236*(2), 105456.

An, Y., Y. Hu, and M. Shum (2010). Estimating first-price auctions with an unknown number of bidders: A misclassification approach. *Journal of Econometrics 157*(2), 328–341.

An, Y., Y. Hu, and R. Xiao (2021). Dynamic decisions under subjective expectations: A structural analysis. *Journal of Econometrics 222*(1), 645–675.

Andrew, A. L., K.-W. E. Chu, and P. Lancaster (1993). Derivatives of eigenvalues and eigenvectors of matrix functions. *SIAM journal on matrix analysis and applications 14*(4), 903–926.

Andrews, D. W. (1994). Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica 62*(1), 43–72.

Aradillas-Lopez, A. (2010). Semiparametric estimation of a simultaneous game with incomplete information. *Journal of Econometrics 157*(2), 409–431.

Bailey, M., D. Johnston, T. Kuchler, J. Stroebel, and A. Wong (2022). Peer effects in product adoption. *American Economic Journal: Applied Economics 14*(3), 488–526.

Bajari, P., H. Hong, J. Krainer, and D. Nekipelov (2010). Estimating static models of strategic interactions. *Journal of Business & Economic Statistics 28*(4), 469–482.

Brock, W. A. and S. N. Durlauf (2001). Discrete choice with social interactions. *The Review of Economic Studies 68*(2), 235–260.

Brock, W. A. and S. N. Durlauf (2007). Identification of binary choice models with social interactions. *Journal of Econometrics 140*(1), 52–75.

Calvó-Armengol, A., E. Patacchini, and Y. Zenou (2009). Peer effects and social networks in education. *The Review of Economic Studies 76*(4), 1239–1267.

Chen, Q. and Z. Fang (2019). Improved inference on the rank of a matrix. *Quantitative Economics 10*(4), 1787–1824.

Cunha, F., J. J. Heckman, and S. M. Schennach (2010). Estimating the technology of cognitive and noncognitive skill formation. *Econometrica 78*(3), 883–931.

De Paula, A. and X. Tang (2012). Inference of signs of interaction effects in simultaneous games with incomplete information. *Econometrica 80*(1), 143–172.

Deer, G., E. Tadesse, Z. Chen, S. Khalid, and C. Gao (2023). The impact of chinese adolescents visual art participation on self-efficacy: A serial mediating role of cognition and emotion. *Plos one 18*(11), e0288379.

Diamond, W. and N. Agarwal (2017). Latent indices in assortative matching models. *Quantitative Economics 8*(3), 685–728.

Duflo, E. and E. Saez (2002). Participation and investment decisions in a retirement plan: The influence of colleagues' choices. *Journal of Public Economics 85*(1), 121–148.

Eble, A. and F. Hu (2022). Gendered beliefs about mathematics ability transmit across generations through children's peers. *Nature Human Behaviour 6*(6), 868–879.

Gaviria, A. and S. Raphael (2001). School-based peer effects and juvenile behavior. *Review of Economics and Statistics 83*(2), 257–268.

Glaeser, E. L., B. Sacerdote, and J. A. Scheinkman (1996). Crime and social interactions. *The Quarterly Journal of Economics 111*(2), 507–548.

Glaeser, E. L., B. I. Sacerdote, and J. A. Scheinkman (2003). The social multiplier. *Journal of the European Economic Association 1*(2-3), 345–353.

Golub, G. H. and C. F. Van Loan (2013). *Matrix Computations*. JHU press.

Horst, U. and J. A. Scheinkman (2006). Equilibria in systems of social interactions. *Journal of Economic Theory 130*(1), 44–77.

Hu, Y. (2008). Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution. *Journal of Econometrics 144*(1), 27–61.

Hu, Y. (2017). The econometrics of unobservables: Applications of measurement error models in empirical industrial organization and labor economics. *Journal of Econometrics 200*(2), 154–168.

Hu, Y., D. McAdams, and M. Shum (2013). Identification of first-price auctions with non-separable unobserved heterogeneity. *Journal of Econometrics 174*(2), 186–193.

Hu, Y. and S. M. Schennach (2008). Instrumental variable treatment of nonclassical measurement error models. *Econometrica 76*(1), 195–216.

Hu, Y. and M. Shum (2012). Nonparametric identification of dynamic models with unobserved state variables. *Journal of Econometrics 171*(1), 32–44.

Kasy, M. (2019). Uniformity and the delta method. *Journal of Econometric Methods 8*(1), 1–19.

Kleibergen, F. and R. Paap (2006). Generalized reduced rank tests using the singular value decomposition. *Journal of Econometrics 133*(1), 97–126.

Leary, M. T. and M. R. Roberts (2014). Do peer firms affect corporate financial policy? *The Journal of Finance 69*(1), 139–178.

Lee, L.-f., J. Li, and X. Lin (2014). Binary choice models with social network under heterogeneous rational expectations. *Review of Economics and Statistics 96*(3), 402–417.

Leung, M. P. (2022). Dependence-robust inference using resampled statistics. *Journal of Applied Econometrics 37*(2), 270–285.

Lewbel, A., X. Qu, and X. Tang (2023). Social networks with unobserved links. *Journal of Political Economy 131*(4), 898–946.

Lewbel, A. and X. Tang (2015). Identification and estimation of games with incomplete information using excluded regressors. *Journal of Econometrics 189*(1), 229–244.

Li, T., I. Perrigne, and Q. Vuong (2000). Conditionally independent private information in ocs wildcat auctions. *Journal of Econometrics 98*(1), 129–161.

Lin, Z. and Y. Hu (2024). Binary choice with misclassification and social interactions, with an application to peer effects in attitude. *Journal of Econometrics 238*(1), 105551.

Lin, Z., X. Tang, and N. N. Yu (2021). Uncovering heterogeneous social effects in binary choices. *Journal of Econometrics 222*(2), 959–973.

Luo, Y., P. Xiao, and R. Xiao (2022). Identification of dynamic games with unobserved heterogeneity and multiple equilibria. *Journal of Econometrics 226*(2), 343–367.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies 60*(3), 531–542.

Manski, C. F. (2013). Identification of treatment response with social interactions. *The Econometrics Journal 16*(1), S1–S23.

Marsden, P. V. (1990). Network data and measurement. *Annual Review of Sociology 16*(1), 435–463.

Menzel, K. (2016). Inference for games with many players. *The Review of Economic Studies 83*(1), 306–337.

Nakajima, R. (2007). Measuring peer effects on youth smoking behaviour. *The Review of Economic Studies 74*(3), 897–935.

Newey, W. K. (1994a). The asymptotic variance of semiparametric estimators. *Econometrica 62*(6), 1349–1382.

Newey, W. K. (1994b). 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. *Handbook of Econometrics 4*, 2111–2245.

Perrigne, I. and Q. Vuong (2011). Nonparametric identification of a contract model with adverse selection and moral hazard. *Econometrica 79*(5), 1499–1539.

Perrigne, I. and Q. Vuong (2012). On the identification of the procurement model. *Economics Letters 114*(1), 9–11.

Rio, E. (2017). *Asymptotic Theory of Weakly Dependent Random Processes*. Springer.

Robin, J.-M. and R. J. Smith (2000). Tests of rank. *Econometric Theory 16*(2), 151–175.

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.

Song, K. (2016). Ordering-free inference from locally dependent data. *arXiv preprint arXiv:1604.00447*.

Sun, L., M. N. Shafiq, M. McClure, and S. Guo (2020). Are there educational and psychological benefits from private supplementary tutoring in mainland china? evidence from the china education panel survey, 2013–15. *International Journal of Educational Development 72*, 102144.

Tamer, E. (2003). Incomplete simultaneous discrete response model with multiple equilibria. *The Review of Economic Studies 70*(1), 147–165.

Wan, Y. and H. Xu (2014). Semiparametric identification of binary decision games of incomplete information with correlated private signals. *Journal of Econometrics 182*(2), 235–246.

Wooldridge, J. M. (1994). Estimation and inference for dependent processes. *Handbook of Econometrics 4*, 2639–2738.

Xiao, R. (2018). Identification and estimation of incomplete information games with multiple equilibria. *Journal of Econometrics 203*(2), 328–343.

Xu, H. (2018). Social interactions in large networks: A game theoretic approach. *International Economic Review 59*(1), 257–284.

Yang, C. and L.-f. Lee (2017). Social interactions under incomplete information with heterogeneous expectations. *Journal of Econometrics 198*(1), 65–83.

Zhang, L. (2020). Spillovers of program benefits with mismeasured networks. *arXiv preprint arXiv:2009.09614*.

Zhang, Y. and Y. Xie (2016). Family background, private tutoring, and children's educational performance in contemporary china. *Chinese sociological review 48*(1), 64–82.

