Essays in Econometrics

Junlong Feng

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2020

ABSTRACT

Essays in Econometrics

Junlong Feng

My dissertation explores two broad areas in econometrics and statistics. The first area is nonparametric identification and estimation with endogeneity using instrumental variables. The second area is related to low-rank matrix recovery and high-dimensional panel data models. The following three chapters study different topics in these areas.

Chapter 1 considers identification and estimation of triangular models with a discrete endogenous variable and an instrumental variable (IV) taking on fewer values. Using standard approaches, the small support set of the IV leads to under-identification due to the failure of the order condition. This chapter develops the first approach to restore identification for both separable and nonseparable models in this case by supplementing the IV with covariates, allowed to enter the model in an arbitrary way. For the separable model, I show that it satisfies a system of linear equations, yielding a simple identification condition and a closed-form estimator. For the nonseparable model, I develop a new identification argument by exploiting its continuity and monotonicity, leading to weak sufficient conditions for global identification. Built on it, I propose a uniformly consistent and asymptotically normal sieve estimator. I apply my approach to an empirical application of the return to education with a binary IV. Though under-identified by the IV alone, I obtain results consistent with the empirical literature using my method. I also illustrate the applicability of the approach via an application of preschool program selection where the supplementation procedure fails.

Chapter 2, written with Jushan Bai, studies low-rank matrix recovery with a non-sparse error matrix. Sparsity or approximate sparsity is often imposed on the error

matrix for low-rank matrix recovery in statistics and machine learning literature. In econometrics, on the other hand, it is more common to impose a location normalization for the stochastic errors. This chapter sheds light on the deep connection between the median zero assumption and the sparsity-type assumptions by showing that the *principal component pursuit* method, a popular approach for low-rank matrix recovery by Candès et al. (2011), consistently estimates the low-rank component under a median zero assumption. The proof relies on a new theoretical argument showing that the median-zero error matrix can be decomposed into a matrix with a sufficient number of zeros and a non-sparse matrix with a small norm that controls the estimation error bound. As no restriction is imposed on the moments of the errors, the results apply to cases when the errors have heavy- or fat-tails.

In Chapter 3, I consider nuclear norm penalized quantile regression for large N and large T panel data models with interactive fixed effects. As the interactive fixed effects form a low-rank matrix, inspired by the median-zero interpretation, the estimator in this chapter extends the one studied in Chapter 2 by incorporating a conditional quantile restriction given covariates. The estimator solves a global convex minimization problem, not requiring pre-estimation of the (number of the) fixed effects. Uniform rates are obtained for both the slope coefficients and the low-rank common component of the interactive fixed effects. The rate of the latter is nearly optimal. To derive the rates, I show new results that establish uniform bounds of norms of certain random matrices of jump processes. The performance of the estimator is illustrated by Monte Carlo simulations.

Table of Contents

List of Tables

List of Figures

Acknowledgments

Over the years of my Ph.D. life, I have owed tremendously to my advisors Jushan Bai and Sokbae (Simon) Lee, for their advice, guidance and support. I am extremely lucky to learn how to be a good researcher and a good person from them. They taught me that details are as important as good ideas, that diligence is the path to meet high standards, and most importantly, that interest and passion are things that I should pursue. Proving a new lemma is fun, but the most unforgettable days in my Ph.D. life were those when I got stuck on a research question. The encouragement from Jushan on that summer night in his house, and the trust and comfort injected from the kind words that Simon said to me during our countless meetings, have been my most vivid memories of the six years.

I am incredibly thankful to my other committee members, Bernard Salanié, José Luis Montiel Olea (Pepe) and Bodhisattva Sen. Bernard's deep thinking in almost every field in economics has largely broadened my vision. Every time I talked with him, I discovered new angles to look at my own research. And thus, I learned how important being open-minded is. I also owe much to him for the huge amount of time he lent me. He read through many versions of the following chapters. His comments have helped improve the quality of them greatly. Pepe has become one of my role models among young econometricians. His enthusiasm and broad interest in research have shown me how fun being an econometrician could be. I benefited a lot from discussions with both Pepe and Bodhi. I appreciate much from the support and feedback they provided me.

I am deeply indebted to Serena Ng and Christoph Rothe. Serena gave me invaluable advice on both my research and being a researcher. It will continue to guide me through

my career. Christoph was my second-year advisor. Chapter 1 has been evolved from back then, and his comments and suggestions helped shape the original idea into what it is today.

I would also like to thank my classmates and friends who have made the six years warm and sweet. Special thanks to Lidan Bai, Yi Cheng, Chun-Che Chi, Zhihan Cui, Qi Dong, Jean-Jacques Forneron, Leonard Goff, Qi Guo, Jingying He, Zhuojun Huang, Chen Jiang, Yang Jiao, Mai Li, Xiaomao Li, Xuan Li, Yating Li, Yuexin Li, Jinyu Liu, Likun Liu, Zihan Lv, Lina Lu, Chengzhen Meng, Zhen Qu, Zhiling Sun, Yinong Tan, Dudian Tang, Shanjie Tang, Shuwei Tang, Nachuan Tian, Chunyan Wang, Fangfei Wang, Lijun (Leo) Wang, Mengxue Wang, Ruimin Wang, Jia Xiang, Yaxin Xiao, Yinxi Xie, Xiao Xu, Yue Yu, Shijia Zhang, Ye Zhang, Xiaofeng Zhou.

Finally, I would like to thank my family. It is impossible to finish my dissertation without the unconditional love and never-ceasing support from my parents and Lioney. And I thank my grandmother, who passed away during these six years. The stories she told me in my childhood illuminated my imagination and sparked my curiosity, which have led me here.

To my grandmother.

Preface

> *Science is built up with facts, as a house is with stones. But a collection of facts is no more a science than a heap of stones is a house.*
>
> – Henri Poincaré, *Hypotheses in Physics*, Tr. G.B. Halsted (1913)

Economic research has become data-intensive over the decades. Economists either hope to trace out causality between variables of interest, or to distill raw data into useful information for prediction. Yet without taming data with appropriate structures, data hardly speak for themselves. This thesis portrays different scenarios where under reasonable structures, desired information is extracted from data that would be otherwise deemed limited or noisy.

The first scenario I consider is the case where some dimension in the data is too limited to answer certain research questions. In Chapter 1, a researcher is assumed to have a multivalued endogenous variable but only a binary instrument. This is common in practice; a valid instrument often comes from a natural experiment or a policy shock, mechanically having a small support. But a researcher may be interested in multiple endogenous alternatives in its own right. The insufficiency of the instrument results in under-identification of the counterfactuals due to the failure of the classical order condition.

To resolve this issue, I impose structures on the selection mechanism to track which groups of individuals have the same level of endogeneity. One of the structures is called "propensity score coherence", meaning that individuals with identical propensity scores would make the same selection. This condition is robust to mis-specification of the se-

lection model, and is indirectly testable. Under this condition, patterns of substitution between the instrument and the covariates in terms of the selection decision can be identified. Consequently, one can identify groups of individuals, indexed by the realizations of the covariates, that have identical selection bias. In this way, to identify the counterfactuals of a given group, one can utilize not only information of the subgroups with different values of the instrument, but also information of a different group that has the same selection bias.

This chapter illustrates how reasonable structures can dramatically improve the information obtained from data. Without imposing any additional structures, the counterfactuals are only partially identified, and the identified set may be non-informative. For instance, if the endogenous variable takes on three values, the identified set is the entire solution set of two equations with three unknowns using a binary instrument, which could be too large to yield any useful information. However, under the conditions I impose, the counterfactuals could be point-identified, and these conditions are jointly testable.

The second scenario I consider is that data are sometimes noisy and conceals the information that a researcher is interested in. To remove the noises, again appropriate structures are called for. Chapter 2 and Chapter 3 study two problems in this scenario where noises arise in different sources.

In Chapter 2 (with Jushan Bai), we hope to recover a low-rank matrix from a high-dimensional panel data matrix. Low-rankness *per se* is a useful structure to alleviate the curse of dimensionality and/or scale. To consistently estimate the low-rank component, methods vary with structures imposed on the residual (or error) matrix. Traditional methods like principal component analysis assume that the data are not "too noisy" by requiring the existence of certain moments of the random errors. To deal with highly noisy data, this assumption or structure may be no longer plausible as the errors may be thought to be heavy- or fat-tailed. Meanwhile, methods in machine learning, such as

principal component pursuit (PCP), allow for large deterministic errors but require the error matrix to be sparse. Again, this structure may be not plausible in many applications in economics and finance where the errors are usually modeled as continuously distributed random variables.

In this chapter, the structure we impose is that the errors have zero median. We find that the median zero assumption is connected to sparsity in the sense that a random matrix with median zero entries can be decomposed into two matrices where one has small enough norm and the other has enough zeros for the argument under sparsity to go through. This decomposition is achieved by a novel theoretical tool we develop, named the Bernoulli device. Using this technique, we show that PCP consistently estimates the low-rank component with median zero errors under other regularity conditions.

Chapter 3 considers quantile regression with interactive fixed effects. Without these fixed effects, the slope coefficients can be obtained by standard quantile regression. Thus, the presence of the interactive fixed effects introduces noises and raises difficulties to utilize the data to extract the slope coefficients.

A useful structure often imposed in the literature is that the matrix of the interactive fixed effects is low-rank. This structure and the median zero restriction in Chapter 2 imply that PCP can be adapted to a quantile restriction to entertain the model in Chapter 3. I thus propose a nuclear norm regularized quantile regression estimator for panel data models with interactive fixed effects. I show that the estimator consistently estimates both the slope coefficients and the low-rank interactive-fixed effect matrix. I also provide the rates of convergence; the rate for the latter is nearly optimal.

Chapter 1

Matching Points: Supplementing Instruments with Covariates in Triangular Models

Junlong Feng[†]

[†]I thank Jushan Bai, Sokbae (Simon) Lee and Bernard Salanié, who were gracious with their advice, support and feedback. I have also greatly benefited from comments and discussions with Karun Adusumilli, Isaiah Andrews, Andres Aradillas-Lopez, Sandra Black, Ivan Canay, Xiaohong Chen, Leonard Goff, Florian Gunsilius, Han Hong, Jessie Li, José Luis Montiel Olea, Ulrich Müller, Whitney Newey, Serena Ng, Christoph Rothe, Jörg Stoye, Matt Taddy, Alexander Torgotivtsky, Quong Vuong, Yulong Wang, Kaspar Wuthrich and participants of the Columbia Econometrics Colloquium and Workshop as well as the participants of the seminar at the 2019 Econometrics Society Asian Meeting in Xiamen. I also thank Research Connections for providing the data of the Head Start Impact Study.

1.1 Introduction

This paper considers identification and estimation of the outcome function $\boldsymbol{g}^* \equiv (g_d^*)_d$ in a triangular model:

$$Y = \sum_d \mathbb{1}(D = d) \cdot g_d^*(\boldsymbol{X}, U)$$

$$D = h(\boldsymbol{X}, Z, \boldsymbol{V})$$

where both the endogenous variable D and the instrumental variable (IV) Z are discrete, $\boldsymbol{X}$ is a vector of covariates, and the disturbances U and $\boldsymbol{V}$ are correlated (see also Newey, Powell and Vella (1999), Chesher (2003), Matzkin (2003), Newey and Powell (2003), Chernozhukov and Hansen (2005), Das (2005), Imbens and Newey (2009), etc.)

It is well-known that in general, $\boldsymbol{g}^*$ is not identified if Z takes on fewer values than D does. In many applications, however, IVs have very small support while endogenous variables may take on more values.

Let us consider an example of the returns to education. Suppose the log wage (Y) is determined by unobserved earning ability U and functions of covariates ($\boldsymbol{X}$) such as parents' education. These functions are heterogeneous in the level of education d: completing high school ($d = 1$), having some college education ($d = 2$), and at least completing college ($d = 3$). The available IV may be only binary. For instance, in Card (1995), Z indicates whether an individual lived near a 4-year college or not.

To see why identification may fail, suppose the unknown function is separable in ability: $g_d^*(\boldsymbol{X}, U) = m_d^*(\boldsymbol{X}) + U$. Then the model can be rewritten as $Y = \alpha(\boldsymbol{X}) + m_2^*(\boldsymbol{X})\mathbb{1}(D = 2) + m_3^*(\boldsymbol{X})\mathbb{1}(D = 3) + U$. The classical order condition thus does not hold: conditional on $\boldsymbol{X}$, there are two endogenous variables, $\mathbb{1}(D = 2)$ and $\mathbb{1}(D = 3)$, but only one binary IV.

Under the standard validity assumptions for the IV, it can be shown that the outcome function $\boldsymbol{m}^*$ satisfies the moment condition $\sum_{d=1}^3 p_d(\boldsymbol{x}_0, Z) m_d^*(\boldsymbol{x}_0) = \mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x}_0, Z)$ for some $\boldsymbol{x}_0$ where $p_d(\cdot, \cdot)$ is the generalized propensity score (e.g. Newey and Powell

(2003)). With a binary Z, we obtain two equations by conditioning on each value that Z can take, but there are three unknowns. To the best of the author's knowledge, no existing method achieves point-identification in such a case.

This paper develops the first approach that obtains point-identification of g^* when the IV takes on fewer values than the discrete endogenous variable. This is achieved by supplementing the IV with variation in $\boldsymbol{X}$. We show that for a fixed $\boldsymbol{x}_0$, there may exist a *matching point* $\boldsymbol{x}_m$ such that the difference between $g^*(\boldsymbol{x}_0, \cdot)$ and $g^*(\boldsymbol{x}_m, \cdot)$ is identified. Controlling for the difference, moment equations like the example above can be evaluated at $\boldsymbol{x}_m$ in addition to $\boldsymbol{x}_0$ without introducing new unknowns. In this way, the effective support set of the IV is enlarged via the matching points, making identification possible.

To see why such outcome function differences may be identified before the outcome functions themselves are, note that in the triangular model, endogeneity is generally due to the dependency between U and $\boldsymbol{V}$. Suppose $\boldsymbol{X}$ and Z generate partitions in the space of $\boldsymbol{V}$ (for instance in an ordered choice model). Selecting into a value of D is determined by which partition $\boldsymbol{V}$ falls into. Hence, if for some $z \neq z'$, $(\boldsymbol{x}_0, z)$ and $(\boldsymbol{x}_m, z')$ generate exactly the same partitions, then the same selection choices would be made across the two schemes for any realization of the latent $\boldsymbol{V}$. The unknown selection biases at these two points would thus be equal and can be differenced out. The relationship between $g^*(\boldsymbol{x}_0, \cdot)$ and $g^*(\boldsymbol{x}_m, \cdot)$ can thus be traced out from the distribution of the observed outcomes at $(\boldsymbol{x}_0, z)$ and $(\boldsymbol{x}_m, z')$.

Let us go back to the return to education example. Suppose we have a single covariate X, the average of parents' years of schooling. Let X and Z enter h via a linear single index $kZ + X$. It implies that k more years of the parents' schooling compensate for not living near a 4-year college ($Z = 0$) in terms of educational attainment choices. For any realization of $\boldsymbol{V}$, individuals with $(X, Z) = (x_0, 0)$ and with $(X, Z) = (x_0 - k, 1)$ would select into the same level of education. These two groups of individuals are equivalent

in terms of selection, so their selection biases are presumably the same. Comparing their (average or distributions of) observed log wages, the biases may be differenced out.

To find the matching points of a given x_0, we do not restrict the dimension of V and no notion of monotonicity is imposed. We propose a condition called *propensity score coherence*. Under it, the matching points can be found by matching the generalized propensity scores at different values of X and Z without specifying the selection model h. We provide examples to illustrate that many widely used discrete choice models satisfy this condition.

Given the matching points, we derive the exact forms of the outcome function differences for two particular models of g^*: additively separable in U, and nonseparable and strictly increasing in U. For each model, we provide sufficient conditions for identification and construct consistent and asymptotically normal estimators.

For the separable model, we show that the outcome function solves a system of linear equations, preserving a similar structure as in the standard IV approach. We thus obtain a closed-form estimator which is easy to implement in practice. We apply it to examine the return to education example using the same extract from 1979 National Longitudinal Surveys (NLS) as in Card (1995). We adopt the proximity-to-college IV and find that living near a four-year college and parents' years of schooling are indeed close substitutes in terms of children's educational attainment. Using the matching points generated from this covariate, we find that the return to education is (a) increasing in the level of schooling with slightly diminishing marginal return, and (b) heterogeneous in parents' education; individuals with less educated parents enjoy higher potential returns. In contrast, the two-stage-least-squares (2SLS) estimates of parametric models using the interaction of parents' years of schooling and the proximity-to-college IV as an extra instrument lead to misleading results.