Appendix A Proofs of main results

Proof of Lemma 2.1

Proof. Define $R(Z_g^*, \nu_g) = \int F_\epsilon \left(\alpha + \beta m_g + \gamma X_i + Z_g^*\right) dF_{X|Z^*}$. By the boundedness of f_ϵ and Leibniz rule, we have

$$\begin{aligned}\frac{\partial R}{\partial m} =& \int f_\epsilon \left(\alpha + \beta m_g + \gamma X_i + Z_g^*\right) \beta dF_{X|Z^*} \\ \leq& |\beta| \int f_\epsilon \left(\alpha + \beta m_g + \gamma X_i + Z_g^*\right) dF_{X|Z^*} \\ \leq& |\beta| \int \sup_e f_\epsilon \left(e\right) dF_{X|Z^*} \\ <& 1\end{aligned}$$

if $|\beta| < \dfrac{1}{\sup_e f_\epsilon(e)}$. Therefore, the fixed point m_g is unique by Banach fixed-point theorem. □

Proof of Lemma 3.1

Proof. First, note that the law of total probability implies

$$f_{Z,Z',Y|X}(z, z', y|x) = \sum_{z^* \in \text{supp}(Z^*)} f_{Z,Z',Y,Z^*|X}(z, z', y, z^*|x),$$

where

$$\begin{aligned}f_{Z,Z',Y,Z^*|X}(z, z', y, z^*|x) &= f_{Z|Z^*,Z',Y,X}(z|z', y, z^*, x) f_{Z'|Z^*,Y,X}(z'|z^*, y, x) f_{Y|Z^*,X}(y|z^*, x) f_{Z^*|X}(z^*|x) \\ &= f_{Z|Z^*,X}(z|z^*, x) f_{Z'|Z^*,X}(z'|z^*, x) f_{Y|Z^*,X}(y|z^*, x) f_{Z^*}(z^*)\end{aligned}$$

by Assumptions 3.1 and 3.2. Similarly, we can show that

$$f_{Z,Z'|X}(z, z'|x) = \sum_{z^* \in \text{supp}(Z^*)} f_{Z|Z^*,X}(z|z^*, x) f_{Z'|Z^*,X}(z'|z^*, x) f_{Z^*}(z^*)$$

□

Proof of Lemma 3.2

Proof. (⇒) By Assumption 3.3, the rank of D_{Z^*} is K. By the rank inequality, for any $p \times n$ matrix A and $n \times q$ matrix B,

$$\text{rank}(A) + \text{rank}(B) - n \leq \text{rank}(AB) \leq \min\{\text{rank}(A), \text{rank}(B)\}.$$

Then, we can show that $M_{Z|Z^*,X}D_{X^*}$ has rank K. By applying the inequality again, we can conclude that the matrix

$$M_{Z,Z'|X} = M_{Z|Z^*,X} D_{Z^*} M^T_{Z'|Z^*,X} \tag{A.1}$$

has rank K.

($\Leftarrow$) Suppose $\operatorname{rank}(M_{Z,Z'|X}) = K$. Then, by (A.1) and the rank inequality above, we have $\operatorname{rank}(M_{Z|Z^*,X}) \geq K$ and $\operatorname{rank}(M_{Z'|Z^*,X}) \geq K$. By Assumption 3.3, both matrices are $K \times K$, which leads to the conclusion. □

Proof of Lemma 3.3

Proof. Since Y is binary,

$$f_{Y|Z^*,X} = \mathbb{E}(Y|Z^*,X)^Y[1 - \mathbb{E}(Y|Z^*,X)]^{1-Y}.$$

Therefore, $f_{Y|Z^*,X}$ is strictly monotonic in Z^* if and only if $\mathbb{E}(Y|Z^*,X)$ is strictly monotonic in Z^*. Since

$$\frac{\partial \mathbb{E}(Y|Z^*,X)}{\partial Z^*} = f_\epsilon(\alpha + \beta m + \gamma X + Z^*)\left(\beta \frac{\partial m}{\partial Z^*} + 1\right).$$

By Leibniz rule, taking the derivative with respect to Z^* on both sides of equation (3.2) gives

$$\frac{\partial m}{\partial Z^*} = \int f_\epsilon\left(\alpha + \beta m + \gamma X + Z^*\right)\left(\beta \frac{\partial m}{\partial Z^*} + 1\right) dF_X.$$

Therefore,

$$\frac{\partial m}{\partial Z^*} = \frac{\int f_\epsilon\left(\alpha + \beta m + \gamma X + Z^*\right) dF_X}{1 - \beta \int f_\epsilon\left(\alpha + \beta m + \gamma X + Z^*\right) dF_X}.$$

It implies that

$$\begin{aligned}\beta \frac{\partial m^*}{\partial Z^*} + 1 &= \frac{\beta \int f_\epsilon\left(\alpha + \beta m + \gamma X + Z^*\right) dF_X}{1 - \beta \int f_\epsilon\left(\alpha + \beta m + \gamma X_i + Z^*\right) dF_X} + 1 \\ &= \frac{1}{1 - \beta \int f_\epsilon\left(\alpha + \beta m + \gamma X + Z^*\right) dF_X} \\ &> 0\end{aligned}$$

by Assumption 2.1. □

Proof of Proposition 3.1

Proof. Lemma 3.1 implies that

$$f_{Z,Z',Y,Z^*|X}(z,z',y,z^*|x) = f_{Z|Z^*,X}(z|z^*,x) f_{Z'|Z^*,X}(z'|z^*,x) f_{Y|Z^*,X}(y|z^*,x) f_{Z^*}(z^*)$$

and

$$f_{Z,Z'|X}(z,z'|x) = \sum_{z^* \in \text{supp}(Z^*)} f_{Z|Z^*,X}(z|z^*,x) f_{Z'|Z^*,X}(z'|z^*,x) f_{Z^*}(z^*)$$

which have the matrix representations

$$M_{Z,Z',Y|X} = M_{Z|Z^*,X} D_{Y|Z^*,X} D_{Z^*} M^T_{Z'|Z^*,X} \tag{A.2}$$

and

$$M_{Z,Z'|X} = M_{Z|Z^*,X} D_{Z^*} M^T_{Z'|Z^*,X}. \tag{A.3}$$

By Assumption 3.4, the matrices $M_{Z|Z^*,X}$ and $M_{Z'|Z^*,X}$ are invertible. Therefore,

$$M^{-1}_{Z,Z'|X} = \left(M^T_{Z'|Z^*,X}\right)^{-1} D^{-1}_{Z^*} \left(M_{Z|Z^*,X}\right)^{-1}. \tag{A.4}$$

Consequently, we post-multiply equation (A.2) by (A.4), which leads to

$$M_{Z,Z',Y|X} M^{-1}_{Z,Z'|X} = M_{Z|Z^*,X} D_{Y|Z^*,X} M^{-1}_{Z|Z^*,X}. \tag{A.5}$$

The matrices on the left-hand side of (A.5) can be directly computed from the data, while the matrices on-the right hand side are of particular interest. Moreover, this representation implies that the matrices of the joint conditional densities on the left-hand side admit an eigenvalue-eigenvector decomposition. Consequently, $M_{Z|Z^*,X}$ can be identified as eigenvectors up to permutation of its columns, and $D_{Y|Z^*,X}$ can be identified as eigenvalues up to permutation of its diagonal entries(Hu 2008). Since each column in $M_{Z|Z^*,X}$ represents an entire distribution, the column sum should be 1, through which normalization can be performed.

Lemma 3.3 ensures that there are no duplicate values for the diagonal elements in $D_{Y|Z^*,X}$, which correspond to $f_{Y|Z^*,X}$. Therefore, the eigenvectors are linearly independent with each other. The next step is to determine which eigenvalues corresponds to $f_{Y|Z^*,X}(\cdot|j,x)$ for $j = 1, 2, ..., K$. Lemma 3.3 directly imposes an ordering for the eigenvalues and hence the eigenvectors. Consequently, we can identify $f_{Y|Z^*,X}$ by checking the ordering of each eigenvalue, and $f_{Z|Z^*,X}$ can be identified similarly.

Next, we need to identify m^*. By the law of total probability and Assumption 3.1

$$f_{Y|Z^*}(\cdot|\cdot) = \sum_{x \in \text{supp}(X)} f_{Y|Z^*,X}(\cdot|\cdot,x) f_X(x). \tag{A.6}$$

Since Y is binary,

$$f_{Y|Z^*} = \mathbb{E}(Y|Z^*)^Y [1 - \mathbb{E}(Y|Z^*)]^{1-Y}.$$

Therefore, m^* can be identified as $f_{Y|Z^*}(1|\cdot)$. □

Proof of Lemma 3.4

Proof. The law total probability implies that the conditional density of Z can be represented as

$$f_{Z|Z',X}(z|z',x) = \sum_{z^*\in \text{supp}(Z^*)} f_{Z|Z^*,Z',X}(z|z^*,z',x) f_{Z^*|Z',X}(z^*|z',x) = \sum_{z^*\in \text{supp}(Z^*)} f_{Z|Z^*,X}(z|z^*,x) f_{Z^*|Z',X}(z^*|z',x),$$

where the last equality is by Assumption 3.2. It has the matrix representation

$$M_{Z|Z',X} = M_{Z|Z^*,X} M_{Z^*|Z',X},$$

where

$$M_{Z|Z',X} = \left[f_{Z|Z',X}(z_l|z_k',x)\right]_{l=1,2,\dots,K;k=1,2,\dots,K}$$

and

$$M_{Z^*|Z',X} = \left[f_{Z^*|Z}(z_l^*|z_k',x)\right]_{l=1,2,\dots,K;k=1,2,\dots,K}.$$

By Assumption 3.4, the matrix $M_{Z|Z^*,X}$ is invertible. Therefore, the conditional density of Z^* can be identified as

$$M_{Z^*|Z',X} = M_{Z|Z^*,X}^{-1} M_{Z|Z',X}, \tag{A.7}$$

where $M_{Z|Z^*,X}$ is identified by Proposition 3.1 and the matrix $M_{Z|Z',X}$ is directly identifiable from data. □

Proof of Proposition 4.1

Proof. We prove this proposition in three steps:

Step 1. First, we show that the estimator $\widehat{m}$ is uniformly consistent at rate $n^{-1/2}$. By Lemma B.1, the frequency estimators $f_{Z,Z',Y,X}$ and f_X are uniformly consistent at rate $n^{-1/2}$. Hence, by Assumption 4.1 (iii) and Slutsky's theorem, we have

$$\left\|\widehat{f}_{Z,Z',Y|X} - f_{Z,Z',Y|X}\right\|_\infty = \left\|\frac{\widehat{f}_{Z,Z',Y}}{\widehat{f}_X} - f_{Z,Z',Y|X}\right\|_\infty = O_p(n^{-1/2}),$$

which implies that the matrix $\widehat{M}_{Z,Z',Y|X}$ is also uniformly consistent for $M_{Z,Z',Y|X}$ at rate $n^{-1/2}$. Similarly, we have

$$\left\|\widehat{f}_{Z,Z'|X} - f_{Z,Z'|X}\right\|_\infty = \left\|\frac{\widehat{f}_{Z,Z'}}{\widehat{f}_X} - f_{Z,Z'|X}\right\|_\infty = O_p(n^{-1/2}).$$

By notation abuse, define $v_0 = (f_{Z,Z',Y|X}, f_{Z,Z'|X})^T$ and $\widehat{v} = (\widehat{f}_{Z,Z',Y|X}, \widehat{f}_{Z,Z'|X})^T$. By Lemma 3 of Hu (2008), there exists a neighborhood of $(\text{vec}(M_{Z,Z',Y|X})^T, \text{vec}(M_{Z,Z'|X})^T)^T$ such that the eigenvalue function $\psi(\cdot)$ is analytical and

$$\sup_{\|\widehat{v}-v_0\|_\infty\leq\varepsilon}\left\|\psi\left(\widehat{M}_{Z,Z',Y|X}\widehat{M}^{-1}_{Z,Z|X}\right)-\psi\left(M_{Z,Z',Y|X}M^{-1}_{Z,Z|X}\right)\right\|_1=O_p(\|\widehat{v}-v_0\|_\infty),$$

where $\varepsilon>0$ is arbitrary and $\|\cdot\|_1$ is the L^1 norm. Hence, we have shown that $\widehat{f}_{Y|Z^*,X}$ is uniformly consistent for $f_{Y|Z^*,X}$ at rate $n^{-1/2}$. Because

$$\widehat{m}=\sum_{x\in\text{supp}(X)}\widehat{f}_{Y|Z^*,X}(1|z^*,x)\widehat{f}_X(x).$$

We can conclude that

$$\|\widehat{m}-m\|_\infty=O_p(n^{-1/2}).\tag{A.8}$$

Step 2. Then, we show that $\widehat{f}_{Z^*|Z',X}$ is uniformly consistent at rate $n^{-1/2}$. By Lemma B.1, the frequency estimators $\widehat{f}_Z$ and $\widehat{f}_{Z',X}$ are $\sqrt{n}$-uniformly consistent. Therefore, by Assumption 4.1 (iii) and Slutsky's theorem,

$$\left\|\widehat{f}_{Z|Z',X}-f_{Z|Z',X}\right\|_\infty=O_p(n^{-1/2}).$$

Again, by Lemma 3 of Hu (2008), there exists a neighborhood of $(\text{vec}(M_{Z,Z',Y|X})^T,\text{vec}(M_{Z,Z'|X})^T)^T$ such that the eigenvector function $\phi(\cdot)$ is analytical and

$$\sup_{\|\widehat{v}-v_0\|_\infty\leq\varepsilon}\left\|\phi\left(\widehat{M}_{Z,Z',Y|X}\widehat{M}^{-1}_{Z,Z|X}\right)-\phi\left(M_{Z,Z',Y|X}M^{-1}_{Z,Z|X}\right)\right\|_1=O_p(\|\widehat{v}-v_0\|_\infty),$$

which implies $\widehat{f}_{Z|Z^*,X}$ is $\sqrt{n}$-uniformly consistent. Since

$$\widehat{M}_{Z^*|Z',X}=\widehat{M}^{-1}_{Z|Z^*,X}\widehat{M}_{Z|Z',X}$$

and

$$\left\|\widehat{f}_{Z|Z',X}-f_{Z|Z',X}\right\|_\infty=\left\|\frac{\widehat{f}_Z}{\widehat{f}_{Z',X}}-f_{Z|Z'X}\right\|_\infty=O_p(n^{-1/2}).$$

we can conclude that

$$\left\|\widehat{f}_{Z^*|Z',X}-f_{Z^*|Z',X}\right\|_\infty=O_p(n^{-1/2}).\tag{A.9}$$

Step 3. (A.8) and (A.9) together indicate that $\|\widehat{\sigma}-\sigma_0\|_\infty=O_p(n^{-1/2})$. □

Proof of Theorem 4.1

Proof. We prove this theorem by verifying conditions (i)-(iv) of Theorem 2.1 in Newey and Mc-

Fadden (1994). Define

$$Q_n(\theta,\sigma) = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - g(Z_i', X_i; \theta, \sigma)\right]^2$$

and

$$Q(\theta,\sigma) = \mathbb{E}\left[Y_i - g(Z_i', X_i; \theta, \sigma)\right]^2.$$

It is straightforward to see that their conditions (ii) and (iii) are satisfied for $Q(\theta,\sigma_0)$ by Assumption 4.2 (i) and (ii). Furthermore, condition (i) is satisfied by Proposition 3.2. Therefore, we only need to verify condition (iv), i.e.,

$$\sup_{\theta\in\Theta}\left|Q_n(\theta,\widehat{\sigma}) - Q(\theta,\sigma_0)\right| = o_p(1). \tag{A.10}$$

By triangle inequality, the left-hand side of (A.10) is bounded as follows:

$$\sup_{\theta\in\Theta}\left|Q_n(\theta,\widehat{\sigma}) - Q(\theta,\sigma_0)\right| \leq \sup_{\theta\in\Theta}\left|Q_n(\theta,\widehat{\sigma}) - Q_n(\theta,\sigma_0)\right| + \sup_{\theta\in\Theta}\left|Q(\theta,\sigma_0) - Q(\theta,\sigma_0)\right|. \tag{A.11}$$

By Lemma B.2, the second term on the right-hand side of (A.11) is $o_p(1)$. Hence, we only need to show that the first term is $o_p(1)$.