For the nonseparable model, we develop a new identification argument by exploiting continuity and monotonicity of $g^*(X, \cdot)$. We show that global identification of $g^*(X, \cdot)$

is achieved in the space of monotonic functions if $g^*(\boldsymbol{X}, u)$ is only locally identified for each u. This new result also applies to the standard IV approach when the IV has large support. Based on our identification strategy, we construct a sieve estimator. We show that its large sample properties are guaranteed by simple low-level conditions, thanks to the nice properties of the monotonic function space.

It is worth noting that the success of our approach hinges on the covariates that are able to offset the impact of Z. For applications where the IV has the dominant effect, covariates may not have comparable effects on the selection, and matching points may not exist. This is testable in some cases. As an illustration, we consider another empirical application on the preschool program selection. We use the administrated Head Start Impact Study (HSIS) dataset following Kline and Walters (2016). The endogenous variable considered also takes on three values: participating in Head Start, in an alternative preschool program, and not participating in any programs. The binary IV indicates whether an individual won a lottery granting access to Head Start. The IV has a very large effect on the choice of preschool programs. From the tests we develop, we find that no available covariate in the sample is able to generate a matching point.

We defer a detailed comparison of our method to the existing literature until Section 1.8. Here we highlight some major differences. Precursory methods that circumvent the problem of having an IV with small support include imposing homogeneity between adjacent levels of D when D is ordered, or specifying a parametric form for g^* and using interactions between Z and $\boldsymbol{X}$ as a second IV by assuming $\boldsymbol{X}$ is exogenous. Torgovitsky (2015, 2017) and D'Haultfœuille and Février (2015) show a binary IV is able to identify nonseparable models with a continuous endogenous variable. Continuity is crucial in their approach and they require the selection function strictly increasing in the scalar unobservable. Similar to this paper, Caetano and Escanciano (2018) also use covariates to identify models when the instruments do not have enough variation. Their approach does not rely on the first stage, but they need the covariates used for identification

purpose to be "separable" in the model in a way that the model can be "inverted" and become free of them. In contrast, the covariates in our approach can enter the model in an arbitrary way. Huang, Khalil and Yıldız (2019) consider identification of separable models with multiple endogenous variable but a single instrument. They focus on a partial linear model, and one of the endogenous variable needs to be continuous to apply their control function technique. Ichimura and Taber (2000) and Vytlacil and Yıldız (2007) use shifts in some observables that compensate for a shift in a target variable to facilitate identification of different parameters than this paper. The shifting variables and the target variable are different from ours. Vuong and Xu (2017) and Feng, Vuong and Xu (2020) study the individual treatment effect of a binary D and develop a concept called the counterfactual mapping. It is also an identifiable function linking two outcome functions but at *different* values of D and the *same* value of X by exploiting the compliers' information.

The rest of the paper is organized as follows. In Section 1.2, we introduce the model, discuss the preliminary assumptions, and introduce the matching points. We also preview the basic idea of the new identification strategy. In Section 1.3, we discuss the existence of the matching points and provide sufficient conditions for identification of the matching points and the outcome functions. In Section 1.4, we propose estimators for them and discuss some implementation issues. Section 1.5 presents results of two empirical applications. Section 1.6 shows the estimators' asymptotic properties. Section 1.7 provides Monte Carlo simulations to illustrate the estimator's finite sample performance. Section 1.8 discusses the relation of our approach to the related work. Section 1.9 concludes. Appendix A.1 discusses general cases including models with multiple discrete endogenous variables. Appendix A.2 contains proofs of the results in Sections 1.2 and 1.3. Appendix A.3 illustrates the propensity coherence condition via various discrete choice models for a single or multiple endogenous variables. In the supplementary appendices, Appendix A.4 provides additional simulation results, and Appendix

A.5 collects proofs of the asymptotic results.

Notation

We use upper-case Latin letters for random variables and the corresponding lower-cases for their realizations. Bold Latin letters denote vectors or matrices. For two generic random variables A and B, denote the conditional expectation of A given $B = b$ by $\mathbb{E}_{A|B}(b)$, with similar notation for conditional distribution functions, densities and variances. Denote the support set of A by $S(A)$, and the support of A given $B = b$ by $S(A|B = b)$, or simply $S(A|b)$ when it does not cause confusion. For a finite set H, $|H|$ denotes the number of elements in it, while for a generic vector $\boldsymbol{c}$, $|\boldsymbol{c}|$ denotes its Euclidean norm. For two generic sets H_1 and H_2, $H_1 \backslash H_2$ denotes the set difference $H_1 \cap H_2^c$. Throughout, we assume all the random variables involved are in a common probability space with the measure function $\mathcal{P}$. Whenever we say almost surely (a.s.) and measurable, we refer to almost surely and measurable with respect to $\mathcal{P}$.

1.2 The Model

To highlight the key features of our approach, we focus on a simple case where the endogenous D takes on three values ($|S(D)| = 3$) and Z is binary ($|S(Z)| = 2$). We will also discuss the usefulness of the approach when $|S(D)| = |S(Z)| = 2$. The general cases for arbitrary $|S(D)| \geq |S(Z)|$ and multiple Ds will be discussed in Appendix A.1.1.

We study the separable model and the nonseparable model respectively:

$$Y = \sum_{d \in S(D)} \mathbb{1}(D = d) \cdot \left(m_d^*(\boldsymbol{X}) + U \right) \tag{SP}$$

and

$$Y = \sum_{d \in S(D)} \mathbb{1}(D = d) \cdot g_d^*(\boldsymbol{X}, U) \tag{NSP}$$

where $S(D) \equiv \{1,2,3\}$, $\boldsymbol{X}$ is a vector of covariates, and U is a scalar unobservable. The goal of this paper is to identify and estimate the outcome functions at a fixed value of $\boldsymbol{X}$: $\boldsymbol{m}^*(\boldsymbol{x}_0) \equiv (m_d^*(\boldsymbol{x}_0))_d$ and $\boldsymbol{g}^*(\boldsymbol{x}_0,\cdot) \equiv (g_d^*(\boldsymbol{x}_0,\cdot))_d$. Note the choice of $S(D)$ is without loss of generality because any set of three element can be one-to-one mapped onto it.

We rewrite the selection model for D as follows:

$$D = d \text{ if and only if } h_d(\boldsymbol{X}, Z, \boldsymbol{V}) = 1 \tag{SL}$$

where for all $d \in S_D$, the selection function $h_d(\boldsymbol{X}, Z, \boldsymbol{V}) \in \{0,1\}$, and $\sum_{d=1}^{3} h_d(\boldsymbol{X}, Z, \boldsymbol{V}) = 1$ a.s. $\boldsymbol{V}$ is a vector of unobservables that is correlated with U. We assume that for every $(\boldsymbol{x}, z) \in S(\boldsymbol{X}, Z)$, $h_d(\boldsymbol{x}, z, \cdot)$ is measurable on $S(\boldsymbol{V})$.

In the rest of this section, we introduce and discuss preliminary assumptions for each model. We also illustrate why a binary Z in general fails to identify the outcome functions. Finally, we introduce the key idea to restore identification.

1.2.1 The Separable Model

Let us begin with the assumption for the separable model-SP:

Assumption E-SP (Exogeneity). $\mathbb{E}_{U|\boldsymbol{X}}(\boldsymbol{x}_0) = 0$, $\mathbb{E}_{U|\boldsymbol{V}\boldsymbol{X}Z}(\boldsymbol{V}, \boldsymbol{x}_0, Z) = \mathbb{E}_{U|\boldsymbol{V}\boldsymbol{X}}(\boldsymbol{V}, \boldsymbol{x}_0)$ *a.s., and* $Z \perp\!\!\!\perp \boldsymbol{V} | \boldsymbol{X} = \boldsymbol{x}_0$.

The first condition in Assumption E-SP is a normalization without which $\boldsymbol{m}^*(\boldsymbol{x}_0)$ can only be identified up to an additive constant. The second and the third conditions are standard in the literature on triangular models (e.g. Newey, Powell and Vella (1999)).

Remark 2.1. *Note that the unobservable U is not d-dependent, so our model is more restrictive than many models in the treatment effects literature. For example local average treatment effect (e.g. Imbens and Angrist (1994) and Angrist and Imbens (1995)), marginal treatment effect (e.g.Heckman and Vytlacil (2005) and Heckman, Urzua and Vytlacil (2008)), and effects of multivalued treatment (Lee and Salanié, 2018). However, these works focus on different parameters*

and/or need much richer variation in the instruments (continuous and multidimensional). In our setup, we can allow U to be d-dependent by making extra assumptions. For example, we can show that our results still hold if $\mathbb{E}_{U_d|DXZ}(D, \boldsymbol{x}_0, Z) = \mathbb{E}_{U_{d'}|DXZ}(D, \boldsymbol{x}_0, Z)$ for any $d \neq d'$. This assumption allows U_d for each d to have different conditional distributions so long as they have the same mean dependence of the endogenous variable.

Proposition 1 (Newey and Powell (2003) equation (2.2); Das (2005) equation (2.5))**.** *Under Assumptions E-SP, the following equation holds for all $z \in S(Z)$,*

$$\sum_{d=1}^{3} p_d(\boldsymbol{x}_0, z) \cdot m_d^*(\boldsymbol{x}_0) = \sum_{d=1}^{3} p_d(\boldsymbol{x}_0, z) \cdot \mathbb{E}_{Y|DXZ}(d, \boldsymbol{x}_0, z) \tag{1.2.1}$$

where $p_d(\boldsymbol{x}_0, z) \equiv \mathbb{P}(D = d | \boldsymbol{X} = \boldsymbol{x}_0, Z = z)$.

Since all the terms in equation (1.2.1) are directly identified from the population except for $\boldsymbol{m}^*(\boldsymbol{x}_0)$, we have two linear equations by letting $z = 0, 1$ but three unknowns: $\boldsymbol{m}^*(\boldsymbol{x}_0)$ is not identified without additional information.

1.2.2 The Nonseparable Model

Compared to the separable model, assumptions for the nonseparable model-NSP are more stringent.

Assumption E-NSP (Exogeneity)**.** $(U|\boldsymbol{X} = \boldsymbol{x}_0) \sim Unif[0, 1]$ *and* $(U, \boldsymbol{V}) \perp\!\!\!\perp Z|\boldsymbol{X} = \boldsymbol{x}_0$.

Assumption FS (Full Support)**.** $(U, \boldsymbol{V})|\boldsymbol{x}_0$ *is continuously distributed and* $S(U|\boldsymbol{V}, \boldsymbol{x}_0) = S(U|\boldsymbol{x}_0)$.

Assumption CM (Continuity and Monotonicity)**.** *For all $(d, \boldsymbol{x}) \in S(D, \boldsymbol{X})$, $g_d^*(\boldsymbol{x}, \cdot)$ is continuous and strictly increasing on $[0, 1]$.*

Assumption E-NSP is the counterpart of Assumption E-SP for nonseparable outcome functions; the first part is a popular normalization for identification of nonseparable

models, while the second part is the same as in Imbens and Newey (2009) which is standard for triangular models. Similar to Model-SP, U is invariant with respect to d. We can relax it by adopting the *rank similarity* condition in Chernozhukov and Hansen (2005).

Assumption FS guarantees that the range of $g_d^*(\boldsymbol{x}, z, \cdot)$ on $[0,1]$ is equal to the conditional support $S(Y|d,\boldsymbol{x})$[1]. The same assumption can be found in related work that also focuses on identification of $\boldsymbol{g}^*(\boldsymbol{x}_0, \cdot)$ on the entire domain as this paper, for instance D'Haultfœuille and Février (2015), Torgovitsky (2015) and Vuong and Xu (2017).

Assumption CM regulates the behavior of the NSP-outcome function $\boldsymbol{g}^*(\boldsymbol{x}, \cdot)$. Continuity on $[0,1]$ and Assumption FS (a) imply that $Y|d,\boldsymbol{x},z$ is continuously distributed and that $S(Y|d,\boldsymbol{x},z)$ is compact. Continuity and strict monotonicity are two standard requirements in the literature of nonseparable models when the unobservable is a scalar. In addition to using these properties to construct moment conditions as in the related literature, in this paper we show that they deliver nice results for identification and for deriving the large sample properties of the estimator we propose.

Under these assumptions, we have the following result:

Proposition 2 (Chernozhukov and Hansen (2005), Theorem 1)**.** *Under Assumptions E-NSP, FS and CM, the following equation holds for all* $z \in \{0,1\}$ *and* $u \in [0,1]$*,*

$$\sum_{d=1}^{3} p_d(\boldsymbol{x}_0, z) \cdot F_{Y|D\boldsymbol{X}Z}(g_d^*(\boldsymbol{x}_0, u)|d, \boldsymbol{x}_0, z) = u \tag{1.2.2}$$

Similar to Model-SP, again we have two equations but three unknowns for a fixed u. Global uniqueness of the solution is in general not guaranteed.

[1]This is because $S(Y|d,\boldsymbol{x}) = S(g_d^*(\boldsymbol{x},U)|h_d(\boldsymbol{x},Z,\boldsymbol{V}) = 1,\boldsymbol{x}) = S(g_d^*(\boldsymbol{x},U)|h_d(\boldsymbol{x},z,\boldsymbol{V}) = 1,\boldsymbol{x},z) = S(g_d^*(\boldsymbol{x},U)|h_d(\boldsymbol{x},z,\boldsymbol{V}) = 1,\boldsymbol{x}) = S(g_d^*(\boldsymbol{x},U))$, which is the range of $g_d^*(\boldsymbol{x},\cdot)$ on [0,1].

1.2.3 The Selection Model and the Matching Points

Now we show how to use covariates $\boldsymbol{X}$ to supplement the binary Z to restore identification when the order condition fails.

The major challenge for identification by varying the conditioning value of $\boldsymbol{X}$ is that it results in unknown changes in the outcome function: while more moment conditions are generated, even more unknowns are introduced into the new system of equations. For instance, consider equation (1.2.1) for Model-SP. Suppose Assumptions E-SP also holds for $\boldsymbol{x}' \neq \boldsymbol{x}_0$. Similar to (1.2.1), we have

$$\sum_{d=1}^{3} p_d(\boldsymbol{x}', z) \cdot m_d^*(\boldsymbol{x}') = \sum_{d=1}^{3} p_d(\boldsymbol{x}', z) \cdot \mathbb{E}_{Y|D\boldsymbol{X}Z}(d, \boldsymbol{x}', z)$$

for $z \in \{0, 1\}$. We then have 4 equations in total: two conditional on $\boldsymbol{X} = \boldsymbol{x}_0$ and two on $\boldsymbol{x}'$, yet the number of the unknowns is increased to 6 at the same time. Therefore, an arbitrarily chosen $\boldsymbol{x}'$ does not help identification.

Instead, we look for a point, denoted by $\boldsymbol{x}_m$, such that the difference between $\boldsymbol{m}^*(\boldsymbol{x}_0)$ and $\boldsymbol{m}^*(\boldsymbol{x}_m)$ can be identified first. To see this is possible, let us take conditional expectation on both sides of equation (SP). For all $d, \boldsymbol{x}, z$ in their support,

$$m_d^*(\boldsymbol{x}) + \mathbb{E}_{U|D\boldsymbol{X}Z}(d, \boldsymbol{x}, z) = \mathbb{E}_{Y|D\boldsymbol{X}Z}(d, \boldsymbol{x}, z).$$

The term $\mathbb{E}_{U|D\boldsymbol{X}Z}(d, \boldsymbol{x}, z)$ captures the selection bias due to endogeneity of D. Then for $\boldsymbol{x}_m \neq \boldsymbol{x}_0$, the difference between $m_d^*(\boldsymbol{x}_m)$ and $m_d^*(\boldsymbol{x}_0)$ satisfies:

$$\begin{aligned} m_d^*(\boldsymbol{x}_m) - m_d^*(\boldsymbol{x}_0) = & \underbrace{\left(\mathbb{E}_{Y|D\boldsymbol{X}Z}(d, \boldsymbol{x}_m, z') - \mathbb{E}_{Y|D\boldsymbol{X}Z}(d, \boldsymbol{x}_0, z)\right)}_{\text{Difference in the Observed Outcomes}} \\ & - \underbrace{\left(\mathbb{E}_{U|D\boldsymbol{X}Z}(d, \boldsymbol{x}_m, z') - \mathbb{E}_{U|D\boldsymbol{X}Z}(d, \boldsymbol{x}_0, z)\right)}_{\text{Difference in the Biases}} \end{aligned}$$

When the two unknown bias terms are equal, they cancel out and the change in the outcome function is identified.

Let us consider the following example for illustration.