By using the identity

$$\widehat{a}^2 - a = (\widehat{a} - a)^2 + 2a(\widehat{a} - a),$$

we can obtain

$$\begin{aligned}Q_n(\theta,\widehat{\sigma}) - Q_n(\theta,\sigma_0) =& \frac{1}{n}\sum_{i=1}^{n}\left[g(Z_i', X_i, \theta, \widehat{\sigma}) - g(Z_i', X_i, \theta, \sigma_0)\right]^2 \\ &+ \frac{1}{n}\sum_{i=1}^{n}2[Y_i - g(Z_i', X_i, \theta, \sigma_0)]\left[g(Z_i', X_i, \theta, \widehat{\sigma}) - g(Z_i', X_i, \theta, \sigma_0)\right] \\ \equiv& A_1 + A_2.\end{aligned} \tag{A.12}$$

Next, we show that

$$\sup_{\theta\in\Theta}\left|g(Z_i', X_i, \theta, \widehat{\sigma}) - g(Z_i', X_i, \theta, \sigma_0)\right| = o_p(1).$$

By the identify

$$\widehat{a}\widehat{b} = (\widehat{a} - a)b + a(\widehat{b} - b) + (\widehat{a} - a)(\widehat{b} - b)$$

and triangle inequality, we have

$$|g(Z_i', X_i, \theta, \widehat{\sigma}) - g(Z_i', X_i, \theta, \sigma_0)| \leq \sum_{z^*} |F_\epsilon(\alpha + \beta\widehat{m}_i + \gamma X_i + z^*) - F_\epsilon(\alpha + \beta m_i + \gamma X_i + z^*)| f_{Z^*|Z',X}(z^*|Z_i', X)$$
$$+ \sum_{z^*} F_\epsilon(\alpha + \beta m_i + \gamma X_i + z^*) \left| \widehat{f}_{Z^*|Z',X}(z^*|Z_i', X_i) - f_{Z^*|Z',X}(z^*|Z_i', X_i) \right|$$
$$+ \sum_{z^*} |F_\epsilon(\alpha + \beta\widehat{m}_i + \gamma X_i + z^*) - F_\epsilon(\alpha + \beta m_i + \gamma X_i + z^*)| \left| \widehat{f}_{Z^*|Z',X}(z^*|Z_i', X_i) - f_{Z^*|Z',X}(z^*|Z_i', X_i) \right|.$$

Therefore,

$$\sup_{\theta \in \Theta} \left| g(Z_i', X_i, \theta, \widehat{\sigma}) - g(Z_i', X_i, \theta, \sigma_0) \right| = o_p(1) \tag{A.13}$$

by Proposition 4.1 and the uniform continuous mapping theorem (see, e.g., Theorem 1 of Kasy 2019). (A.13) implies that

$$\sup_{\theta \in \Theta} A_1 = O_p(\|\widehat{\sigma} - \sigma_0\|_\infty) = o_p(1).$$

Furthermore, since $Y_i - g(Z_i, X_i, \theta, \sigma_0)$ is uniformly bounded,

$$\sup_{\theta \in \Theta} A_2 = O_p(\|\widehat{\sigma} - \sigma_0\|_\infty) = o_p(1).$$

Consequently, (A.12) implies that

$$\sup_{\theta \in \Theta} |Q_n(\theta, \widehat{\sigma}) - Q_n(\theta, \sigma_0)| = o_p(1).$$

Thus, (A.10) is verified and we can conclude by Theorem 2.1 of Newey and McFadden (1994) that $\widehat{\sigma} \xrightarrow{p} \sigma_0$. □

Proof of Theorem 4.2

Proof. From Assumption 4.3 (i) and the first order condition of the optimization problem (4.3), $\widehat{\theta}$ solves

$$\frac{1}{n}\sum_{i=1}^n [Y_i - g(Z_i', X_i; \widehat{\theta}, \widehat{\sigma})] \nabla_{\theta^T} g(Z_i', X_i; \widehat{\theta}, \widehat{\sigma}) = 0.$$

Define $s(W_i; \theta, \sigma) = [Y_i - g(Z_i', X_i; \theta, \sigma)] \nabla_{\theta^T} g(Z_i', X_i; \theta, \sigma)$. Then, by the mean value theorem we can obtain

$$\frac{1}{n}\sum_{i=1}^n s(W_i; \theta_0, \widehat{\sigma}) + \frac{1}{n}\sum_{i=1}^n \nabla_\theta s(W_i; \widetilde{\theta}, \widehat{\sigma})(\widehat{\theta} - \theta_0) = 0, \tag{A.14}$$

where $\widetilde{\theta}$ is between $\widehat{\theta}$ and θ_0. Following the proof of Theorem 8.1 in Newey and McFadden (1994), the major step in this proof is to show

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s(W_i;\theta_0,\widehat{\sigma}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[s(W_i;\theta_0,\sigma_0)+\rho(W_i)]+o_p(1). \tag{A.15}$$

(A.15) means $1/\sqrt{n}\sum_{i=1}^{n} s(W_i;\theta_0,\widehat{\sigma})$ has the same asymptotic distribution as $1/\sqrt{n}\sum_{i=1}^{n}[s(W_i;\theta_0,\sigma_0)+\rho(W_i)]$, which converges to a normal distribution. Consider

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s(W_i;\theta_0,\widehat{\sigma}) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n} s(W_i;\theta_0,\sigma_0) =& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{[Y_i - g(Z_i',X_i;\theta_0,\widehat{\sigma})]\nabla_{\theta^T} g(Z_i',X_i;\theta_0,\widehat{\sigma})\\ &- [Y_i - g(Z_i',X_i;\theta_0,\widehat{\sigma})]\nabla_{\theta^T} g(Z_i',X_i;\theta_0,\sigma)\}\end{aligned} \tag{A.16}$$

By the identity

$$\widehat{a}\widehat{b} = (\widehat{a}-a)b + a(\widehat{b}-b) + (\widehat{a}-a)(\widehat{b}-b),$$

the right-hand side of (A.16) equals

$$\begin{aligned}=&-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[g(Z_i',X_i;\theta_0,\widehat{\sigma}) - g(Z_i',X_i;\theta_0,\sigma_0)\right]\nabla_{\theta^T} g(Z_i',X_i;\theta_0,\sigma_0)\\ &+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[Y_i - g(Z_i',X_i;\theta_0,\sigma_0)]\left[\nabla_{\theta^T} g(Z_i',X_i;\theta_0,\widehat{\sigma}) - \nabla_{\theta^T} g(Z_i',X_i;\theta_0,\sigma_0)\right]\\ &-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[g(Z_i',X_i;\theta_0,\widehat{\sigma}) - g(Z_i',X_i;\theta_0,\sigma_0)\right]\left[\nabla_{\theta^T} g(Z_i',X_i;\theta_0,\widehat{\sigma}) - \nabla_{\theta^T} g(Z_i',X_i;\theta_0,\sigma_0)\right]\\ &\equiv B_1 + B_2 + B_3.\end{aligned}$$

First, we show the term B_3 is $o_p(1)$. Following the proof of Theorem 4.1, we have

$$|g(Z_i',X_i;\theta_0,\widehat{\sigma}) - g(Z_i',X_i;\theta_0,\sigma_0)| = O_p(\|\widehat{\sigma}-\sigma_0\|_\infty).$$

Define $\widetilde{W}_i = (1, m_i, X_i^T, z^*)^T$ and $\widehat{W}_i = (1, \widehat{m}_i, X_i^T, z^*)^T$. By the identify $\widehat{a}\widehat{b} = (\widehat{a}-a)b + a(\widehat{b}-b) +$

$(\widehat{a}-a)(\widehat{b}-b)$, we can obtain

$$
\begin{aligned}
&\nabla_{\theta^T} g(Z_i', X_i; \theta_0, \widehat{\sigma}) - \nabla_{\theta^T} g(Z_i', X_i; \theta_0, \sigma_0) \\
=&\sum_{z^*} f_\epsilon(\alpha + \beta\widehat{m}_i + \gamma X_i + z^*)\widehat{W}_i \widehat{f}_{Z^*|Z',X}(z^*|Z_i', X_i) - \sum_{z^*} f_\epsilon(\alpha + \beta m_i + \gamma X_i + z^*)\widetilde{W}_i f_{Z^*|Z',X}(z^*|Z_i', X_i) \\
=&\sum_{z^*} \left[f_\epsilon(\alpha + \beta\widehat{m}_i + \gamma X_i + z^*)\widehat{W}_i - f_\epsilon(\alpha + \beta m_i + \gamma X_i + z^*)\widetilde{W}_i \right] f_{Z^*|Z',X}(z^*|Z_i', X_i) \\
&+ \sum_{z^*} f_\epsilon(\alpha + \beta m_i + \gamma X_i + z^*)\widetilde{W}_i \left[\widehat{f}_{Z^*|Z',X}(z^*|Z_i', X_i) - f_{Z^*|Z',X}(z^*|Z_i', X_i) \right] \\
&+ \sum_{z^*} \left[f_\epsilon(\alpha + \beta\widehat{m}_i + \gamma X_i + z^*)\widehat{W}_i - f_\epsilon(\alpha + \beta m_i + \gamma X_i + z^*)\widetilde{W}_i \right] \left[\widehat{f}_{Z^*|Z',X_i}(z^*|Z_i', X_i) - f_{Z^*|Z',X}(z^*|Z_i', X_i) \right].
\end{aligned}
\tag{A.17}
$$

Furthermore, by the identity $\widehat{a}\widehat{b} = (\widehat{a}-a)b + a(\widehat{b}-b) + (\widehat{a}-a)(\widehat{b}-b)$, we have

$$
\begin{aligned}
&f_\epsilon(\alpha + \beta\widehat{m}_i + \gamma X_i + z^*)\widehat{W}_i - f_\epsilon(\alpha + \beta m_i + \gamma X_i + z^*)\widetilde{W}_i \\
=&\left[f_\epsilon(\alpha + \beta\widehat{m}_i + \gamma X_i + z^*) - f_\epsilon(\alpha + \beta m_i + \gamma X_i + z^*) \right] \widetilde{W}_i + f_\epsilon(\alpha + \beta m_i + \gamma X_i + z^*) \left(\widehat{W}_i - \widetilde{W}_i \right) \\
&+ \left[f_\epsilon(\alpha + \beta\widehat{m}_i + \gamma X_i + z^*) - f_\epsilon(\alpha + \beta m_i + \gamma X_i + z^*) \right] \left(\widehat{W}_i - \widetilde{W}_i \right).
\end{aligned}
\tag{A.18}
$$

Hence, by Assumption 4.3 (vi) and the uniform continuous mapping theorem,

$$
\left\| f_\epsilon(\alpha + \beta\widehat{m}_i + \gamma X_i + z^*)\widehat{W}_i - f_\epsilon(\alpha + \beta m_i + \gamma X_i + z^*)\widetilde{W}_i \right\| = O_p(\|\widehat{m} - m\|_\infty).
$$

Therefore, (A.17) implies that

$$
\left\| \nabla_{\theta^T} g(Z_i', X_i; \theta_0, \widehat{\sigma}) - \nabla_{\theta^T} g(Z_i', X_i; \theta_0, \sigma_0) \right\| = O_p(\|\widehat{\sigma} - \sigma_0\|_\infty).
$$

Consequently,

$$
\|B_3\| = O_p(\sqrt{n}\|\widehat{\sigma} - \sigma_0\|_\infty^2) = o_p(1) \tag{A.19}
$$

by Proposition 4.1. Next, by a second-order Taylor expansion,

$$
\begin{aligned}
B_1 &= -\frac{1}{\sqrt{n}} \sum_{i=1}^n \left[\nabla_{v^T} g(Z_i', X_i; \theta_0, \sigma_0)(\widehat{v} - v_0) + R \right] \nabla_{\theta^T} g(Z_i', X_i; \theta_0, \sigma_0) \\
&= -\frac{1}{\sqrt{n}} \sum_{i=1}^n \left[\nabla_{v^T} g(Z_i', X_i; \theta_0, \sigma_0)(\widehat{v} - v_0) \right] \nabla_{\theta^T} g(Z_i', X_i; \theta_0, \sigma_0) + o_p(1),
\end{aligned}
$$

where $R = O_p(\|\widehat{v} - v_0\|_\infty^2)$ by Assumption 4.3 (ii). Note that the second equality is due to Assumption 4.2 and Lemma B.1. Similarly, we can show that

$$B_2 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[Y_i - g(Z_i', X_i; \theta_0, \sigma_0)]\left[\nabla_{v^T}\nabla_{\theta^T} g(Z_i', X_i; \theta_0, \sigma_0)(\widehat{v} - v_0)\right] + o_p(1).$$

Therefore,

$$\begin{aligned} B_1 + B_2 = &\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{[Y_i - g(Z_i', X_i; \theta_0, \sigma_0)]\nabla_{v^T}\nabla_{\theta^T} g(Z_i', X_i; \theta_0, \sigma_0) \\ &- \nabla_{\theta^T} g(Z_i', X_i; \theta_0, \sigma_0)\nabla_{v^T} g(Z_i', X_i; \theta_0, \sigma_0)\}(\widehat{v} - v_0) + o_p(1) \\ = &\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\nabla_{v^T}\{[Y_i - g(Z_i', X_i; \theta_0, \sigma_0)]\nabla_{\theta^T} g(Z_i', X_i; \theta_0, \sigma_0)\}(\widehat{v} - v_0) + o_p(1). \end{aligned}$$

Define

$$G(w; \widehat{v} - v_0) = \nabla_{v^T}\{[y - g(z', x; \theta_0, \sigma_0)]\nabla_{\theta^T} g(z', x; \theta_0, \sigma_0)\}(\widehat{v} - v_0). \tag{A.20}$$

Next, similar to the proof of Theorem 8.1 in Newey and McFadden (1994), we need to prove two conditions: *stochastic equicontinuity* and *mean-square differentiability*. By Lemmas B.3 and B.4, these two conditions are verified. Hence, (A.15) is proved with $\rho(W_i) = \widetilde{l}(W_i) - \mathbb{E}[\widetilde{l}(W_i)]$ and

$$\widetilde{l}(w) = \mathbb{E}\left[\nabla_{v^T}\{[y - g(z', x; \theta_0, \sigma_0)]\nabla_{\theta^T} g(z', x; \theta_0, \sigma_0)\}\boldsymbol{\iota}_{d_v}\middle| W = w\right].$$

By Assumption 4.3 and the central limit theorem for weakly dependent random processes (e.g., Corollary 4.1 of Rio 2017),

$$-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[s(W_i; \theta_0, \sigma_0) + \rho(W_i)] \xrightarrow{d} N(0, D), \tag{A.21}$$

where

$$D = \lim_{n\to\infty}\frac{1}{n}\operatorname{Var}\left\{\sum_{i=1}^{n}[s(W_i; \theta_0, \sigma_0) + \rho(W_i)]\right\}$$

exists and is positive semi-definite. Furthermore,

$$\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta s(W_i; \widetilde{\theta}, \widehat{\sigma}) \xrightarrow{p} \mathbb{E}[\nabla_\theta s(W_i; \theta_0, \sigma_0)] \equiv H \tag{A.22}$$

by Lemma B.5. Finally, by Slutsky's theorem, (A.14), (A.21), (A.22) and Assumption 4.3 (v) together imply that

$$\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{d} N(0, H^{-1}DH^{-1}).$$

□

Proof of Proposition 4.2

Proof. The proof follows that of Leung (2022, Theorem A.1). First, we have the following decomposition

$$
\begin{aligned}
T_M(\theta_0;\pi) &= T_M(\theta_0;\pi) - \mathbb{E}\left[T_M(\theta_0;\pi)|\boldsymbol{W}\right] + \mathbb{E}\left[T_M(\theta_0;\pi)|\boldsymbol{W}\right] \\
&= \frac{1}{\sqrt{R_n}}\sum_{r=1}^{R_n}\left[\widehat{V}^{-1/2}\chi_{\pi_r(1)} - \mathbb{E}\left(\widehat{V}^{-1/2}\chi_{\pi_r(1)}|\boldsymbol{W}\right)\right] + \frac{1}{\sqrt{R_n}}\sum_{r=1}^{R_n}\mathbb{E}\left(\widehat{V}^{-1/2}\chi_{\pi_r(1)}|\boldsymbol{W}\right) \\
&\equiv C_1 + C_2. \qquad \text{(A.23)}
\end{aligned}
$$

According to the definition of π_r,

$$
\begin{aligned}
C_2 &= \sqrt{R_n}\widehat{V}^{-1/2}\frac{1}{|\Pi|}\sum_{\pi\in\Pi}\chi_{\pi(1)} = \sqrt{R_n}\widehat{V}^{-1/2}\frac{1}{n!}\sum_{\pi\in\Pi}\chi_{\pi(1)} \\
&= \sqrt{R_n}\widehat{V}^{-1/2}\frac{1}{n!}\sum_{i=1}^{n}\chi_i(n-1)! = \sqrt{\frac{R_n}{n}}\widehat{V}^{-1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\chi_i
\end{aligned}
$$

Note that $1/\sqrt{n}\sum_{i=1}^{n}\chi_i = O_p(1)$ by the proof of Theorem 4.2 and Lemma B.2. Furthermore, $\widehat{V} - V = o_p(1)$ can be proved analogously to Lemma B.5. Therefore, C_2 is $O_p\left(\sqrt{R_n/n}\right) = o_p(1)$ by conditions of this proposition.