Example OC (Ordered Choice)**.** *Suppose D is ordered and there is only one covariate. Let $h_1(X,Z,V) = \mathbb{1}(V < \kappa_1 + \beta X + \alpha Z)$, $h_3(X,Z,V) = \mathbb{1}(V \geq \kappa_2 + \beta X + \alpha Z)$, and $h_2 = 1 - h_1 - h_3$. Assume $\alpha \cdot \beta \neq 0$, $\kappa_1 < \kappa_2$, and $(X,Z) \perp\!\!\!\perp V$ where V is continuously distributed on $\mathbb{R}$. Fix x_0, it is straightforward to see that $(x_0, 0)$ and $(x_0 - \frac{\alpha}{\beta}, 1)$ generate exactly the same partitions on $\mathbb{R}$. Then taking $d = 1$ as an example, we have*

$$\mathbb{E}_{U|DXZ}(1, x_0 - \frac{\alpha}{\beta}, 1) = \mathbb{E}_{U|VXZ}(V < \kappa_1 + \beta x_0, x_0 - \frac{\alpha}{\beta}, 1)$$
$$\mathbb{E}_{U|DXZ}(1, x_0, 0) = \mathbb{E}_{U|VXZ}(V < \kappa_1 + \beta x_0, x_0, 0)$$

When the dependency of (U,V) on $(X,Z) = (x_0, 0)$ and $(x_0 - \frac{\alpha}{\beta}, 1)$ are identical, the two bias terms are equal.

From the example, we can see that in order to difference out the bias, $(\boldsymbol{x}_m, z')$ and $(\boldsymbol{x}_0, z)$ should (a) generate the same partitions of $S(\boldsymbol{V})$, and (b) have the same level of dependency with respect to the unobservables. The following conditions formally characterize these ideas.

Definition MP (Matching Points and Matching Pairs)**.** *A point $\boldsymbol{x}_m \in S(\boldsymbol{X})$ is a matching point of $\boldsymbol{x}_0 \in S(\boldsymbol{X})$ if there exist $z \neq z' \in S(Z)$ such that for all $d \in S(D)$,*

$$h_d(\boldsymbol{x}_0, z, \boldsymbol{V}) = h_d(\boldsymbol{x}_m, z', \boldsymbol{V}) \text{ a.s.}, \tag{1.2.3}$$

and for Model-SP,

$$\mathbb{E}_{U|VXZ}(\boldsymbol{V}, \boldsymbol{x}_m, Z) = \mathbb{E}_{U|VXZ}(\boldsymbol{V}, \boldsymbol{x}_0, Z) \text{ a.s. and } (\boldsymbol{V}|\boldsymbol{x}_m, Z) \sim (\boldsymbol{V}|\boldsymbol{x}_0, Z), \tag{1.2.4}$$

or for Model-NSP,

$$((U, \boldsymbol{V})|\boldsymbol{x}_m, Z) \sim ((U, \boldsymbol{V})|\boldsymbol{x}_0, Z). \tag{1.2.5}$$

$(\boldsymbol{x}_0, z)$ *and* $(\boldsymbol{x}_m, z')$ *are called a matching pair.*

Equation (1.2.3) guarantees that the matching pair generate exactly the same partitions on $S(\boldsymbol{V})$. Equation (1.2.4) and (1.2.5) imply that U and $\boldsymbol{V}$ have the same level of dependence given $\boldsymbol{X} = \boldsymbol{x}_0$ or $\boldsymbol{x}_m$. A sufficient condition for these two equations is that $(Z, \boldsymbol{X})$ are jointly exogenous. This assumption is commonly made in practice, for instance Carneiro, Heckman and Vytlacil (2011). Also, it only needs to be satisfied by the covariates that are used to generate the matching points. Finally, results in the following theorem are testable implications for these conditions; after the estimates are obtained, an over-identification type test can be performed to see whether the null that all the conditions hold is true.

From Definition MP, it can be verified that if Assumptions E-SP or Assumptions E-NSP and FS hold at $\boldsymbol{x}_0$, they also hold at the matching points of $\boldsymbol{x}_0$. The following theorem thus shows that the changes in the outcome functions from $\boldsymbol{x}_0$ to a matching point are identified:

Theorem MEQ (Matching Equation)**.** *Suppose* $\boldsymbol{x}_m \in S(\boldsymbol{X})$ *is a matching point for* $\boldsymbol{x}_0 \in S(\boldsymbol{X})$*, then the following claims hold for all* $d \in S(D)$*:*

(a) Model-SP. Under Assumptions E-SP, $p_d(\boldsymbol{x}_m, z') = p_d(\boldsymbol{x}_0, z)$ *and*

$$m_d^*(\boldsymbol{x}_m) = m_d^*(\boldsymbol{x}_0) + \left(\mathbb{E}_{Y|D\boldsymbol{X}Z}(d, \boldsymbol{x}_m, z') - \mathbb{E}_{Y|D\boldsymbol{X}Z}(d, \boldsymbol{x}_0, z)\right). \tag{1.2.6}$$

(b) Model-NSP. Under Assumptions E-NSP, FS and CM, $p_d(\boldsymbol{x}_m, z') = p_d(\boldsymbol{x}_0, z)$ *and*

$$F_{Y|D\boldsymbol{X}Z}(g_d^*(\boldsymbol{x}_m, u)|d, \boldsymbol{x}_m, z') = F_{Y|D\boldsymbol{X}Z}(g_d^*(\boldsymbol{x}_0, u)|d, \boldsymbol{x}_0, z). \tag{1.2.7}$$

The matching equation (1.2.6) directly establishes an identified one-to-one mapping

from $\boldsymbol{m}^*(\boldsymbol{x}_0)$ to $\boldsymbol{m}^*(\boldsymbol{x}_m)$.

For the matching equation (1.2.7), by strict motononicity of the conditional CDFs of Y (implied by Assumptions FS and CM), we have

$$g_d^*(\boldsymbol{x}_m, u) = Q_{Y|DXZ}\big(F_{Y|DXZ}(g_d^*(\boldsymbol{x}_0, u)|d, \boldsymbol{x}_0, z)\big|d, \boldsymbol{x}_m, z'\big) \equiv \varphi_d(g_d^*(\boldsymbol{x}_0, u); \boldsymbol{x}_m, z') \quad (1.2.8)$$

where $\varphi_d(\cdot; \boldsymbol{x}_m, z') : S(Y|d, \boldsymbol{x}_0) \mapsto S(Y|d, \boldsymbol{x}_m)$ is continuous and strictly increasing. Later we may use the shorthand notation $\varphi_d(\cdot)$ for brevity.

Theorem MEQ allows us to condition on $\boldsymbol{X} = \boldsymbol{x}_m$ to help identify $\boldsymbol{m}^*(\boldsymbol{x}_0)$ and $\boldsymbol{g}^*(\boldsymbol{x}_0, u)$. We use Model-SP as an example for illustration. By Proposition 2,

$$\sum_{d=1}^{3} p_d(\boldsymbol{x}_0, z) \cdot m_d^*(\boldsymbol{x}_0) = \sum_{d=1}^{3} p_d(\boldsymbol{x}_0, z) \cdot \mathbb{E}_{Y|DXZ}(d, \boldsymbol{x}_0, z) \quad (1.2.9)$$

$$\sum_{d=1}^{3} p_d(\boldsymbol{x}_0, z') \cdot m_d^*(\boldsymbol{x}_0) = \sum_{d=1}^{3} p_d(\boldsymbol{x}_0, z') \cdot \mathbb{E}_{Y|DXZ}(d, \boldsymbol{x}_0, z') \quad (1.2.10)$$

$$\sum_{d=1}^{3} p_d(\boldsymbol{x}_m, z) \cdot m_d^*(\boldsymbol{x}_m) = \sum_{d=1}^{3} p_d(\boldsymbol{x}_m, z) \cdot \mathbb{E}_{Y|DXZ}(d, \boldsymbol{x}_m, z) \quad (1.2.11)$$

$$\sum_{d=1}^{3} p_d(\boldsymbol{x}_m, z') \cdot m_d^*(\boldsymbol{x}_m) = \sum_{d=1}^{3} p_d(\boldsymbol{x}_m, z') \cdot \mathbb{E}_{Y|DXZ}(d, \boldsymbol{x}_m, z') \quad (1.2.12)$$

Substitute equation (1.2.6) into equations (1.2.11) and (1.2.12) for all d. Then (1.2.12) is redundant with (1.2.9) as they become identical. With the extra equation (1.2.11), we end up with three equations and three unknowns; identification becomes possible.

Further, the augmentation of the moment conditions does not necessarily end here. Given $\boldsymbol{x}_m$, one would expect that if it also has a matching point $\boldsymbol{x}_m' \neq \boldsymbol{x}_0$, then the mapping between the outcome functions at $\boldsymbol{x}_m'$ and $\boldsymbol{x}_m$ is identified. Consequently, the mapping between those at $\boldsymbol{x}_m'$ and $\boldsymbol{x}_0$ is identified, too. The following example illustrates this possibility.

Example OC Cont'd. *Under the setup in Example OC, for any fixed $x_0 \in S(X)$, it has the*

following two matching points by equation (1.2.3) if they are in $S(X)$:

$$(z=0,z'=1):\beta x_{m1}+\alpha\cdot 1=\beta x_0+\alpha\cdot 0 \implies x_{m1}=x_0-\frac{\alpha}{\beta} \tag{1.2.13}$$

$$(z=1,z'=0):\beta x_{m2}+\alpha\cdot 0=\beta x_0+\alpha\cdot 1 \implies x_{m2}=x_0+\frac{\alpha}{\beta} \tag{1.2.14}$$

Similarly, each of x_{m1} and x_{m2} also has two matching points: One is x_0, and the other is $x_0-2\frac{\alpha}{\beta}$ and $x_0+2\frac{\alpha}{\beta}$ respectively. This process can be continued until the boundaries of $S(X)$ are reached, illustrated by the following figure:

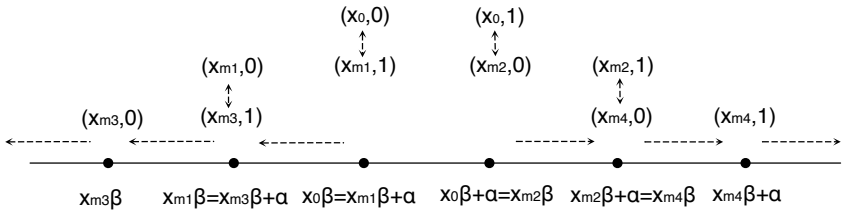

Figure 1.1: The Pyramid of Matching Points

The horizontal axis is the value of the single index $x\beta+z\alpha$. Starting from $(x_0,0)$ and $(x_0,1)$, we obtain x_{m1} and x_{m2} by solving the equations below the horizontal axis. Then we repeat this procedure to match $(x_{m1},0)$ with $(x_{m3},1)$ and match $(x_{m2},1)$ with $(x_{m4},0)$. Continuing the process, one can expect to see that the dotted points on this axis extend to both directions, until they reach the boundaries of $S(X)$.

These additional points in Example OC Cont'd are not the matching points of $\boldsymbol{x}_0$. But by recursively applying Theorem MEQ, the outcome functions at these points and at $\boldsymbol{x}_0$ are still linked by identified one-to-one mappings. To formalize this idea, we introduce the following concepts.

Definition MC (M-Connected Set). *A set $\mathcal{X}_{MC}(\boldsymbol{x}_0)\subseteq S(\boldsymbol{X})$ is called the m-connected set of $\boldsymbol{x}_0$ if $\boldsymbol{x}_0\in\mathcal{X}_{MC}(\boldsymbol{x}_0)$ and for any $\boldsymbol{x}\in\mathcal{X}_{MC}(\boldsymbol{x}_0)$, there exists $\boldsymbol{x}_1,\boldsymbol{x}_2,...,\boldsymbol{x}_{k(x)}\in\mathcal{X}_{MC}(\boldsymbol{x}_0)$ such*

that $\boldsymbol{x}_j$ is a matching point of $\boldsymbol{x}_{j-1}$, $j = 1, ..., k(\boldsymbol{x})$, and $\boldsymbol{x}$ is a matching point of $\boldsymbol{x}_{k(\boldsymbol{x})}$. Any two points in the m-connected set are said to be m-connected.

By definition, the m-connected set is the largest subset of $S(\boldsymbol{X})$ such that the outcome functions' relationship at any two elements in it is identified by recursively applying Theorem MEQ. Coupled with $S(Z)$, the set $\mathcal{Z}(\boldsymbol{x}_0) \equiv \mathcal{X}_{MC}(\boldsymbol{x}_0) \times S(Z)$ contains every possible value of $(\boldsymbol{X}, Z)$ that may be conditioned on to identify $\boldsymbol{m}^*(\boldsymbol{x}_0)$ or $\boldsymbol{g}^*(\boldsymbol{x}_0, \cdot)$.

1.3 Identification

In this section we first discuss the existence and identification of the matching points of a given $\boldsymbol{x}_0$. A similar argument holds for other points in $\boldsymbol{x}_0$'s m-connected set. Then we provide sufficient conditions under which the SP- and the NSP-outcome functions are identified.

1.3.1 The Existence and Identification of the Matching Points

The existence and identification of the matching points are closely related to the selection function $\boldsymbol{h}$ and features of $\boldsymbol{X}$, for example its dimensionality and support. Different ways to find them are available depending on how much we know about $\boldsymbol{h}$,

When the form of $\boldsymbol{h}$ or some of its structures are known, the matching points $\boldsymbol{x}_m$ may be obtained by directly applying the definition: $h_d(\boldsymbol{x}_m, z', \boldsymbol{v}) - h_d(\boldsymbol{x}_0, z, \boldsymbol{v}) = 0$ for all $\boldsymbol{v} \in S(\boldsymbol{V})$. For instance, in Example OC Cont'd, we know D is determined by a single-index ordered choice model. Then the matching points are obtained via equations (1.2.13) and (1.2.14) when α and β are identified (up to a multiplicative constant).

When the model that determines D is unknown, as is common in many economic applications, solving equation (1.2.3) is infeasible. On the other hand, the generalized propensity scores are usually directly identified from the population. Under the exogeneity assumption for Z, a matching point $\boldsymbol{x}_m$ necessarily solves the following equation

for $z' \neq z \in S(Z)$:

$$\big(p_1(\boldsymbol{x}, z') - p_1(\boldsymbol{x}_0, z)\big)^2 + \big(p_2(\boldsymbol{x}, z') - p_2(\boldsymbol{x}_0, z)\big)^2 = 0 \tag{1.3.1}$$

In principle, the existence of a solution to equation (1.3.1) depends on how much $\boldsymbol{X}$, within its support, can affect the propensity scores at $z' \in S(Z)$. For example if the propensity scores at z' have full support, i.e., $(p_1(\cdot, z'), p_2(\cdot, z')) : S(\boldsymbol{X}|z') \mapsto [0,1] \times [0,1]$ is surjective, then a solution always exists. In general, the higher the dimension of $\boldsymbol{X}$ is, the larger its effect on the propensity score, and the larger its support is, the more likely a solution is to exist. For instance, in Example OC, $x_0 \pm \frac{\alpha}{\beta} \in S(X)$ if $S(X)$ is large and/or α/β, i.e., the relative effect of Z with respect to X, is small.

If the converse is also true, the solutions to equation (1.3.1) are candidates of the matching points of $\boldsymbol{x}_0$.

Definition PSC (Propensity Score Coherence, PSC)**.** *Suppose $\boldsymbol{p}(\boldsymbol{x}, z) = \boldsymbol{p}(\boldsymbol{x}', z')$. The selection model is said to be propensity score coherent at $(\boldsymbol{x}, z)$ and $(\boldsymbol{x}', z')$ if $\boldsymbol{h}(\boldsymbol{x}, z, \boldsymbol{V}) = \boldsymbol{h}(\boldsymbol{x}', z', \boldsymbol{V})$ a.s.*

Note that if $\boldsymbol{h}$ is identified by propensity scores at $(\boldsymbol{x}_0, z)$, that is, there does not exist $\boldsymbol{h}'(\boldsymbol{x}_0, z, \boldsymbol{V}) \neq \boldsymbol{h}(\boldsymbol{x}_0, z, \boldsymbol{V})$ with positive probability such that the propensity scores at $(\boldsymbol{x}_0, z)$ based on $\boldsymbol{h}$ and $\boldsymbol{h}'$ are equal, then PSC holds at $(\boldsymbol{x}_0, z)$ and $(\boldsymbol{x}, z')$ for any $\boldsymbol{x}$ that solves equation (1.3.1). Many familiar discrete choice models are identified by propensity scores. We present examples in Appendix A.3.

On the other hand, the existence of pairs that satisfy PSC does not require $\boldsymbol{h}(\boldsymbol{x}_0, z, \cdot)$ to be identified by the propensity scores. The following example adapted from the *two-way flow* model in Lee and Salanié (2018) illustrates it.