Next, we apply the Lyapunov central limit theorem to C_1, which is a sum of conditionally i.i.d. random vectors. First ,we have

$$
C_1 = \frac{1}{\sqrt{R_n}}\sum_{r=1}^{R_n}\left(\widehat{V}^{-1/2}\chi_{\pi_r(1)} - \widehat{V}^{-1/2}\frac{1}{n}\sum_{i=1}^{n}\chi_i\right) = \frac{1}{\sqrt{R_n}}\sum_{r=1}^{R_n}\left[\widehat{V}^{-1/2}\left(\chi_{\pi_r(1)} - \overline{\chi}\right)\right].
$$

Then,

$$
\begin{aligned}
\mathrm{Var}(C_1|\boldsymbol{W}) = \mathbb{E}\left(C_1C_1^T|\boldsymbol{W}\right) &= \frac{1}{R_n}\sum_{r=1}^{R_n}\mathbb{E}\left[\widehat{V}^{-1/2}\left(\chi_{\pi_r(1)} - \overline{\chi}\right)\left(\chi_{\pi_r(1)} - \overline{\chi}\right)^T\left(\widehat{V}^{-1/2}\right)^T\middle|\boldsymbol{W}\right] \\
&= \frac{1}{|\Pi|}\sum_{\pi\in\Pi}\widehat{V}^{-1/2}\left(\chi_{\pi(1)} - \overline{\chi}\right)\left(\chi_{\pi(1)} - \overline{\chi}\right)^T\left(\widehat{V}^{-1/2}\right)^T \\
&= \frac{1}{n!}\sum_{i=1}^{n}\widehat{V}^{-1/2}\left(\chi_i - \overline{\chi}\right)\left(\chi_i - \overline{\chi}\right)^T\left(\widehat{V}^{-1/2}\right)^T(n-1)! \\
&= I_{d_\theta}.
\end{aligned}
$$

Similarly, we can show that for some $\lambda > 0$,

$$\mathbb{E}\left[\left\|\widehat{V}^{-1/2}\left(\chi_i - \overline{\chi}\right)\right\|^{2+\lambda}|\boldsymbol{W}\right] = \frac{1}{n}\sum_{i=1}^{n}\left\|\widehat{V}^{-1/2}\left(\chi_i - \overline{\chi}\right)\right\|^{2+\lambda} = \frac{1}{n}\sum_{i=1}^{n}\left\|V^{-1/2}\left(\widetilde{\chi}_i - \overline{\widetilde{\chi}}\right)\right\|^{2+\lambda} + o_p(1)$$

by the uniform continuous mapping theorem. Note that this is $O_p(1)$ by conditions (i) and (ii) and Minkowski's inequality, which verifies the Lyapunov condition. Hence, $C_1 \xrightarrow{d} N(0, I_{d_\theta})$ conditional on $\boldsymbol{W}$ by Lyapunov's central limit theorem. □

Appendix B Auxiliary lemmas

This section introduces useful lemmas that are used in the proofs of Appendix A.

Lemma B.1. *Under Assumption 4.1, the frequency estimator $\widehat{f}_W$ is a uniformly consistent estimator at rate $n^{-1/2}$, i.e.,*

$$\|\widehat{f}_W - f_W\|_\infty = O_p(n^{-1/2}).$$

Proof. We need to show that

$$\lim_{\varepsilon\to\infty}\limsup_{n\to\infty} Pr\left(\sqrt{n}\|\widehat{f}_W - f_W\|_\infty > \varepsilon\right) = 0.$$

Let K be the number of elements in the support of W, which is finite because W is discrete. Then,

$$\begin{aligned} Pr\left(\|\widehat{f}_W - f_W\|_\infty > \frac{\varepsilon}{\sqrt{n}}\right) &\leq \sum_{w\in\text{supp}(W)} Pr\left\{\left[\widehat{f}_W(w) - f_W(w)\right]^2 > \frac{\varepsilon^2}{n}\right\} \\ &\leq \sum_{w\in\text{supp}(W)} \mathbb{E}\left[\widehat{f}_W(w) - f_W(w)\right]^2 \frac{n}{\varepsilon^2} \\ &\leq \frac{Kn}{\varepsilon^2}\max_{w\in\text{supp}(W)} \mathbb{E}\left[\widehat{f}_W(w) - f_W(w)\right]^2, \end{aligned}$$

where the second inequality is by Chebyshev's inequality. It remains to show that $\mathbb{E}[\widehat{f}_W(w) - f_W(w)]^2$ is $O(n^{-1})$ for any $w \in \text{supp}(W)$. By Assumption 4.1,

$$\mathbb{E}\left[\widehat{f}_W(w) - f_W(w)\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(W_i = w) - f_W(w)\right] = 0.$$

Define $\mathbb{1}_i = \mathbb{1}(W_i = w) - f_W(w)$. Hence, by Assumption 4.1 we have

$$\begin{aligned}
\mathbb{E}\left[\widehat{f}_W(w) - f_W(w)\right]^2 &= \operatorname{Var}\left[\widehat{f}_W(w) - f_W(w)\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \operatorname{Var} \mathbb{1}_i + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}^{n} \operatorname{Cov}(\mathbb{1}_i, \mathbb{1}_j) \\
&= \frac{1}{n} \operatorname{Var} \mathbb{1}_i + \frac{1}{n^2}\sum_{g=1}^{G}\sum_{i\in\mathcal{N}_g}^{n_g}\sum_{j\neq i, j\in\mathcal{N}_g}^{n_g} \operatorname{Cov}(\mathbb{1}_i, \mathbb{1}_j) \\
&= O\left(n^{-1}\right),
\end{aligned}$$

Therefore,

$$\|\widehat{f}_W - f_W\|_\infty = O_p(n^{-1/2}).$$

□

Lemma B.2. *Suppose Assumption 4.1 (i) and (ii) hold. For any function $b : \operatorname{supp}(W) \times \Theta \mapsto \mathbb{R}^p$, if the following condition hold (i) Θ is compact; (ii) $b(w, \theta)$ is measurable of w for each $\theta \in \Theta$ and continuously differentiable in θ for all $w \in \operatorname{supp}(W)$; (iii) There exists a function $h(w)$ with $\mathbb{E}[h(w)] < \infty$ such that $b(w;\theta)^2 \leq h(w)$ and $\|\nabla_{\theta^T} b(w;\theta)\|^2 \leq h(w)$ for all w. Then,*

$$\sup_{\theta\in\Theta}\left\|\frac{1}{n}\sum_{i=1}^{n} b(W_i, \theta) - \mathbb{E}[b(W_i, \theta)]\right\| = o_p(1).$$

Proof. This proof is based on that of Theorem 4.2 in Wooldridge (1994). By notation abuse, define $Q_n(\theta) = 1/n\sum_{i=1}^{n} b(W_i, \theta)$ and $Q(\theta) = \mathbb{E}[b(W_i, \theta)]$. Let η be a positive number to be set later. Since Θ is compact, there exists a finite covering of Θ, say $B_\eta(\theta_j), j = 1, 2, ..., K(\eta)$, where $B_\eta(\theta_j)$ is the sphere of radius η about θ_j, i.e.,

$$B_\eta(\theta_j) = \{\theta \in \Theta | \; \|\theta - \theta_j\| < \eta\}.$$

Since $\Theta \subset \bigcup_{j=1}^{K} B_\eta(\theta_j)$, it follows that for all $\varepsilon > 0$,

$$\begin{aligned}
Pr\left[\sup_{\theta\in\Theta}\|Q_n(\theta) - Q(\theta)\| > \varepsilon\right] \leq & Pr\left[\max_{1\leq j\leq K(\eta)} \sup_{\theta\in B_\eta(\theta_j)} \|Q_n(\theta) - Q(\theta)\| > \varepsilon\right] \\
\leq & \sum_{j=1}^{K(\eta)} Pr\left[\sup_{\theta\in B_\eta(\theta_j)} \|Q_n(\theta) - Q(\theta)\| > \varepsilon\right]. \qquad \text{(B.1)}
\end{aligned}$$

We will show that each probability in the summand of (*B*.1) is $o(1)$. Define $b_i(\theta) = b(W_i, \theta)$. For

$\theta \in B_\eta(\theta_j)$,

$$
\begin{aligned}
\|Q_n(\theta) - Q(\theta)\| &\leq \|Q_n(\theta) - Q(\theta_j)\| + \|Q_n(\theta_j) - Q(\theta_j)\| + \|Q(\theta_j) - Q(\theta)\| \\
&\leq \frac{1}{n}\sum_{i=1}^n \|b_i(\theta) - Q(\theta_j)\| + \left\|\frac{1}{n}\sum_{i=1}^n b_i(\theta_j) - Q(\theta_j)\right\| + \frac{1}{n}\sum_{i=1}^n \|Q(\theta) - Q(\theta_j)\|
\end{aligned}
$$

by triangle inequality. By Conditions (i)-(ii) and the monotonicity of expectations,

$$
\|b_i(\theta) - b_i(\theta_j)\| \leq c_i \|\theta - \theta_j\| \leq \eta c_i
$$

and

$$
\|Q(\theta) - Q(\theta_j)\| \leq \overline{c}_i \|\theta - \theta_j\| \leq \eta \overline{c}_i,
$$

where $c_i = \sup_{\theta \in \Theta} \|\nabla_{\theta^T} b_i(\theta)\|$ and $\overline{c}_i = \mathbb{E}(c_i)$. Therefore, we have

$$
\begin{aligned}
\sup_{\theta \in B_\eta(\theta_j)} |Q_n(\theta) - Q(\theta)| &\leq \eta \left(\frac{1}{n}\sum_{i=1}^n c_i + \frac{1}{n}\sum_{i=1}^n \overline{c}_i\right) + \left\|\frac{1}{n}\sum_{i=1}^n b_i(\theta_j) - Q(\theta_j)\right\| \\
&\leq \frac{2\eta}{n}\sum_{i=1}^n \overline{c}_i + \eta \left\|\frac{1}{n}\sum_{i=1}^n c_i - \frac{1}{n}\sum_{i=1}^n \overline{c}_i\right\| + \left\|\frac{1}{n}\sum_{i=1}^n b_i(\theta_j) - Q(\theta_j)\right\| \\
&= 2\eta\overline{c}_i + \eta \left\|\frac{1}{n}\sum_{i=1}^n c_i - \overline{c}_i\right\| + \left\|\frac{1}{n}\sum_{i=1}^n b_i(\theta_j) - Q(\theta_j)\right\|.
\end{aligned}
$$

where the second inequality is by triangle inequality and the last equality is by Assumption 4.1. Since $\overline{c}_i \leq \overline{C} < \infty$ by dominated convergence theorem. It follows that

$$
Pr\left[\sup_{\theta \in B_\eta(\theta_j)} \|Q_n(\theta) - Q(\theta)\| > \varepsilon\right] \leq Pr\left[\eta \left\|\frac{1}{n}\sum_{i=1}^n c_i - \overline{c}_i\right\| + \left\|\frac{1}{n}\sum_{i=1}^n b_i(\theta_j) - Q(\theta_j)\right\| > \varepsilon - 2\eta\overline{C}\right].
$$