Example TWF (Two-Way Flow)**.** *Let $h_1(\boldsymbol{X}, Z, V) = \mathbb{1}\big(V_1 \leq \gamma_1(\boldsymbol{X}, Z), V_2 \leq \gamma_2(\boldsymbol{X}, Z)\big)$, $h_2(\boldsymbol{X}, Z, V) = \mathbb{1}\big(V_1 \geq \gamma_1(\boldsymbol{X}, Z), V_2 \geq \gamma_2(\boldsymbol{X}, Z)\big)$, and $h_3 = 1 - h_1 - h_2$. V_1 and V_2 are two scalar random variables that are continuously distributed.*

It can be seen that the model is not identified by the propensity scores. For example, there may exist $\gamma_1'(\boldsymbol{x}_0, z) < \gamma_1(\boldsymbol{x}_0, z)$ and $\gamma_2'(\boldsymbol{x}_0, z) > \gamma_1(\boldsymbol{x}_0, z)$, but the propensity scores are equal. In the meanwhile, for some $\boldsymbol{x}'$, as long as $\gamma_1(\boldsymbol{x}_0, z) = \gamma_1(\boldsymbol{x}', z')$ and $\gamma_2(\boldsymbol{x}_0, z) = \gamma_2(\boldsymbol{x}', z')$, PSC holds at $(\boldsymbol{x}_0, z)$ and $(\boldsymbol{x}', z')$.

Another related concept is "index sufficiency" in the literature on local instrumental variable (LIV) and marginal treatment effect (MTE) (e.g. Heckman and Vytlacil (1999, 2001, 2005), Heckman, Urzua and Vytlacil (2006), etc.). For separable model as an example, index sufficiency says that the following equation holds for any $d \in S(D)$,

$$\mathbb{E}_{U|D\boldsymbol{X}Z}(d, \boldsymbol{X}, Z) = \mathbb{E}_{U|D\boldsymbol{X}Z}\left(d, \boldsymbol{X}, \boldsymbol{p}(\boldsymbol{X}, Z)\right)$$

This literature focuses on continuous Z and as they noted, Z needs to contain at least two variables for the condition to have empirical content. For index sufficiency to hold in that scenario, they essentially require that $(\boldsymbol{X}, Z)$ enters the selection model only through $\boldsymbol{p}(\boldsymbol{X}, Z)$, i.e., there exists an indicator function $\tilde{\boldsymbol{h}}$ such that $\boldsymbol{h}(\boldsymbol{X}, Z, \cdot) = \tilde{\boldsymbol{h}}(\boldsymbol{p}(\boldsymbol{X}, Z), \cdot)$.

In our context, Z is a scalar and binary. Index sufficiency trivially holds if Z is "relevant" (i.e. $\boldsymbol{p}(\boldsymbol{X}, z) \neq \boldsymbol{p}(\boldsymbol{X}, z')$ a.s.). This is because given $\boldsymbol{X}$, $\boldsymbol{p}(\boldsymbol{X}, Z)$ and Z are then one-to-one. Hence, index sufficiency itself does not have identification power in our setup. However, the condition that $(\boldsymbol{X}, Z)$ enters the selection model only through $\boldsymbol{p}(\boldsymbol{X}, Z)$ is sufficient for our purpose. With the selection model $\tilde{\boldsymbol{h}}(\boldsymbol{p}(\boldsymbol{X}, Z), \cdot)$, if $\boldsymbol{p}(\boldsymbol{x}_0, z) = \boldsymbol{p}(\boldsymbol{x}_m, z')$, then under exogeneity of the instrument and condition (1.2.4), we again have the matching equation:

$$\mathbb{E}_{U|D\boldsymbol{X}Z}(d, \boldsymbol{x}_0, z) = \mathbb{E}_{U|D\boldsymbol{X}\boldsymbol{p}}\left(d, \boldsymbol{x}_0, \boldsymbol{p}(\boldsymbol{x}_0, z)\right) = \mathbb{E}_{U|D\boldsymbol{X}\boldsymbol{p}}\left(d, \boldsymbol{x}_m, \boldsymbol{p}(\boldsymbol{x}_m, z')\right) = \mathbb{E}_{U|D\boldsymbol{X}Z}(d, \boldsymbol{x}_m, z')$$

In the LIV and MTE literature, for $\boldsymbol{h}$ to have the desired structure and index sufficiency to hold, $\boldsymbol{V}$ is required to be separable in the selection model. In our language, PSC needs to hold everywhere. However, this is unnecessary for our purpose. In the

example of the two-way flow model, a valid matching point may still be obtained by propensity score matching although PSC does not hold everywhere. PSC is a more "local" concept in the sense that as long as there exists one point that jointly satisfies the propensity score equation and matches the selection function, that point is a potential matching point so that identification of the outcome function could be achieved.

Similar to conditions (1.2.4) and (1.2.5), the matching equations (1.2.6) and (1.2.7) in Theorem MEQ are testable implications for PSC by plugging in matching points yielded by propensity score matching.

1.3.2 Identification of the SP-Outcome Functions

Given the m-connected set of $\boldsymbol{x}_0$, we are ready to study the identification of the outcome functions. We begin with the separable model-SP.

For illustrative purposes, let us only consider one matching point $\boldsymbol{x}_m$ of $\boldsymbol{x}_0$ first. Substituting the matching equation (1.2.6) into (1.2.11) and (1.2.12) for each d and deleting the redundant equation, we can see $\boldsymbol{m}^*(\boldsymbol{x}_0)$ satisfies the following system of equations:

$$\begin{aligned}
&\Pi_{SP} \cdot \boldsymbol{m}^*(\boldsymbol{x}_0) \\
&= \begin{pmatrix}
\sum_{d=1}^{3} \mathbb{E}_{Y|DXZ}(d, \boldsymbol{x}_0, z) p_d(\boldsymbol{x}_0, z) \\
\sum_{d=1}^{3} \mathbb{E}_{Y|DXZ}(d, \boldsymbol{x}_0, z') p_d(\boldsymbol{x}_0, z') \\
\underbrace{\sum_{d=1}^{3} \mathbb{E}_{Y|DXZ}(d, \boldsymbol{x}_m, z) p_d(\boldsymbol{x}_m, z)}_{=\sum_{d=1}^{3} m_d^*(\boldsymbol{x}_m) p_d(\boldsymbol{x}_m, z)} + \sum_{d=1}^{3} \Big[\underbrace{\mathbb{E}_{Y|DXZ}(d, \boldsymbol{x}_0, z) - \mathbb{E}_{Y|DXZ}(d, \boldsymbol{x}_m, z')}_{m_d^*(\boldsymbol{x}_0) - m_d^*(\boldsymbol{x}_m)} \Big] p_d(\boldsymbol{x}_m, z)
\end{pmatrix}
\end{aligned} \tag{1.3.2}$$

where $\Pi_{SP} = \begin{pmatrix} p_1(\boldsymbol{x}_0, z) & p_2(\boldsymbol{x}_0, z) & p_3(\boldsymbol{x}_0, z) \\ p_1(\boldsymbol{x}_0, z') & p_2(\boldsymbol{x}_0, z') & p_3(\boldsymbol{x}_0, z') \\ p_1(\boldsymbol{x}_m, z) & p_2(\boldsymbol{x}_m, z) & p_3(\boldsymbol{x}_m, z) \end{pmatrix}$.

The first two equations in the system are directly from Proposition 1. In the third

equation, we condition on $\boldsymbol{x}_m$ instead of $\boldsymbol{x}_0$; the first term on the right hand side again follows from Proposition 1. The second term, obtained from Theorem MEQ, then accounts for the difference sending $\boldsymbol{m}^*(\boldsymbol{x}_m)$ back to $\boldsymbol{m}^*(\boldsymbol{x}_0)$. Since the system of equations (1.3.2) is linear in $\boldsymbol{m}^*(\boldsymbol{x}_0)$, it is identified if Π_{SP} is full rank.

More generally, recall the augmented set of conditioning points $\mathcal{Z}(\boldsymbol{x}_0) \equiv \mathcal{X}_{MC}(\boldsymbol{x}_0) \times S(Z)$ introduced in Section 1.2.3. The equation system (1.3.2) can be easily adapted for any point in $\mathcal{Z}(\boldsymbol{x}_0)$. Then $\boldsymbol{m}^*(\boldsymbol{x}_0)$ is identified if Π_{SP} constructed by any three points in $\mathcal{Z}(\boldsymbol{x}_0)$ is full rank. Further, once $\boldsymbol{m}^*(\boldsymbol{x}_0)$ is identified, $\boldsymbol{m}^*(\cdot)$ at any other points in $\mathcal{X}_{MC}(\boldsymbol{x}_0)$ is also identified.

Theorem ID-SP. *Under Assumptions E-SP, if there exists $\tilde{\boldsymbol{z}}_1, \tilde{\boldsymbol{z}}_2, \tilde{\boldsymbol{z}}_3 \in \mathcal{Z}(\boldsymbol{x}_0)$ such that*

$$\Pi_{SP} = \begin{pmatrix} p_1(\tilde{\boldsymbol{z}}_1) & p_2(\tilde{\boldsymbol{z}}_1) & p_3(\tilde{\boldsymbol{z}}_1) \\ p_1(\tilde{\boldsymbol{z}}_2) & p_2(\tilde{\boldsymbol{z}}_2) & p_3(\tilde{\boldsymbol{z}}_2) \\ p_1(\tilde{\boldsymbol{z}}_3) & p_2(\tilde{\boldsymbol{z}}_3) & p_3(\tilde{\boldsymbol{z}}_3) \end{pmatrix} \text{ is full rank,}$$

then $\boldsymbol{m}^*(\boldsymbol{x})$ is identified for all $\boldsymbol{x} \in \mathcal{X}_{MC}(\boldsymbol{x}_0)$.

Remark 3.2. *Note that the conditioning values in the theorem does not necessarily include $(\boldsymbol{x}_0, z)$ and $(\boldsymbol{x}_0, z')$. For instance, in Example OC Cont'd, they can be any three of the dotted points in Figure 1.1.*

Since Π_{SP} does not contain any components of $\boldsymbol{m}^*$, whether the full-rank condition holds or not solely depends on the selection model. In what follows, we provide sufficient and necessary conditions for Π_{SP} to be full-rank.

The Full Rank Condition

For simplicity, we go back to the case in the beginning of this section and consider one matching point $\boldsymbol{x}_m$ such that $\boldsymbol{p}(\boldsymbol{x}_m, z') = \boldsymbol{p}(\boldsymbol{x}_0, z)$. Recall that in this case,

$$\Pi_{SP} = \begin{pmatrix} p_1(\boldsymbol{x}_0, z) & p_2(\boldsymbol{x}_0, z) & p_3(\boldsymbol{x}_0, z) \\ p_1(\boldsymbol{x}_0, z') & p_2(\boldsymbol{x}_0, z') & p_3(\boldsymbol{x}_0, z') \\ p_1(\boldsymbol{x}_m, z) & p_2(\boldsymbol{x}_m, z) & p_3(\boldsymbol{x}_m, z) \end{pmatrix}$$

Since the sum of the three columns in Π_{SP} is equal to vector $\mathbf{1}$, it can be shown that Π_{SP} is full rank if and only if

$$\begin{aligned} &\big(p_1(\boldsymbol{x}_m, z) - p_1(\boldsymbol{x}_0, z)\big)\big(p_3(\boldsymbol{x}_0, z) - p_3(\boldsymbol{x}_0, z')\big) \\ \neq &\big(p_1(\boldsymbol{x}_0, z) - p_1(\boldsymbol{x}_0, z')\big)\big(p_3(\boldsymbol{x}_m, z) - p_3(\boldsymbol{x}_0, z)\big) \end{aligned} \tag{1.3.3}$$

Inequality (1.3.3) does not hold if both sides are simultaneously zero. This is the case when Z has no effect on $\boldsymbol{p}$ at $\boldsymbol{X} = \boldsymbol{x}_0$ or $\boldsymbol{X}$ has no effect on $\boldsymbol{p}$ at $Z = z$. Both can be ruled out by a local relevance condition saying that $\boldsymbol{X}$ and Z have nonzero effects on the propensity scores at $(\boldsymbol{x}_0, z)$.

Now suppose neither side is 0. By $\boldsymbol{p}(\boldsymbol{x}_0, z) = \boldsymbol{p}(\boldsymbol{x}_m, z')$, inequality (1.3.3) can be rewritten as

$$\frac{p_1(\boldsymbol{x}_m, z) - p_1(\boldsymbol{x}_0, z)}{p_3(\boldsymbol{x}_m, z) - p_3(\boldsymbol{x}_0, z)} \neq \frac{p_1(\boldsymbol{x}_m, z') - p_1(\boldsymbol{x}_0, z')}{p_3(\boldsymbol{x}_m, z') - p_3(\boldsymbol{x}_0, z')} \tag{1.3.4}$$

The inequality generally holds unless the propensity score differences are locally uniform. For example, one can verify that the inequality is satisfied in the ordered choice model in Example OC for almost all x_0 in its support unless V is (locally) uniformly distributed. In particular, it holds for widely applied Logit and Probit models. The following example provides sufficient and necessary conditions for inequality (1.3.4) in an ordered choice model with multi-dimensional unobservables.

Example OC Cont'd 2. *Suppose now there are two unobservables in the ordered choice model:* $h_1(X,Z,\boldsymbol{V}) = \mathbb{1}(V_1 \leq \kappa_1 + \alpha Z + \beta X)$, $h_3(X,Z,\boldsymbol{V}) = \mathbb{1}(V_2 > \kappa_2 + \alpha Z + \beta X)$, *and* $h_2(X,Z,\boldsymbol{V}) = 1 - h_1(X,Z,\boldsymbol{V}) - h_3(X,Z,\boldsymbol{V})$. *To guarantee* $V_1 < V_2$ *a.s., we assume both are continuously distributed on* $S(V_1) \equiv (-\infty, c]$ *and* $S(V_2) \equiv [c, \infty)$ *respectively where* $c \in \mathbb{R}$. *Finally, assume* $\alpha \cdot \beta \neq 0$ *and we only consider the matching point* $x_m = x_0 - \frac{\alpha}{\beta}$.

Theorem ID-OC (Identification under Example OC Cont'd 2)**.** *Under the setup in Example OC Cont'd 2,* Π_{SP} *is full rank if and only if the single index* $X\beta + Z\alpha$ *evaluated at* $(x_0, 0)$, $(x_0, 1)$ *and* $(x_m, 0)$ *do not all fall into* $S(V_1)$ *or* $S(V_2)$ *at the same time.*

1.3.3 Identification of the NSP-Outcome Functions

As before, let us start from one matching point $\boldsymbol{x}_m$ such that, $\boldsymbol{p}(\boldsymbol{x}_m, z') = \boldsymbol{p}(\boldsymbol{x}_0, z)$. Similar to Section 1.3.2, we can substitute equation (1.2.8) into equation (1.2.2) for $(\boldsymbol{x}_m, z)$. Then $g^*(\boldsymbol{x}_0, u)$ solves the following system of equations for every $u \in [0,1]$:

$$\sum_{d=1}^{3} p_d(\boldsymbol{x}_0, z) \cdot F_{Y|DXZ}(g_d^*(\boldsymbol{x}_0, u)|d, \boldsymbol{x}_0, z) = u \tag{1.3.5}$$

$$\sum_{d=1}^{3} p_d(\boldsymbol{x}_0, z') \cdot F_{Y|DXZ}(g_d^*(\boldsymbol{x}_0, u)|d, \boldsymbol{x}_0, z') = u \tag{1.3.6}$$

$$\sum_{d=1}^{3} p_d(\boldsymbol{x}_m, z) \cdot F_{Y|DXZ}(\varphi_d(g_d^*(\boldsymbol{x}_0, u); \boldsymbol{x}_m, z')|d, \boldsymbol{x}_m, z) = u \tag{1.3.7}$$

Unlike identification of nonseparable models with a continuous D (e.g. Chernozhukov, Imbens and Newey (2007), Chen et al. (2014)), here we do not face the ill-posed problem due to the discreteness of D.

As the system is nonlinear in finite dimensional unknowns for a fixed u, it is well-known that the Jacobian of the system being full-rank at $g^*(\boldsymbol{x}_0, u)$ only implies local identification of $g^*(\boldsymbol{x}_0, u)$ (see Chernozhukov and Hansen (2005) and Chen et al. (2014) for examples). In what follows, we show that by continuity and monotonicity of $g^*(\boldsymbol{x}_0, \cdot)$, local identification at all $u \in [0,1]$ actually implies global identification of $g^*(\boldsymbol{x}_0, \cdot)$ in the

class of monotonic functions.

Let us first define a solution path, a concept widely adopted in differential equations.

Definition SolP (Solution Paths)**.** *For a system of equations* $\boldsymbol{M}(\boldsymbol{y}, u) = \mathbf{0}$*, where* $\boldsymbol{y}$ *is a real vector and* $u \in \mathcal{U}$*, a solution path* $\boldsymbol{y}^*(\cdot)$ *is a function on* $\mathcal{U}$ *such that* $\boldsymbol{M}(\boldsymbol{y}^*(u), u) = \mathbf{0}$ *for all* $u \in \mathcal{U}$*.*

Then we have the following lemma.

Lemma UNQ. *Let* $\mathcal{Y} = \{\boldsymbol{y} : [0,1] \mapsto S \equiv \prod_{l=1}^{L} S_l \subset \mathbb{R}^L | \forall 0 \leq u_1 \leq u_2 \leq 1, \boldsymbol{y}(u_1) \leq \boldsymbol{y}(u_2)\}$ *where for every* l*,* S_l *is a compact interval. Let* $\boldsymbol{M}(\cdot,\cdot) : S \times [0,1] \mapsto \mathbb{R}^d$ *be a continuous function and differentiable in the first* L *arguments where the Jacobian* $\nabla \boldsymbol{M}(\cdot,\cdot)$ *is also continuous. Suppose there exists a continuous solution path* $\boldsymbol{y}^* \in \mathcal{Y}$ *to* $\boldsymbol{M}(\cdot, u) = \mathbf{0}$ *on* $[0,1]$*. If for every* u*,* $\boldsymbol{M}(\cdot, u)$ *is strictly increasing in each argument and* $\nabla \boldsymbol{M}(\boldsymbol{y}^*(u), u)$ *is full-rank, then* $\boldsymbol{y}^*$ *is the unique solution path in* $\mathcal{Y}$*.*

Note that the domain $[0,1]$ of functions in $\mathcal{Y}$ can be replaced by any compact intervals in $\mathcal{R}$. Also, functions in $\mathcal{Y}$ can be decreasing. To see this, let $\tilde{\boldsymbol{M}}((-\boldsymbol{y}), u) = -\boldsymbol{M}(-(-\boldsymbol{y}), u)$. If $\boldsymbol{y}(u)$ is decreasing, $-\boldsymbol{y}(u)$ is increasing and $\tilde{\boldsymbol{M}}(\cdot, u)$ as a function of $-\boldsymbol{y}$ is strictly increasing in every argument. Similarly, $\boldsymbol{M}(\cdot, u)$ can be strictly decreasing as well.

Lemma UNQ shows that monotonicity and continuity simplify the sufficient conditions that are usually required for the global uniqueness of a solution path to a system of nonlinear equations at each u (see variants of Hadamard's theorem, e.g. Ambrosetti and Prodi (1995)). Here, full-rankness of the Jacobian matrix is only required along the unique solution path. For any fixed u, a full-rank Jacobian matrix only guarantees that the solution is locally unique, but from the lemma, local uniqueness of a solution at every u implies global uniqueness of a solution path under monotonicity and continuity. Note that the result not only holds for the class of increasing and continuous functions. Discontinuous functions are allowed in $\mathcal{Y}$.

Now let us stack the left hand side of equations (1.3.5) to (1.3.7) into a vector denoted by $\Psi(\boldsymbol{g}^*(\boldsymbol{x}_0, u))$. Denote the vector $(u, u, u)'$ by $\boldsymbol{u}$. Then $\boldsymbol{g}^*(\boldsymbol{x}_0, \cdot)$ is one solution path to $\boldsymbol{M}(\boldsymbol{y}, \boldsymbol{u}) \equiv \Psi(\boldsymbol{y}) - \boldsymbol{u} = 0$. By Assumption CM, $\boldsymbol{g}^*(\boldsymbol{x}_0, \cdot)$ is continuous and each component is strictly increasing on $[0, 1]$. We set $\mathcal{G}$ to be the set of all increasing functions defined on $[0, 1]$:

$$\mathcal{G} \equiv \{\boldsymbol{g} : [0, 1] \mapsto \mathbb{R}^3 \text{ and is weakly increasing}\}.$$

Recall that $\mathcal{Z}(\boldsymbol{x}_0) \equiv \mathcal{X}_{MC}(\boldsymbol{x}_0) \times S(Z)$ contains all the points that can be conditioned on to identify $\boldsymbol{g}^*(\boldsymbol{x}_0, \cdot)$. Let $\Psi(\cdot; \tilde{\boldsymbol{z}}_1, \tilde{\boldsymbol{z}}_2, \tilde{\boldsymbol{z}}_3)$ be the moment equations adapted from equations (1.3.5) to (1.3.7) by conditioning on $\tilde{\boldsymbol{z}}_1, \tilde{\boldsymbol{z}}_2, \tilde{\boldsymbol{z}}_3 \in \mathcal{Z}(\boldsymbol{x}_0)$. For example, the k-th component in Ψ is $\sum_{d=1}^{3} p_d(\tilde{\boldsymbol{z}}_d) \cdot F_{Y|D\boldsymbol{X}Z}(\varphi_d(\cdot)|d, \tilde{\boldsymbol{z}}_k)$. By Lemma UNQ and the special structures of CDFs, the following theorem provides sufficient conditions that guarantee global identification of $\boldsymbol{g}^*(\boldsymbol{x}_0, \cdot)$ in $\mathcal{G}$.

Theorem ID-NSP. *Under Assumptions E-NSP, FS, and CM, if there exist $\tilde{\boldsymbol{z}}_1, \tilde{\boldsymbol{z}}_2, \tilde{\boldsymbol{z}}_3 \in \mathcal{Z}(\boldsymbol{x}_0)$ such that $\Psi(\cdot; \tilde{\boldsymbol{z}}_1, \tilde{\boldsymbol{z}}_2, \tilde{\boldsymbol{z}}_3)$ is continuously differentiable on $\prod_{d=1}^{3} S(Y|d, \boldsymbol{x}_0)$, and that its Jacobian matrix at $\boldsymbol{g}^*(\boldsymbol{x}_0, u)$, $\Pi_{NSP}(\boldsymbol{g}^*(\boldsymbol{x}_0, u))$, is full-rank for all $u \in [0, 1]$, then $\boldsymbol{g}^*(\boldsymbol{x}_0, \cdot)$ is the unique solution path (up to $u = 0, 1$) to $\Psi(\cdot; \tilde{\boldsymbol{z}}_1, \tilde{\boldsymbol{z}}_2, \tilde{\boldsymbol{z}}_3) - \boldsymbol{u} = 0$ in $\mathcal{G}$.*

Note that we allow a larger parameter space in Theorem ID-NSP than in Lemma UNQ; functions whose ranges are outside the conditional support $\prod_d S(Y|d, \boldsymbol{x}_0)$ are allowed. This is due to nice properties of CDFs. By construction, any candidate $\boldsymbol{g}(u)$ enters Ψ through the conditional CDFs. Outside the support, a CDF is constant and equal to the value at the corresponding boundary of the support. Therefore, for any $\boldsymbol{g}(u)$ taking values outside the support, there exists a function taking values within the support (including the boundaries) that yields the same Ψ. Hence, the uniqueness of the solution path among increasing functions ranging within the support also holds among all increasing functions with arbitrary ranges, with the only exceptions at the endpoints

$u = 0, 1$; points that lies outside $\prod_d S(Y|d, \boldsymbol{x}_0)$ trivially satisfy the moment equations at $u = 0$ or 1.

Allowing the parameter space to contain functions outside the conditional support is useful in estimation because we no long need to accurately estimate the boundaries of the conditional support to obtain a consistent estimator of $\boldsymbol{g}^*(\boldsymbol{x}_0, \cdot)$.

Remark 3.3. *By definition,* $\varphi_d(g_d^*(\boldsymbol{x}_0, \cdot); \boldsymbol{x}_0, z) = g_d^*(\boldsymbol{x}_0, \cdot)$. *So Theorem ID-NSP also applies to the standard IV approach when D is discrete with* $|S(Z)| \geq |S(D)|$, *for example, Chernozhukov and Hansen (2005).*

Before we close this section, let us emphasize that the identification notion in Theorem ID-NSP is in terms of the uniqueness of monotonic solution path. It does not rule out the possibility that at certain u, the solution to $\Psi(\cdot) = \boldsymbol{u}$ is not unique. This is expected because the conditions we require are much weaker than the sufficient conditions for global invertibility of $\Psi(\cdot)$ on its entire domain $\Pi_{d=1}^3 S(Y|d, \boldsymbol{x}_0)$. Under this weaker notion of identification, estimation cannot be conducted for a fixed u; as will be seen in the next section, we will estimate $\boldsymbol{g}^*(\boldsymbol{x}_0, \cdot)$ at multiple nodes jointly by imposing monotonicity and assuming the number of the nodes grows to infinity with the sample size.

1.4 Estimation

In this section, we propose estimators for the matching points and the outcome functions given an independently and identically distributed sample $(Y_i, D_i, \boldsymbol{X}_i, Z_i)_{i=1}^n$. We also discuss some practical issues for implementation.

The estimation strategy follows our constructive identification. From Section 1.3, the matching points can be obtained by either matching the propensity scores or matching the selection functions, depending on the assumptions made on $\boldsymbol{h}$. Meanwhile, the moment conditions for $\boldsymbol{m}^*(\boldsymbol{x}_0)$ and $\boldsymbol{g}^*(\boldsymbol{x}_0, u)$ essentially can be constructed by condition-

ing on any three values in $\mathcal{Z}(\boldsymbol{x}_0)$. For illustrative purpose, we focus on the following benchmark case to highlight the key features of the estimation procedure.

1. $\boldsymbol{X}$ is one-dimensional, denoted by X.

2. Two matching pairs exist: $(x_0, 0)$, $(x_{m1}, 1)$ and $(x_0, 1)$, $(x_{m2}, 0)$. PSC holds at each pair.

The benchmark conditions setup the simplest scenario while both the matching points (due to Condition 1) and the outcome functions (due to Condition 2) are over-identified, allowing us to construct over-identification tests. Extending Condition 1 to multivariate $\boldsymbol{X}$ is straightforward. Condition 2 is testable by the over-identification test.

The Matching Points

Let $\hat{\boldsymbol{p}}(\cdot, z)$, $z = 0, 1$, be a consistent estimator of $\boldsymbol{p}(\cdot, z)$ uniformly on $S_0(X)$, a compact interior subset of $S(X)$. We assume both matching points are in $S_0(X)$. Let $\Delta\hat{\boldsymbol{p}}(x_1, x_2) \equiv \big(\hat{p}_1(x_1, 1) - \hat{p}_1(x_0, 0), \hat{p}_2(x_1, 1) - \hat{p}_2(x_0, 0), \hat{p}_1(x_2, 0) - \hat{p}_1(x_0, 1), \hat{p}_2(x_2, 0) - \hat{p}_2(x_0, 1)\big)'$. Finally for some weighting matrix $\boldsymbol{W}_{xn}$ with positive definite probability limit, let $\hat{Q}_x(x_1, x_2) \equiv \Delta\hat{\boldsymbol{p}}(x_1, x_2)'\boldsymbol{W}_{xn}\Delta\hat{\boldsymbol{p}}(x_1, x_2)$. Under PSC, the estimator $(\hat{x}_{m1}, \hat{x}_{m2})$ we propose are points in $S_0^2(X)$ such that for some $a_n = o(1)$,

$$\hat{Q}_x(\hat{x}_{m1}, \hat{x}_{m2}) \leq \inf_{S_0^2(X)} \hat{Q}_x(x_1, x_2) + a_n \tag{1.4.1}$$

When $a_n = 0$, $(\hat{x}_{m1}, \hat{x}_{m2})$ is the minimizer of $\hat{Q}_x(x_1, x_2)$. In general, the minimizer of $\hat{Q}_x(x_1, x_2)$ is consistent of (x_{m1}, x_{m2}) only if the latter is the unique minimizer of the population objective function Q_x. When $Q_x(\cdot, \cdot)$ has multiple minima, which is allowed for the purpose of identification, the set of the minimizers of $\hat{Q}_x$ tends to be smaller than the true set, and its probability limit may not exist. For example, suppose Q_x has two global minima on $S_0^2(X)$ but $\hat{Q}_x$ may only have one. As $n \to \infty$, the minimum of $\hat{Q}_x$ may

jump across the neighborhoods of the two minima of Q_x. The probability limit of being in any one particular neighborhood may thus be strictly smaller than one. To handle the general multiple minima case, we let $a_n > 0$ and converge to o at an appropriate rate similar to Chernozhukov, Hong and Tamer (2007). We discuss the general case in Appendix A.1.2. For simplicity, here we focus on the case where (x_{m1}, x_{m2}) is unique and let $a_n = 0$.

For concreteness, we consider the following kernel estimator for the propensity scores. We can use other nonparametric estimators for conditional probability too.

$$\hat{p}_d(x,z) = \frac{\sum_{i=1}^N \mathbb{1}(D_i = d) K(\frac{X_i - x}{h_x}) \mathbb{1}(Z_i = z)}{\sum_{i=1}^N K(\frac{X_i - x}{h_x}) \mathbb{1}(Z_i = z)} \tag{1.4.2}$$

where $K(\cdot)$ is a kernel function and h_x is the bandwidth converging to 0. Regularity conditions for K and the convergence rate of h_x will be given in Section 1.6.

The SP-Outcome Functions

For the SP-outcome functions, by linearity of the moment conditions (1.3.2), we obtain the following closed-form estimator by inverting the estimated Π_{SP} matrix (weighted by $\boldsymbol{W}_{mn}$):

$$\hat{\boldsymbol{m}}(x_0) = (\widehat{\Pi}_{SP}' \boldsymbol{W}_{mn} \widehat{\Pi}_{SP})^{-1} \cdot \widehat{\Pi}_{SP}' \boldsymbol{W}_{mn} \widehat{\Phi}(\hat{x}_{m1}, \hat{x}_{m2}) \tag{1.4.3}$$

where in this case $\widehat{\Pi}_{SP} = \begin{pmatrix} \hat{p}_1(x_0,0) & \hat{p}_2(x_0,0) & \hat{p}_3(x_0,0) \\ \hat{p}_1(x_0,1) & \hat{p}_2(x_0,1) & \hat{p}_3(x_0,1) \\ \hat{p}_1(\hat{x}_{m1},0) & \hat{p}_2(\hat{x}_{m1},0) & \hat{p}_3(\hat{x}_{m1},0) \\ \hat{p}_1(\hat{x}_{m2},1) & \hat{p}_2(\hat{x}_{m2},1) & \hat{p}_3(\hat{x}_{m2},1) \end{pmatrix}$ and

$$\widehat{\Phi}(\hat{x}_{m1},\hat{x}_{m2}) = \begin{pmatrix} \sum_{d=1}^{3} \widehat{\mathbb{E}}_{Y|DXZ}(d,x_0,0)\hat{p}_d(x_0,0) \\ \sum_{d=1}^{3} \widehat{\mathbb{E}}_{Y|DXZ}(d,x_0,1)\hat{p}_d(x_0,1) \\ \sum_{d=1}^{3} \left[\widehat{\mathbb{E}}_{Y|DXZ}(d,\hat{x}_{m1},0) + \widehat{\mathbb{E}}_{Y|DXZ}(d,x_0,0) - \widehat{\mathbb{E}}_{Y|DXZ}(d,\hat{x}_{m1},1)\right]\hat{p}_d(\hat{x}_{m1},0) \\ \sum_{d=1}^{3} \left[\widehat{\mathbb{E}}_{Y|DXZ}(d,\hat{x}_{m2},1) + \widehat{\mathbb{E}}_{Y|DXZ}(d,x_0,1) - \widehat{\mathbb{E}}_{Y|DXZ}(d,\hat{x}_{m2},0)\right]\hat{p}_d(\hat{x}_{m2},1) \end{pmatrix}.$$

The estimated propensity scores in $\widehat{\Pi}_{SP}$ and $\widehat{\Phi}$ follow equation (1.4.2). The conditional expectations are estimated by the standard Nadaraya-Waston estimator:

$$\widehat{\mathbb{E}}_{Y|DXZ}(d,x,z) = \frac{\sum_{i=1}^{N} Y_i \mathbb{1}(D_i = d) K(\frac{X_i - x}{h_m}) \mathbb{1}(Z_i = z)}{\sum_{i=1}^{N} \mathbb{1}(D_i = d) K(\frac{X_i - x}{h_m}) \mathbb{1}(Z_i = z)} \tag{1.4.4}$$

The NSP-Outcome Functions

In the benchmark case, we have the following moment functions:

$$\Psi(\boldsymbol{g}(u)) = \begin{pmatrix} \sum_{d=1}^{3} p_d(x_0,0) \cdot F_{Y|DXZ}(g_d(u)|d,x_0,0) \\ \sum_{d=1}^{3} p_d(x_0,1) \cdot F_{Y|DXZ}(g_d(u)|d,x_0,1) \\ \sum_{d=1}^{3} p_d(x_{m1},0) \cdot F_{Y|DXZ}(\varphi_d(g_d(u);x_{m1},1)|d,x_{m1},0) \\ \sum_{d=1}^{3} p_d(x_{m2},1) \cdot F_{Y|DXZ}(\varphi_d(\tilde{g}_d(u);x_{m2},0)|d,x_{m2},1) \end{pmatrix}.$$

Let $\boldsymbol{u} = (u,u,u)'$ and $Q_{NSP}(\boldsymbol{g},u) \equiv \left[(\Psi(\boldsymbol{g}(u)) - \boldsymbol{u})' \boldsymbol{W}_g(u) (\Psi(\boldsymbol{g}(u)) - \boldsymbol{u})\right]$ for any positive definite matrix $\boldsymbol{W}_g(u)$. Our identification result for $\boldsymbol{g}^*(\boldsymbol{x}_0,\cdot)$ implies that it is the unique minimizer (up to $u = 0,1$) to the following minimization problem:

$$\min_{\boldsymbol{g} \in \mathcal{G}_0} \int_0^1 Q_{NSP}(\boldsymbol{g}(u),u)du \tag{1.4.5}$$

where $\mathcal{G}_0 \subseteq \mathcal{G}$ contains increasing functions on $[0,1]$ with ranges contained in $\prod_{d=1}^{3} S(Y|d)$.