Now choose $\eta \leq 1$ such that $\varepsilon - 2\eta\overline{C} > \varepsilon/2$. Then

$$
Pr\left[\sup_{\theta \in B_\eta(\theta_j)} \|Q_n(\theta) - Q(\theta)\| > \varepsilon\right] \leq Pr\left[\left\|\frac{1}{n}\sum_{i=1}^n c_i - \overline{c}_i\right\| + \left\|\frac{1}{n}\sum_{i=1}^n b_i(\theta_j) - Q(\theta_j)\right\| > \frac{\varepsilon}{2}\right].
$$

Next, choose N so that

$$
Pr\left[\left\|\frac{1}{n}\sum_{i=1}^n c_i - \overline{c}_i\right\| + \left\|\frac{1}{n}\sum_{i=1}^n b_i(\theta_j) - Q(\theta_j)\right\| > \frac{\varepsilon}{2}\right] \leq \frac{\varepsilon}{K(\eta)}, \tag{B.2}
$$

for all $n \geq N$ and all $j = 1, 2, ..., K(\delta)$. Note that this is possible because under Assumption 4.1

and Condition (iii),

$$\left\| \frac{1}{n} \sum_{i=1}^{n} c_i - \bar{c}_i \right\| = o_p(1)$$

and

$$\left\| \frac{1}{n} \sum_{i=1}^{n} b_i(\theta_j) - Q(\theta_j) \right\| = o_p(1)$$

by using variance calculation similar to the proof of Lemma B.1. Consequently, (B.1) and (B.2) together imply that for all $n \geq N$ and $\varepsilon > 0$

$$Pr\left[\sup_{\theta \in \Theta} \| Q_n(\theta) - Q(\theta) \| > \varepsilon \right] \leq \varepsilon,$$

which proves the result. □

Lemma B.3. *Let $G(w; \widehat{v} - v_0)$ be defined as in (A.20). Then, under Assumptions 4.1 and 4.3,*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[G(W_i; \widehat{v} - v_0) - \int G(w; \widehat{v} - v_0) dF_W(w) \right] = o_p(1).$$

Proof. Since $G(w; \widehat{v} - v_0)$ is linear in $\widehat{v} - v_0$, we can rewrite it as $G(w; \widehat{v} - v_0) = \widetilde{l}(w)(\widehat{v} - v_0)$. Consequently,

$$\begin{aligned}
&\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[G(W_i; \widehat{v} - v_0) - \int G(w; \widehat{v} - v_0) dF_W(w) \right] \right\| \\
=& \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \widetilde{l}(W_i) - \mathbb{E}[\widetilde{l}(W_i)] \right\} (\widehat{v} - v_0) \right\| \\
\leq& \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \widetilde{l}(W_i) - \mathbb{E}[\widetilde{l}(W_i)] \right\} \right\| \|\widehat{v} - v_0\|_{\infty}.
\end{aligned}$$

Then,

$$\begin{aligned}
&\mathbb{E} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \widetilde{l}(W_i) - \mathbb{E}[\widetilde{l}(W_i)] \right\} \right)^2 \\
=& \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left\{ \widetilde{l}(W_i) - \mathbb{E}[\widetilde{l}(W_i)] \right\}^2 + \frac{1}{n} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \mathbb{E} \left\{ \widetilde{l}(W_i) - \mathbb{E}[\widetilde{l}(W_i)] \right\} \{ l(W_j) - \mathbb{E}[l(W_j)] \} \\
=& \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left\{ \widetilde{l}(W_i) - \mathbb{E}[\widetilde{l}(W_i)] \right\}^2 + \frac{1}{n} \sum_{g=1}^{G} \sum_{i \in \mathcal{N}_g}^{n_g} \sum_{j \neq i, j \in \mathcal{N}_g}^{n_g} \mathbb{E} \{ h(W_i) - \mathbb{E}[h(W_i)] \} \{ h(W_j) - \mathbb{E}[h(W_j)] \} \\
=& O(1)
\end{aligned}$$

by Assumptions 4.1 and 4.3. Hence, by Proposition 4.1 we can obtain

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[G(W_i;\widehat{v}-v_0)-\int G(w;\widehat{v}-v_0)dF_W(w)\right]=o_p(1).$$

□

Lemma B.4. *There exists a function $\rho:\mathrm{supp}(W)\mapsto\mathbb{R}^{d_\theta}$ such that*

$$\int G(w;\widehat{v}-v_0)dF_W(w)=\int\rho(w)d\widehat{F}_W(w),$$

where $\widehat{F}_W$ is the empirical distribution of W_i.

Proof. By the linearity of $G(w;v)$ in v and the law of iterated expectations, we have

$$\int G(w;v)dF_W(w)=\int\widetilde{l}(w)v(w)dw,$$

where

$$\widetilde{l}(w)=\mathbb{E}\left[\nabla_{v^T}\{[y-g(z,x;\theta_0,\sigma_0)]\nabla_{\theta^T}g(z,x;\theta_0,\sigma_0)\}\boldsymbol{\iota}_{d_v}|W=w\right].$$

Note that $\boldsymbol{\iota}_{d_v}$ is a $d_v\times 1$ vector of ones and d_v is the dimension of v. Furthermore, define $\rho(w)=\widetilde{l}(w)-\mathbb{E}[\widetilde{l}(w)]$, we can easily verify that

$$\int G(w;\widehat{v}-v_0)dF_W(w)=\int\rho(w)d\widehat{F}_W(w).$$

□

Lemma B.5. *Suppose the assumptions of Proposition 4.1 hold. Then, under Assumptions 4.1-4.3,*

$$\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta s(W_i;\widetilde{\theta},\widehat{\sigma})\xrightarrow{p}\mathbb{E}[\nabla_\theta s(W_i;\theta_0,\sigma_0)].$$

Proof. By triangle inequality,

$$\begin{aligned}&\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta s(W_i;\widetilde{\theta},\widehat{\sigma})-\mathbb{E}[\nabla_\theta s(W_i;\theta_0,\sigma_0)]\right\|\\ \leq&\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta s(W_i;\widetilde{\theta},\widehat{\sigma})-\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta s(W_i;\theta_0,\sigma_0)\right\|+\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta s(W_i;\theta_0,\sigma_0)-\mathbb{E}[\nabla_\theta s(W_i;\theta_0,\sigma_0)]\right\|.\end{aligned}\tag{B.3}$$

By Assumptions 4.1 and 4.3 (v) and using variance calculation similar to the proof of Lemma B.1, the second term on the right-hand side of (B.3) is $o_p(1)$. Therefore, we need to show that the first term is $o_p(1)$. Again, by triangle inequality we have

$$\begin{aligned}
&\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta s(W_i;\widetilde{\theta},\widehat{\sigma})-\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta s(W_i;\theta_0,\sigma_0)\right\| \\
\leq&\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta s(W_i;\widetilde{\theta},\widehat{\sigma})-\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta s(W_i;\widetilde{\theta},\sigma_0)\right\|+\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta s(W_i;\widetilde{\theta},\sigma_0)-\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta s(W_i;\theta_0,\sigma_0)\right\| \\
\equiv&\|D_1\|+\|D_2\|.
\end{aligned}$$

Note that $\|D_2\|=o_p(1)$ by Theorem 4.1 and Assumption 4.2 (ii). Now consider the term D_1. Since