Construct $\hat{Q}_{NSP}(g,u)$ by plugging in estimators of Ψ and $W_g(u)$. Let $u_j = \frac{j}{J}$ where $1 \leq j \leq J$ and $J \to \infty$. We estimate $g^*(x_0,\cdot)$ by solving the following minimization problem:

$$\min_{\underline{y} \leq g(u_1) \leq \ldots \leq g(u_J) \leq \bar{y}} \frac{1}{J} \sum_{j=1}^{J} \hat{Q}_{NSP}(g(u_j),u_j) + \lambda \sum_{j=2}^{J} \left(g(u_j) - g(u_{j-1})\right)' \left(g(u_j) - g(u_{j-1})\right) \tag{1.4.6}$$

Let us begin with the constraint. By linearly connecting the nodes $g(u_j)$ and $g(u_{j-1})$ for all $j = 2,..J$, the constraint induces a finite dimensional sieve space $\widehat{\mathcal{G}}$ of piecewise affine increasing functions defined on $[0,1]$. By sending $J \to \infty$, elements in the sieve space are able to approximate any continuous and increasing functions that are bounded by $\underline{y}$ and $\bar{y}$, the lower and upper bounds of $\prod_d S(Y|d)$. As D is discrete, for each d the bounds can be estimated by $\underline{y}_d = \min(Y_i|D_i = d)$ and $\bar{y}_d = \max(Y_i|D_i = d)$. We treat the bounds as known parameters as these estimators converge faster than the nonparametric rate[2].

The second term in equation (1.4.6) is a penalty making the estimator smoother in finite samples. We let $\lambda \to 0$ fast enough so the the penalty does not affect the estimator's asymptotic behavior.

As for $\widehat{\Psi}$, the conditional CDFs are estimated by the following smoothed kernel estimator (e.g. Hansen (2004) and Li and Racine (2008)):

$$\widehat{F}_{Y|DXZ}(y|d,x,z) = \frac{\sum_{i=1}^{N} L(\frac{y-Y_i}{h_0}) \mathbb{1}(D_i = d) K(\frac{X_i - x}{h_g}) \mathbb{1}(Z_i = z)}{\sum_{i=1}^{N} \mathbb{1}(D_i = d) K(\frac{X_i - x}{h_g}) \mathbb{1}(Z_i = z)} \tag{1.4.7}$$

[2]Alternatively, we could shrink $\mathcal{G}_0$ so that only functions bounded within $\prod_d S(Y|d,x_0)$ are included and $\underline{y}$ and $\bar{y}$ are boundaries estimators for this smaller support set. In this space $g^*(x_0,\cdot)$ is unqiue including at the end points. Also, as will be seen in Section 1.6.3, nicer boundary properties can be obtained. However, since X is continuous, boundary estimators of $\prod_d S(Y|d,x_0)$ (e.g. Guerre, Perrigne and Vuong (2000)) involve extra tuning parameters. For simplicity, we do not adopt this approach.

where $L(\cdot)$ is a smooth CDF supported on a bounded interval and $h_0 \mapsto 0$ faster than h_g. Another component in $\widehat{\Psi}$ is the function $\hat{\varphi}_d$:

$$\hat{\varphi}_d(y; x_m, z') = \arg \min_{y' \in [\underline{y}_d, \bar{y}_d]} \left(\widehat{F}_{Y|DXZ}(y'|d, \hat{x}_m, z') - \widehat{F}_{Y|DXZ}(y|d, x_0, z) \right)^2 \tag{1.4.8}$$

Remark 4.1. *Under pointwise identification, estimation can be simplified; one can minimize $\hat{Q}_{NSP}(\boldsymbol{g}(u), u)$ at each u of interest separately (e.g. Lewbel (2007)). The inequality constraints can be dropped. The dimension of each individual minimization problem is smaller. Computation is thus made easier. Under pathwise identification, joint estimation under the constraint of monotonicity is necessary because it is possible that $\boldsymbol{g}^*(\boldsymbol{x}_0, u)$ is not the unique solution to the moment equations for some u. The minimizers of $\hat{Q}_{NSP}(\boldsymbol{g}(u), u)$ at these u are then inconsistent.*

1.5 Empirical Applications

Before we move into the asymptotic theory of the estimators, we consider two applications to illustrate the usefulness and limitation of our approach. The first application is the return to education example we discussed earlier. The second studies preschool program selection using the administrative Head Start Impact Study (HSIS) dataset.

1.5.1 The Return to Schooling: A Binary D

We use the same extract from the 1979 NLS as in Card (1995). The outcome variable Y is the log wage. We adopt the same IV which indicates whether an individual grew up near an accredited four-year college. In this subsection, we assume that the latent selection mechanism yields two outcomes: $D = 1$ if the years of schooling is greater than 12 and $D = 0$ otherwise. We will consider a three-valued D in the next subsection. We use the average across parents' years of schooling as the matching covariate X. Finally, we drop the observations who were still enrolled in a school at the time of the survey. The remaining sample size is 2000.

We assume the log wage is determined by Model-SP. As $\boldsymbol{m}^*$ is identified by the standard IV approach with the binary Z, we can compare the results using the standard IV method and our approach.

No Covariates

Let us first consider the following case assuming no covariates are in the outcome functions:

$$Y = \sum_{d=0}^{1} \mathbb{1}(D = d) m_d^* + U.$$

This model sets up a clean benchmark because m_0^* and m_1^* are identified by Z so no extra steps for propensity score matching are needed. In fact they can be estimated by the simple Wald estimator. The results provide us with references about the magnitudes of the outcome function and the effects. As shown in Table 1.1 (standard errors in parentheses), the return to education is increasing in D. The wage for individuals receiving post-high school education is on average 1.35% higher than those with at most high school education.

Table 1.1: IV Estimates

$\hat{m}_0$	5.58
	(0.18)
$\hat{m}_1$	6.93
	(0.16)

Covariates and Matching

Let us first illustrate the process of finding a matching point. Figure 1.2 depicts $\hat{p}_0(x, 0)$ and $\hat{p}_0(x, 1)$. The black dashed lines illustrate the how we find $\hat{x}_{m1}$: For the fixed x_0 in the left panel, we find the value of the propensity score $\hat{p}_0(x_0, 0)$, and find $\hat{x}_{m1}$ in the right panel such that $\hat{p}_0(\hat{x}_{m1}, 1) = \hat{p}_0(x_0, 0)$. Similarly, the blue dash-dot line starts

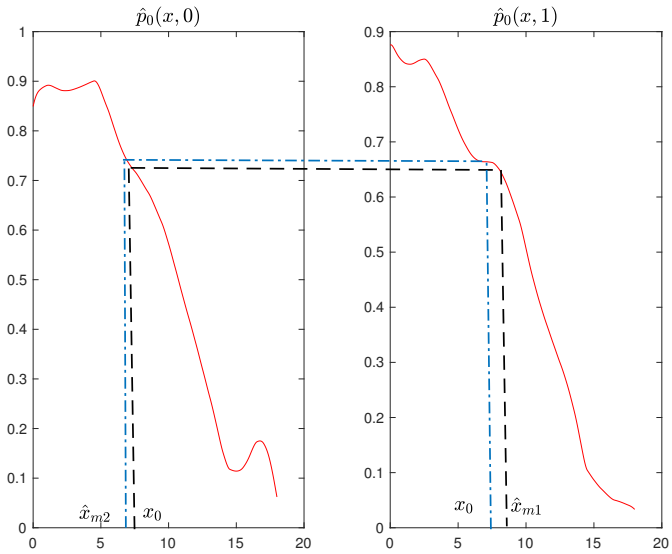

Figure 1.2: Propensity Scores: $|S(D)| = 2$

from x_0 in the right panel for $Z = 1$, and the second matching point $\hat{x}_{m2}$ is found in the left panel. In Figure 1.3, the red solid curves in the left and right panels are $\hat{p}_0(x,1) - \hat{p}_0(12,0)$ and $\hat{p}_0(x,0) - \hat{p}_0(12,1)$ respectively. These propensity score differences clearly intersect with zero. The intersection points are the estimated matching points. The patterns for other values of x_0 we consider are similar and are thus omitted.

Figures 1.2 and 1.3 imply that individuals whose parents have more years of schooling are more likely to attain post-high school education. Also, from the values of the matching points, living close to a four-year college and parents education are substitutes. At $X = 12$, an increase of about half a year in parents' education compensates for not living near a college.

Now let us turn to the outcome function estimates at $x_0 = 10, 11, 12$, shown in Table 1.2. The second row "Matching" indicates whether the matching points are estimated and used. When not using the matching points, we estimate $(m_0^*(x_0), m_0^*(x_1))$ by inverting the following moment conditions:

$$\begin{pmatrix} \hat{p}_0(x_0,0), \hat{p}_1(x_0,0) \\ \hat{p}_0(x_0,1), \hat{p}_1(x_0,1) \end{pmatrix} \begin{pmatrix} \hat{m}_0(x_0) \\ \hat{m}_1(x_0) \end{pmatrix} = \begin{pmatrix} \hat{p}_0(x_0,0)\hat{\mathbb{E}}_{Y|DXZ}(0,x_0,0) + \hat{p}_1(x_0,0)\hat{\mathbb{E}}_{Y|DXZ}(1,x_0,0) \\ \hat{p}_0(x_0,1)\hat{\mathbb{E}}_{Y|DXZ}(0,x_0,1) + \hat{p}_1(x_0,1)\hat{\mathbb{E}}_{Y|DXZ}(1,x_0,1) \end{pmatrix}.$$

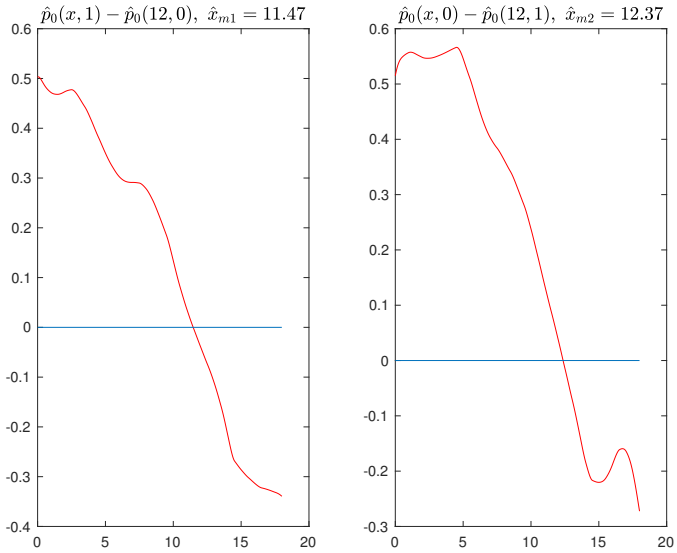

Figure 1.3: Propensity Score Differences: $x_0 = 12$, $|S(D)| = 2$

When we use our approach, we first estimate the matching points by grid search over 500 nodes. The standard errors in parentheses are computed using the asymptotic variance estimators derived in the next section. Note that with the matching points, the model is over-identified because we have four moment conditions and two unknowns. Hence, we can perform the over-identification test with the null hypothesis that all the moment conditions are valid. The test is as in the standard GMM framework and we will provide details in Section 1.6.2. The p-values of the test results are in the last row.

Table 1.2: Binary D with X

	$x_0 = 10$		$x_0 = 11$		$x_0 = 12$	
Matching:	✗	✓	✗	✓	✗	✓
$\hat{m}_0(x_0)$	5.63	5.64	5.59	5.56	5.35	5.37
	(0.28)	(0.17)	(0.33)	(0.20)	(0.60)	(0.33)
$\hat{m}_1(x_0)$	7.15	7.13	6.90	6.92	6.90	6.89
	(0.31)	(0.19)	(0.25)	(0.15)	(0.32)	(0.18)
Over-Id p value	N.A.	0.98	N.A.	0.99	N.A.	0.80

From Table 1.2, we can make three observations. First, the estimates using the two approaches are very close, but the variances are lower using our approach. Similar point estimates provide extra evidence in addition to the insignificant over-identification test statistics that the extra moment conditions brought in by the matching points are valid. Variance reduction is due to the use of more moment conditions. Consequently, the estimated effects are more significant. For instance, it can be computed that $\hat{m}_1(12) - \hat{m}_0(12)$ is significant at 10% level using the IV approach, but is significant at 1% level using our approach. Second, the outcome function is increasing in the level of education and heterogeneous in parents' education. Individuals with less educated parents have higher returns of education at both levels. Finally, for each level of education, the range of the heterogeneous estimates cover the results in Table 1.1, indicating the results in Table 1.2 are in a reasonable range.

1.5.2 The Return to Schooling: A Three-Valued D

In this subsection, we assume the underlying selection model yields three outcomes; we recode high school education by $D = 1$ and divide post-high school education into two groups: $D = 2$ if $12 <$ years of schooling ≤ 15 (some college), and $D = 3$ if years of schooling > 15 (college and above). In this case, no existing method can identify and estimate $\boldsymbol{m}^*(x_0)$ without imposing extra structures on it.

Again, let us first illustrate how we find a matching point. Figure 1.4 depicts the propensity score functions at $Z = 0$ and $Z = 1$. Starting from x_0 on the left panel, we need to match both $\hat{p}_1(x_0, 0)$ and $\hat{p}_3(x_0, 0)$ with $\hat{p}_1(x, 1)$ and $\hat{p}_3(x, 1)$ at the same x. If such x exists, it is the estimated matching point $\hat{x}_{m1}$. Evidently, with a scalar X, x_{m1} is over-identified, so in the finite sample, it is very likely that we cannot exactly match both propensity scores, but we need the difference to be small enough. In Section 1.6.1, we propose an over-identification test to see whether all the propensity scores can be indeed matched with a single covariate.

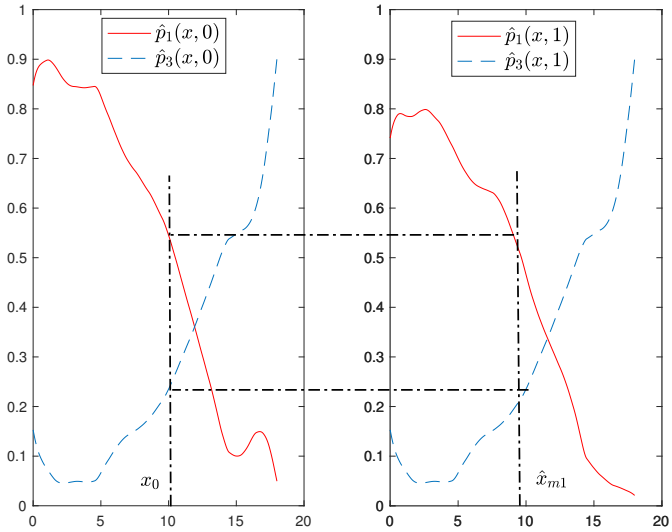

Figure 1.4: Propensity Scores: $|S(D)| = 3$

Figure 1.5 illustrates the matching points for $x_0 = 12$. Again we omit other values of x_0 as the patterns are similar. The solid red curves in the two panels in Figure 1.5 are $\hat{p}_1(x,1) - \hat{p}_1(x_0,0)$ and $\hat{p}_1(x,0) - \hat{p}_1(x_0,1)$, while the dashed blue curves are $\hat{p}_3(x,1) - \hat{p}_3(x_0,0)$ and $\hat{p}_3(x,0) - \hat{p}_3(x_0,1)$. Matching is successful if the solid curve and the dashed curve intersect with the horizontal line of zero at the same point. From the figure, the intersection points are indeed very close in both panels. This is also supported by the over-identification tests; $\mathcal{J}_{x1}$ and $\mathcal{J}_{x2}$ on top are insignificant in both cases. Finally, since the baseline level here (years of schooling ≤ 12) is the same as that in the previous case, the propensity scores at the baseline level of D are equal. Hence, the estimated matching points in these two cases should be similar. Here the estimates are 11.54 and 12.34, indeed very close to those when D is binary (11.47 and 12.37).