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta s(W_i;\widetilde{\theta},\widehat{\sigma})-\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta s(W_i;\widetilde{\theta},\sigma_0) \\
=&\frac{1}{n}\sum_{i=1}^{n}\left\{[Y_i-g(Z_i',X_i;\widetilde{\theta},\widehat{\sigma})]\left[\nabla_\theta\nabla_{\theta^T}g(Z_i',X_i;\widetilde{\theta},\widehat{\sigma})\right]-[Y_i-g(Z_i',X_i;\widetilde{\theta},\sigma_0)]\left[\nabla_\theta\nabla_{\theta^T}g(Z_i',X_i;\widetilde{\theta},\sigma_0)\right]\right\} \\
&-\frac{1}{n}\sum_{i=1}^{n}\left[\nabla_{\theta^T}g(Z_i',X_i;\widetilde{\theta},\widehat{\sigma})\nabla_\theta g(Z_i',X_i;\widetilde{\theta},\widehat{\sigma})-\nabla_{\theta^T}g(Z_i',X_i;\widetilde{\theta},\sigma_0)\nabla_\theta g(Z_i',X_i;\widetilde{\theta},\sigma_0)\right].
\end{aligned}$$

Using the identity

$$\widehat{a}\widehat{b}=(\widehat{a}-a)b+a(\widehat{b}-b)+(\widehat{a}-a)(\widehat{b}-b),$$

we have for any $\widetilde{\theta}\in\Theta$,

$$\begin{aligned}
D_1=&-\frac{1}{n}\sum_{i=1}^{n}\left[g(Z_i',X_i;\widetilde{\theta},\widehat{\sigma})-g(Z_i',X_i;\widetilde{\theta},\sigma_0)\right]\nabla_\theta\nabla_{\theta^T}g(Z_i',X_i;\widetilde{\theta},\sigma_0) \\
&+\frac{1}{n}\sum_{i=1}^{n}[Y_i-g(Z_i',X_i;\widetilde{\theta},\sigma_0)]\left[\nabla_\theta\nabla_{\theta^T}g(Z_i',X_i;\widetilde{\theta},\widehat{\sigma})-\nabla_\theta\nabla_{\theta^T}g(Z_i',X_i;\widetilde{\theta},\sigma_0)\right] \\
&-\frac{1}{n}\sum_{i=1}^{n}\left[g(Z_i',X_i;\widetilde{\theta},\widehat{\sigma})-g(Z_i',X_i;\widetilde{\theta},\sigma_0)\right]\left[\nabla_\theta\nabla_{\theta^T}g(Z_i',X_i;\widetilde{\theta},\widehat{\sigma})-\nabla_\theta\nabla_{\theta^T}g(Z_i',X_i;\widetilde{\theta},\sigma_0)\right] \\
&-\frac{1}{n}\sum_{i=1}^{n}\left[\nabla_{\theta^T}g(Z_i',X_i;\widetilde{\theta},\widehat{\sigma})-\nabla_{\theta^T}g(Z_i',X_i;\widetilde{\theta},\sigma_0)\right]\nabla_\theta g(Z_i',X_i;\widetilde{\theta},\sigma_0) \\
&-\frac{1}{n}\sum_{i=1}^{n}\nabla_{\theta^T}g(Z_i',X_i;\widetilde{\theta},\sigma_0)\left[\nabla_\theta g(Z_i',X_i;\widetilde{\theta},\widehat{\sigma})-\nabla_\theta g(Z_i',X_i;\widetilde{\theta},\sigma_0)\right] \\
&-\frac{1}{n}\sum_{i=1}^{n}\left[\nabla_{\theta^T}g(Z_i',X_i;\widetilde{\theta},\widehat{\sigma})-\nabla_{\theta^T}g(Z_i',X_i;\widetilde{\theta},\sigma_0)\right]\left[\nabla_\theta g(Z_i',X_i;\widetilde{\theta},\widehat{\sigma})-\nabla_\theta g(Z_i',X_i;\widetilde{\theta},\sigma_0)\right].
\end{aligned}$$

Following the proof of Theorem 4.1, we can conclude that

$$\left| g(Z_i', X_i; \widetilde{\theta}, \widehat{\sigma}) - g(Z_i', X_i; \widetilde{\theta}, \sigma_0) \right| = O_p(\|\widehat{\sigma} - \sigma_0\|_\infty) \tag{B.4}$$

and

$$\left\| \nabla_{\theta^T} g(Z_i', X_i; \widetilde{\theta}, \widehat{\sigma}) - \nabla_{\theta^T} g(Z_i', X_i; \widetilde{\theta}, \sigma_0) \right\| = O_p(\|\widehat{\sigma} - \sigma_0\|_\infty). \tag{B.5}$$

Besides, by using the identity $\widehat{a}\widehat{b} = (\widehat{a} - a)b + a(\widehat{b} - b) + (\widehat{a} - a)(\widehat{b} - b)$ again, we have

$$\begin{aligned}
&\nabla_\theta \nabla_{\theta^T} g(Z_i', X_i; \widetilde{\theta}, \widehat{\sigma}) - \nabla_\theta \nabla_{\theta^T} g(Z_i', X_i; \widetilde{\theta}, \sigma_0) \\
=& \sum_{z^*} \nabla_\theta f_\epsilon(\widetilde{\theta}\widehat{W}_i)\widehat{W}_i\widehat{W}_i^T \widehat{f}_{Z^*|Z',X}(z^*|Z_i, X_i) - \sum_{z^*} \nabla_\theta f_\epsilon(\widetilde{\theta}\widetilde{W}_i)\widetilde{W}_i\widetilde{W}_i^T f_{Z^*|Z',X}(z^*|Z_i', X_i) \\
=& \sum_{z^*} \left[\nabla_\theta f_\epsilon(\widetilde{\theta}\widehat{W}_i)\widehat{W}_i\widehat{W}_i^T - \nabla_\theta f_\epsilon(\widetilde{\theta}\widetilde{W}_i)\widetilde{W}_i\widetilde{W}_i^T \right] f_{Z^*|Z',X}(z^*|Z_i', X_i) \\
&+ \sum_{z^*} f_\epsilon(\widetilde{\theta}\widetilde{W}_i)\widetilde{W}_i\widetilde{W}_i^T \left[\widehat{f}_{Z^*|Z',X}(z^*|Z_i', X_i) - f_{Z^*|Z',X}(z^*|Z_i', X_i) \right] \\
&+ \sum_{z^*} \left[\nabla_\theta f_\epsilon(\widetilde{\theta}\widehat{W}_i)\widehat{W}_i\widehat{W}_i^T - \nabla_\theta f_\epsilon(\widetilde{\theta}\widetilde{W}_i)\widetilde{W}_i\widetilde{W}_i^T \right] \left[\widehat{f}_{Z^*|Z',X}(z^*|Z_i', X_i) - f_{Z^*|Z',X}(z^*|Z_i', X_i) \right]
\end{aligned} \tag{B.6}$$

Furthermore,

$$\begin{aligned}
&\nabla_\theta f_\epsilon(\widetilde{\theta}\widehat{W}_i)\widehat{W}_i\widehat{W}_i^T - \nabla_\theta f_\epsilon(\widetilde{\theta}\widetilde{W}_i)\widetilde{W}_i\widetilde{W}_i^T \\
=& \left[\nabla_\theta f_\epsilon(\widetilde{\theta}\widehat{W}_i) - \nabla_\theta f_\epsilon(\widetilde{\theta}\widetilde{W}_i) \right] \widetilde{W}_i\widetilde{W}_i^T + \nabla_\theta f_\epsilon(\widetilde{\theta}\widetilde{W}_i) \left[\widehat{W}_i\widehat{W}_i^T - \widetilde{W}_i\widetilde{W}_i^T \right] \\
&+ \left[\nabla_\theta f_\epsilon(\widetilde{\theta}\widehat{W}_i) - \nabla_\theta f_\epsilon(\widetilde{\theta}\widetilde{W}_i) \right] \left[\widehat{W}_i\widehat{W}_i^T - \widetilde{W}_i\widetilde{W}_i^T \right]
\end{aligned} \tag{B.7}$$

by the identity $\widehat{a}\widehat{b} = (\widehat{a} - a)b + a(\widehat{b} - b) + (\widehat{a} - a)(\widehat{b} - b)$. By Assumption 4.3 and the uniform continuous mapping theorem, (B.6) and (B.7) together imply that

$$\left\| \nabla_\theta \nabla_{\theta^T} g(Z_i', X_i; \widetilde{\theta}, \widehat{\sigma}) - \nabla_\theta \nabla_{\theta^T} g(Z_i', X_i; \widetilde{\theta}, \sigma_0) \right\| = O_p(\|\widehat{\sigma} - \sigma_0\|_\infty). \tag{B.8}$$

Consequently, (B.4), (B.5) and (B.8) together indicate that the term $\|D_1\|$ can be bounded as

$$\|D_1\| = O_p(\|\widehat{\sigma} - \sigma_0\|_\infty) = o_p(1)$$

by Proposition 4.1. Hence, we can conclude by (B.3) that

$$\frac{1}{n} \sum_{i=1}^n \nabla_\theta s(W_i; \widetilde{\theta}, \widehat{\sigma}) \xrightarrow{p} \mathbb{E}[\nabla_\theta s(W_i; \theta_0, \sigma_0)].$$

□