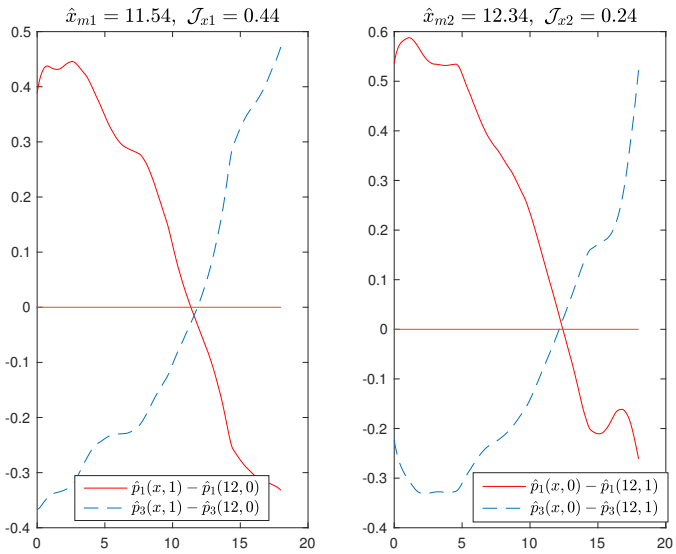

Figure 1.5: Propensity Score Differences: $x_0 = 12$, $|S(D)| = 3$

Next, let us turn to the estimates of the outcome function shown in Table 1.3. Since both the outcome function and the matching points are over-identified in this case, we present the p-values for each over-identification test in the bottom panel. First, we can see that none of the over-identification test statistics $\boldsymbol{m}^*(x_0)$ are significant at any reasonable level, similar to Table 1.2 in the binary case. Also, the joint over-identification tests for the matching points are also insignificant, confirming that all the propensity scores are matched. Second, the return to education is monotonic and the marginal return is slightly decreasing. Third, the returns are heterogeneous in parents education; similar to the binary case, individuals whose parents are less educated have slightly higher returns of education at almost all levels.

Table 1.3: Three-valued D

	$x_0 = 10$	$x_0 = 11$	$x_0 = 12$
$\hat{m}_1(x_0)$	5.62	5.56	5.33
	(0.23)	(0.22)	(0.38)
$\hat{m}_2(x_0)$	7.03	6.47	6.39
	(3.18)	(1.39)	(1.45)
$\hat{m}_3(x_0)$	7.28	7.32	7.31
	(2.72)	(1.09)	(1.00)
Over-Id p-value			
$\boldsymbol{m}^*(x_0)$	0.89	0.87	0.58
$\boldsymbol{x}_m$	0.54	0.36	0.41

1.5.3 Validity of the Exogeneity Assumption

In this subsection, we continue with this empirical example to show that in fact the covariate we choose may not be exogenous in the sense of $\mathbb{E}_{U|XZ}(X, Z) \neq 0$, so methods such as 2SLS using for example ZX as an extra instrument may not deliver correct estimates, even though they do not rely on the selection model. In contrast, our approach only imposes local exogeneity assumptions with respect to U and $\boldsymbol{V}$; though stronger than the standard nonparametric IV approach, we can still obtain informative results.

The conditional mean independence $\mathbb{E}_{U|XZ}(X, Z) = 0$ requires that $\mathbb{E}_{U|XZ}(x_0, Z) = 0$ for almost all $x_0 \in S(X)$. Since the latter equation is also required by our approach for fixed x_0, we can use our over-identification test to check if there are values of x_0 such that the condition may not hold. Specifically, we re-estimate the outcome function at $x_0 = 8$ and $x_0 = 14$, for both the binary D and the three-valued D. In each case, we conduct the over-identification test for the outcome function, and for three-valued D, we

also conduct the test for the matching points as they are also over-identified then. The results are in Table 1.4.

Table 1.4: Values of x_0 Where Exogeneity May Fail

Over-Id	$x_0 = 8$		$x_0 = 14$	
p-value	$\lvert S(D)\rvert = 2$	$\lvert S(D)\rvert = 3$	$\lvert S(D)\rvert = 2$	$\lvert S(D)\rvert = 3$
$\boldsymbol{m}^*(x_0)$	0.05	0.01	0.01	0.00
$\boldsymbol{x}_m$	N.A.	0.23	N.A.	0.14

The results imply that not all the moment conditions are valid at these two values of X, although matching is still successful. It suggests that the invalidity of the moment conditions is more likely to be driven by violations of the exogeneity assumption at these choices of x_0.

For verification, now let us turn to the 2SLS estimates under the conditional mean independence assumption. We consider the following two specifications:

Setup 1: $Y = \beta_0 + \mathbb{1}(D = 2)\beta_1 + \mathbb{1}(D = 3)\beta_2 + X\beta_3 + U$

Setup 2: $Y = \beta_0 + \mathbb{1}(D = 2)\beta_1 + \mathbb{1}(D = 3)\beta_2 + X\beta_3 + \mathbb{1}(D = 2)X\beta_4 + \mathbb{1}(D = 3)X\beta_5 + U$

Under $\mathbb{E}_{U|XZ}(X, Z) = 0$, polynomials of X and their interactions with Z are all valid IVs by the law of iterated expectation. We present the results for $x_0 = 12$ in Table 1.5. The estimates in Column (1) are obtained using our approach. Note that we do not utilize the parametric form, so they are the same as in Table 1.5. Columns (2)-(5) contain the fitted values for $x_0 = 12$ using the 2SLS estimates. Estimates under Setup 1 are in Columns (2)-(4), using (Z, ZX), (Z, ZX, X^2) and (Z, ZX, X^2, ZX^2) as instruments respectively. Column (5) contains results under Setup 2, using (Z, ZX, X^2, ZX^2) as instruments. The last row reports the p-values of the over-identification tests for our approach and for the standard 2SLS.

Table 1.5: Comparison with 2SLS

	MP	2SLS			
	(1)	(2)	(3)	(4)	(5)
$\hat{m}_1(12)$	5.33	4.87	5.63	5.76	5.35
	(0.38)	(0.71)	(0.37)	(0.29)	(0.80)
$\hat{m}_2(12)$	6.39	7.24	7.56	7.34	5.20
	(1.45)	(0.73)	(0.54)	(0.41)	(2.10)
$\hat{m}_3(12)$	7.31	7.08	6.24	6.26	8.24
	(1.00)	(0.63)	(0.17)	(0.14)	(2.00)
Over-Id p-value	0.58	N.A.	0.07	0.07	N.A.

Table 1.5 shows that the 2SLS over-identification tests, when available, are indeed significant at 10% level, rejecting the null hypothesis that all these instruments are valid, which in turn rejects the conditional mean independence assumption $\mathbb{E}_{U|XZ}(X, Z) = 0$. From the estimates, the results from the 2SLS are misleading: they suggest that the return to education is not monotonic. In Columns (2)-(4), it first increases then decreases in the level of education. In Column (5), it decreases first and then increases to a level that is outside the range of Y in the sample ($\max(Y_i) = 7.78$). In contrast, the over-identification test is insignificant in our approach because we only need it to hold locally in x_0, and our results are consistent with the literature on returns to education.

This example shows that although our approach relies on stronger assumptions than the standard IV approach, when the latter is not possible due to the failure of the order condition, our approach may still obtain informative results. By contrast, alternative approaches that make stronger assumptions on exogeneity with respect to the outcome heterogeneity may not work well in some applications.

1.5.4 When Does Matching Fail?

In this section we illustrate two possibilities where a matching point does not exist. The first case is that the covariate only matches one propensity score at a time. We use the IQ score in place of parents' education for illustration. The second possibility is that

the IV has dominant effects on the propensity scores such that the covariates cannot compensate for the shift in the IV. Using the HSIS dataset, we illustrate it by examining the impact of a randomly assigned lottery granting access to the Head Start preschool program compared to the impacts of other covariates.

Covariates Too Few

Recall that when D is three-valued, there are two propensity scores to be matched. One covariate may fail to match both even if it has large effects on each of them.

For illustration, we keep the setup in Section 1.5.2 but replace parents' education with the IQ score; IQ is a reasonable candidate for the matching point because it is likely to affect both an individual's educational attainment and her wage. However, as shown in the following figure, it is unable to generate a matching point that match all the propensity scores.

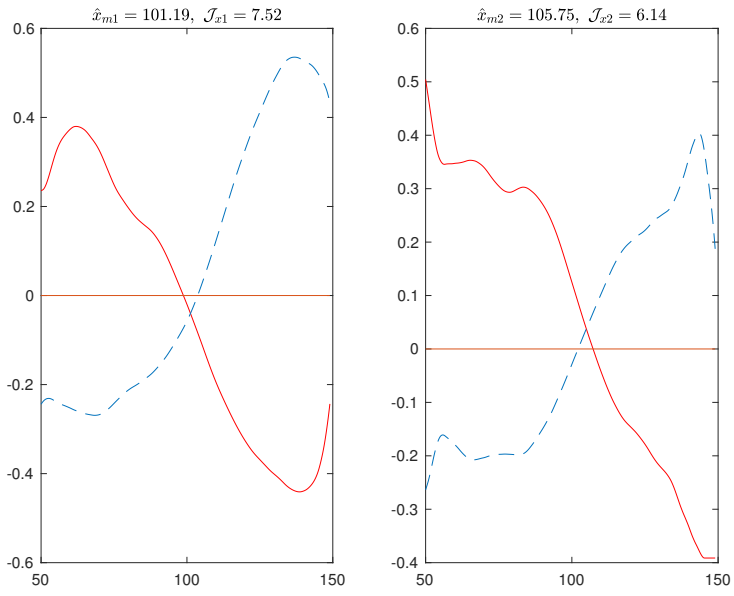

Figure 1.6: Propensity Score Differences: $X = \text{IQ}$, $x_0 = \text{med}(\text{IQ})$

In Figure 1.6, the solid red curves are $\hat{p}_1(x,1) - \hat{p}_1(x_0,0)$ and $\hat{p}_1(x,0) - \hat{p}_1(x_0,1)$, and the dashed blue curves are $\hat{p}_3(x,1) - \hat{p}_3(x_0,0)$ and $\hat{p}_3(x,0) - \hat{p}_3(x_0,1)$. All the four curves

indeed intersect with 0, so a solution does exist for each propensity score matching equation. The problem is that the intersection points do not coincide, so the two propensity score differences cannot be 0 at the same time. Indeed, the individual over-identification test statistics reject the null that both propensity scores are matched at 1% and 5% level. This type of matching failure is likely to be resolved by using more covariates that also have large effects on the propensity scores for matching.

Covariates Too Weak

Another reason for matching failure is that the effects of Z on the propensity scores dominate those of X, making it difficult for the covariate to compensate the change in Z within its support. The extreme of this scenario is that no covariates enter the selection model, and matching points obviously do not exist. For illustration, let us consider an application on preschool program selection, following Kline and Walters (2016) using the HSIS dataset. D takes on three values: participating in Head Start (h), participating in another competing preschool program (c), and not participating in any preschool programs (n). The binary instrument Z is a lottery granting access to Head Start. Available candidates for X are family income, baseline test scores and the centers' quality index.

Figure 1.7 shows the propensity scores with $X =$ the baseline test score and x_0= the sample median. Findings for other values of X and other covariates are similar. We see that when an individual won the lottery, the probability attending the Head Start program is very high, and not much affected by the baseline scores. On the contrary, when not winning the lottery, she would most likely not participate in any programs, and in particular, the probability of attending the Head Start is lowest for almost any baseline scores. Apparently a matching point does not exist in this example because varying X never offsets the dominant effect of Z on the propensity scores.

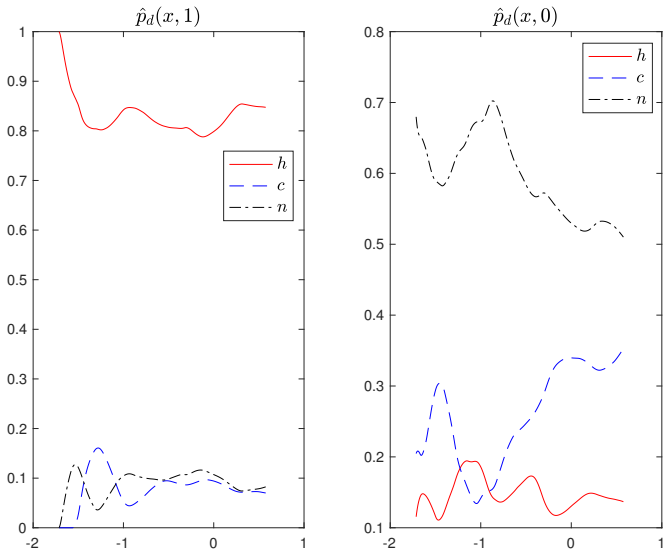

Figure 1.7: Propensity Scores of Different Preschool Program Choices

1.6 Asymptotic Properties

In this section we present the asymptotic properties of the estimators. We also discuss inference procedures and develop some specification tests.

1.6.1 The Matching Points

We start with consistency and asymptotic normality of $(\hat{x}_{m1}, \hat{x}_{m2})$. Recall that we focus on the simple benchmark case where (x_{m1}, x_{m2}) is unique and $a_n = 0$ in equation (1.4.1). The asymptotic property for the general case is presented in Appendix A.1.2. We make the following regularity conditions.

Assumption Reg-MP. *For every $d \in S(D)$ and $z \in \{0,1\}$, $p_d(\cdot, z)$ is twice continuously differentiable on $S_0(X)$ with bounded derivatives. The density of X exists and is bounded away from 0 on $S_0(X)$.*

Assumption Reg-K. *The kernel $K(\cdot)$ is symmetric at 0 with finite second moment and twice continuously bounded derivatives on $[-1,1]$.*

Under the regularity conditions, $\hat{Q}_x$ is uniformly consistent for Q_x. Consistency and asymptotic normality follow from the standard argument for GMM estimators.

Theorem Cons-MP. *Under Assumptions Reg-MP, Reg-K and the benchmark conditions,* $|(\hat{x}_{m1}, \hat{x}_{m2}) - (x_{m1}, x_{m2})| = o_p(1)$.

Denote the gradient of $\Delta \boldsymbol{p}(x_1, x_2)$ evaluated at x_{m1} and x_{m2} by $\partial_{x'} \Delta \boldsymbol{p}(x_{m1}, x_{m2})$. Let $\tilde{z}_1, ..., \tilde{z}_4$ be $(x_0, 0)$, $(x_0, 1)$, $(x_{m1}, 1)$ and $(x_{m2}, 0)$ respectively.

Theorem AsymDist-MP. *Under the conditions in Theorem Cons-MP, if* (x_{m1}, x_{m2}) *is in the interior of* $S_0(X)$, $\Pi_x \equiv \partial_{x'} \Delta \boldsymbol{p}(x_{m1}, x_{m2}) \boldsymbol{W}_x \partial_x \Delta \boldsymbol{p}(x_{m1}, x_{m2})$ *is nonsingular, and* $h_x^2 \cdot \sqrt{nh_x} = o(1)$, *we have*

$$\sqrt{nh_x} \begin{pmatrix} \hat{x}_{m1} - x_{m1} \\ \hat{x}_{m2} - x_{m2} \end{pmatrix} \xrightarrow{d} \mathcal{N}(0, \Pi_x^{-1} \partial_{x'} \Delta \boldsymbol{p}(x_{m1}, x_{m2}) \boldsymbol{W}_x \Sigma_x \boldsymbol{W}_x \partial_x \Delta \boldsymbol{p}(x_{m1}, x_{m2}) \Pi_x^{-1}) \quad (1.6.1)$$

where $\Sigma_x = \kappa \begin{pmatrix} \Sigma_{x1} & \mathbf{0} \\ \mathbf{0} & \Sigma_{x2} \end{pmatrix}$, $\kappa \equiv \int v^2 K(v) dv$, *and*

$$\Sigma_{x1} = \begin{pmatrix} \frac{p_1(\tilde{z}_3)(1-p_1(\tilde{z}_3))}{f_{XZ}(\tilde{z}_3)} + \frac{p_1(\tilde{z}_1)(1-p_1(\tilde{z}_1))}{f_{XZ}(\tilde{z}_1)} & -\frac{p_1(\tilde{z}_3)p_2(\tilde{z}_3)}{f_{XZ}(\tilde{z}_3)} - \frac{p_1(\tilde{z}_1)p_2(\tilde{z}_1)}{f_{XZ}(\tilde{z}_1)} \\ -\frac{p_1(\tilde{z}_3)p_2(\tilde{z}_3)}{f_{XZ}(\tilde{z}_3)} - \frac{p_1(\tilde{z}_1)p_2(\tilde{z}_1)}{f_{XZ}(\tilde{z}_1)} & \frac{p_2(\tilde{z}_3)(1-p_2(\tilde{z}_3))}{f_{XZ}(\tilde{z}_3)} + \frac{p_2(\tilde{z}_1)(1-p_2(\tilde{z}_1))}{f_{XZ}(\tilde{z}_1)} \end{pmatrix},$$

$$\Sigma_{x2} = \begin{pmatrix} \frac{p_1(\tilde{z}_4)(1-p_1(\tilde{z}_4))}{f_{XZ}(\tilde{z}_4)} + \frac{p_1(\tilde{z}_2)(1-p_1(\tilde{z}_2))}{f_{XZ}(\tilde{z}_2)} & -\frac{p_1(\tilde{z}_4)p_2(\tilde{z}_4)}{f_{XZ}(\tilde{z}_4)} - \frac{p_1(\tilde{z}_2)p_2(\tilde{z}_2)}{f_{XZ}(\tilde{z}_2)} \\ -\frac{p_1(\tilde{z}_4)p_2(\tilde{z}_4)}{f_{XZ}(\tilde{z}_4)} - \frac{p_1(\tilde{z}_2)p_2(\tilde{z}_2)}{f_{XZ}(\tilde{z}_2)} & \frac{p_2(\tilde{z}_4)(1-p_2(\tilde{z}_4))}{f_{XZ}(\tilde{z}_4)} + \frac{p_2(\tilde{z}_2)(1-p_2(\tilde{z}_2))}{f_{XZ}(\tilde{z}_2)} \end{pmatrix}.$$

It is easy to verify that the optimal weighting matrix that achieves the smallest asymptotic variance given (1.6.1) is $\boldsymbol{W}_x^* = \Sigma_x^{-1}$. It can be estimated by adopting the standard two-step or multiple-step GMM approach. Denote the estimator using the estimated

optimal weighting matrix by $(\hat{x}^*_{m1}, \hat{x}^*_{m2})$, it is straightforward that

$$\sqrt{nh_x}\begin{pmatrix}\hat{x}^*_{m1} - x_{m1}\\ \hat{x}^*_{m2} - x_{m2}\end{pmatrix} \xrightarrow{d} \mathcal{N}\big(0, (\partial_{x'}\Delta\boldsymbol{p}(x_{m1}, x_{m2})\Sigma_x^{-1}\partial_x\Delta\boldsymbol{p}(x_{m1}, x_{m2}))^{-1}\big) \tag{1.6.2}$$

Note that in this benchmark case, only a single covariate is present, so (x_{m1}, x_{m2}) is over-identified. The null hypothesis $\mathbb{H}_0 : \Delta\boldsymbol{p}(x_{m1}, x_{m2}) = \mathbf{0}$ can be tested by an over-identification test. For example, consider J-test $\mathcal{J}_x = nh_x\Delta\hat{\boldsymbol{p}}(\hat{x}^*_{m1}, \hat{x}^*_{m2})'\widehat{\Sigma}_x^{-1}\Delta\hat{\boldsymbol{p}}(\hat{x}^*_{m1}, \hat{x}^*_{m2})$. Under the null, it can be verified that $\mathcal{J}_x \xrightarrow{d} \chi^2_2$.

In addition to jointly testing whether (x_{m1}, x_{m2}) solves the propensity score matching equations, we can separately test either one of them if needed. By block-diagonality of the asymptotic variance in equation (1.6.2), $\hat{x}^*_{m1}$ and $\hat{x}^*_{m2}$ are asymptotically independent, and thus it is equivalent to estimate them separately. In each separate problem the matching point is still over-identified, so let

$$\mathcal{J}_{x1} = nh_x\Delta\hat{\boldsymbol{p}}(\hat{x}^*_{m1})'\widehat{\Sigma}_{x1}^{-1}\Delta\hat{\boldsymbol{p}}(\hat{x}^*_{m1}), \tag{1.6.3}$$

and

$$\mathcal{J}_{x2} = nh_x\Delta\hat{\boldsymbol{p}}(\hat{x}^*_{m2})'\widehat{\Sigma}_{x2}^{-1}\Delta\hat{\boldsymbol{p}}(\hat{x}^*_{m2}), \tag{1.6.4}$$

Under the null, each of two test statistics converges in distribution to χ^2_1.

1.6.2 The SP-Outcome Functions

From equation (1.4.3), consistency of $\hat{\boldsymbol{m}}(x_0)$ directly follows from consistency of each component in its formula, guaranteed by the following regularity conditions.

Assumption Reg-SP. *For every d, z, x, $Y|d, x, z$ has finite second moment. $\mathbb{E}_{Y|DXZ}(d, \cdot, z)$ is twice continuously differentiable on $S(X)$ with bounded derivatives.*

Theorem Cons-SP. *Under the conditions in Theorem ID-SP, Assumptions Reg-MP, Reg-SP, Reg-K, and the benchmark conditions,* $\hat{\boldsymbol{m}}(x_0) - \boldsymbol{m}^*(x_0) = o_p(1)$.

For the asymptotic distribution, we let $h_m/h_x \to 0$ so that the impacts of estimating (x_{m1}, x_{m2}) and the propensity scores are negligible. Let $\tilde{\boldsymbol{z}}_1, ..., \tilde{\boldsymbol{z}}_6$ be $(x_0, 0)$, $(x_{m1}, 0)$, $(x_{m1}, 1)$, $(x_0, 1)$, $(x_{m2}, 1)$ and $(x_{m2}, 0)$. We have

Theorem AsymDist-SP. *Under the conditions in Theorem Cons-SP, suppose* $h_m/h_x \to 0$ *where* h_x *satisfies the conditions in Theorem AsymDist-MP, then*

$$\sqrt{nh_m}(\hat{\boldsymbol{m}}(x_0) - \boldsymbol{m}(x_0)) \xrightarrow{d} \mathcal{N}(0, (\Pi_{SP}'\boldsymbol{W}_m\Pi_{SP})^{-1}\Pi_{SP}'\boldsymbol{W}_m\Sigma_{SP}\boldsymbol{W}_m'\Pi_{SP}(\Pi_{SP}'\boldsymbol{W}_m\Pi_{SP})^{-1}) \tag{1.6.5}$$

where $\Sigma_{SP} = \kappa(\Sigma_{SP,1} + \Sigma_{SP,2} + \Sigma_{SP,3})$ *and* $\Sigma_{SP,d}$ $(d = 1, 2, 3)$ *equals*

$$\begin{pmatrix} \frac{p_d(\tilde{z}_1)^2\mathbb{V}_{Y|DXZ}(d,\tilde{z}_1)}{f_{DXZ}(d,\tilde{z}_1)} & 0 & \frac{p_d(\tilde{z}_1)p_d(\tilde{z}_2)\mathbb{V}_{Y|DXZ}(d,\tilde{z}_1)}{f_{DXZ}(d,\tilde{z}_1)} & 0 \\ 0 & \frac{p_d(\tilde{z}_4)^2\mathbb{V}_{Y|DXZ}(d,\tilde{z}_4)}{f_{DXZ}(d,\tilde{z}_4)} & 0 & \frac{p_d(\tilde{z}_4)p_d(\tilde{z}_5)\mathbb{V}_{Y|DXZ}(d,\tilde{z}_4)}{f_{DXZ}(d,\tilde{z}_4)} \\ \frac{p_d(\tilde{z}_1)p_d(\tilde{z}_2)\mathbb{V}_{Y|DXZ}(d,\tilde{z}_1)}{f_{DXZ}(d,\tilde{z}_1)} & 0 & \sum_{k=1}^{3}\frac{p_d(\tilde{z}_2)^2\mathbb{V}_{Y|DXZ}(d,\tilde{z}_k)}{f_{DXZ}(d,\tilde{z}_k)} & 0 \\ 0 & \frac{p_d(\tilde{z}_4)p_d(\tilde{z}_5)\mathbb{V}_{Y|DXZ}(d,\tilde{z}_4)}{f_{DXZ}(d,\tilde{z}_4)} & 0 & \sum_{k=4}^{6}\frac{p_d(\tilde{z}_5)^2\mathbb{V}_{Y|DXZ}(d,\tilde{z}_k)}{f_{DXZ}(d,\tilde{z}_k)} \end{pmatrix}$$

Again, the optimal weighting matrix is $\boldsymbol{W}_m^* = \Sigma_{SP}^{-1}$. Note that Σ_{SP} does not depend on $\boldsymbol{m}^*(x_0)$, so it can be directly estimated by plugging in the estimated matching points and consistent estimators for the conditional variances and densities. Denote the estimator under the estimated optimal weighting matrix by $\hat{\boldsymbol{m}}^*(x_0)$, then

$$\sqrt{nh_m}(\hat{\boldsymbol{m}}^*(x_0) - \boldsymbol{m}^*(x_0)) \xrightarrow{d} \mathcal{N}\left(0, (\Pi_{SP}'\Sigma_{SP}^{-1}\Pi_{SP})^{-1}\right) \tag{1.6.6}$$

A consistent estimator of the asymptotic variance in equation (1.6.6) is straightforward to compute. Alternatively, bootstrap inference can be implemented by fixing $\hat{x}_{m1}$, $\hat{x}_{m2}$ and $\hat{\boldsymbol{p}}$ and only re-estimate the conditional expectations in each bootstrap sample.

As with the matching points, the over-identifying restrictions can be tested; construct

the test statistic:

$$\mathcal{J}_{SP} = nh_m\big(\widehat{\Pi}_{SP}(\hat{x}_{m1},\hat{x}_{m2})\hat{\boldsymbol{m}}^*(x_0) - \widehat{\Phi}_{SP}(\hat{x}_{m1},\hat{x}_{m2})\big)'\widehat{\Sigma}_m^{-1}\big(\widehat{\Pi}_{SP}(\hat{x}_{m1},\hat{x}_{m2})\hat{\boldsymbol{m}}^*(x_0) - \widehat{\Phi}_{SP}(\hat{x}_{m1},\hat{x}_{m2})\big)$$

Under the null that all the moment conditions hold, $\mathcal{J}_{SP} \xrightarrow{d} \chi^2_1$.

It is worth noting that the test in our approach not only examines exogeneity of Z, but also jointly tests whether PSC holds at the selected conditioning points and whether all the conditions in Definition MP are satisfied.

1.6.3 The NSP-Outcome Functions

In this subsection, we provide sufficient conditions that deliver uniform consistency and asymptotic normality of $\hat{\boldsymbol{g}}(x_0,u)$. It turns out that monotonicity of $\boldsymbol{g}^*(x_0,\cdot)$ and the structure of our estimator simplify the general theory of sieve estimators (e.g. Chen and Pouzo (2012, 2015)); simple low level conditions suffice.

Let us begin by establishing the following key condition for consistency: For any closed interval $\mathcal{U}_0$ in the interior of $[0,1]$,

$$\inf_{\substack{\boldsymbol{g}\in\mathcal{G}_0:\\ \sup_{u\in\mathcal{U}_0}|\boldsymbol{g}(u)-\boldsymbol{g}^*(x_0,u)|\geq\delta}} \Big|\int_0^1 Q_{NSP}(\boldsymbol{g}(u),u)du - \int_0^1 Q_{NSP}(\boldsymbol{g}^*(x_0,u),u)du\Big| > 0 \qquad (1.6.7)$$

When $\boldsymbol{g}^*$ is the unique solution to $Q_{NSP}(\cdot)=0$, inequality (1.6.7) holds if $\mathcal{G}$ is compact in sup-norm (see for instance Newey and McFadden (1994)). It is common that function spaces are not compact. Thus for sieve estimators, zero on the right hand side of equation (1.6.7) is usually replaced by a positive sequence that converges to zero. We show that for the space $\mathcal{G}_0$ of uniformly bounded monotonic functions on a compact domain, inequality (1.6.7) does hold under Theorem ID-NSP.

Theorem ID-Sup. *Under the conditions in Theorem ID-NSP, inequality (1.6.7) is true.*

Consistency of $\hat{\boldsymbol{g}}$ in sup-norm then follows from the uniform convergence of $\hat{Q}_{NSP}$

and the existence of an element in the sieve space $\hat{G}$ that converges to $g^*(x_0, \cdot)$ in sup-norm (see Chen (2007), Chen and Pouzo (2012, 2015), etc). The latter is straightforward because any continuous increasing function can be approximated by piecewise affine increasing functions arbitrarily well provided that the number of the nodes is sufficiently large. The former holds if the proposed CDF estimators are uniformly consistent, guaranteed by the following regularity assumptions.

Assumption Reg-NSP. *For every $d \in S(D)$ and $z \in \{0, 1\}$, $F_{Y|DXZ}(\cdot|d, \cdot, z)$ is twice continuously differentiable on the support with bounded derivatives. The conditional density $f_{Y|DXZ}(\cdot|d, \cdot, z)$ is continuous and uniformly bounded away from 0 over the support for all d and z.*

Assumption Reg-L. *$L(\cdot)$ is a continuously differentiable CDF supported on $[-1, 1]$ with bounded derivatives.*

Theorem Cons-NSP. *Under the conditions in Theorem ID-NSP, Assumptions Reg-L, Reg-K, Reg-MP, Reg-NSP and the benchmark conditions, if $J \to \infty$ and $\lambda \cdot J \to 0$,*

$$\sup_{u \in \mathcal{U}_0} \left| \hat{g}(x_0, u) - g^*(x_0, u) \right| = o_p(1). \tag{1.6.8}$$

In particular, if $S(Y|d, x) = S(Y|d)$ for $x = x_0, x_{m1}, x_{m2}$, equation (1.6.8) holds for $\mathcal{U}_0 = [0, 1]$.

Now let us derive the asymptotic distribution of $\hat{g}(x_0, u_0) - g^*(x_0, u_0)$ for a fixed u_0. Recall that by construction, $\underline{y} \leq \hat{g}(x_0, u_1) \leq \hat{g}(x_0, u_2) \leq ... \leq \hat{g}(x_0, u_J) \leq \bar{y}$. If all these inequalities are strict, the estimator at each node satisfies the first order condition due to the smoothness of the CDF estimators and the penalty function.

Theorem Cons-NSP implies that for large enough n, $\hat{g}(x_0, u_j)$ for all $u_j \in \mathcal{U}_0$ are uniformly close to the true function values at the corresponding nodes. Meanwhile, the differences between $g^*(x_0, \cdot)$ at adjacent nodes converge to 0 because $|g^*(x_0, u) - g^*(x_0, u \pm 1/J)| = O(1/J)$ if $g^*(x_0, \cdot)$ is differentiable and its derivative is bounded away

from o. Therefore, for any node $u \in \mathcal{U}_0$, by the triangle inequality, if $\hat{g}(x_0, u)$ converges to $g(x_0, u)$ faster than $1/J$, for large enough n, $\hat{g}(x_0, u)$ is strictly greater than $\hat{g}(x_0, u - 1/J)$ and strictly smaller than $\hat{g}(x_0, u + 1/J)$. The following theorem provides the uniform rate of convergence of $\hat{g}(x_0, \cdot)$ at each node in $\mathcal{U}_0$.

Theorem RoC-NSP (Rate of Convergence). *Let* $r_n = \sqrt{\log(n)/nh_g} + h_g$. *Suppose* $h_g/h_x \to 0$, $h_0/h_g \to 0$, *and* $\lambda = o(r_n^2)$. *Under all the conditions in Theorem Cons-NSP,*

$$\max_{u_j \in \mathcal{U}_0} |\hat{g}(x_0, u_j) - g^*(x_0, u_j)| = O_p(\sqrt{J} r_n) \tag{1.6.9}$$

Corollary RoC-NSP. *Under the conditions in Theorem RoC-NSP, suppose* $J \cdot \sqrt{J} r_n \to 0$. *If the derivative of* $g_d^*(x_0, \cdot)$ *is bounded away from* 0 *on* $\mathcal{U}_0$ *for all* d, *then* $\hat{g}(x_0, \cdot)$ *on the nodes in* $\mathcal{U}_0$ *are strictly increasing with probability approaching one.*

Remark 5.1. *The order in equation* (1.6.9) *the square root of the uniform convergence rate of* $J\hat{Q}_{NSP}$. *Note that the bias is of the order of* h_g, *instead of* h_g^2 *in the standard case. This is because of the nonsmoothness of* $F_{Y|DXZ}(\cdot|d, x, z)$ *at the boundaries; symmetry of the kernel function cannot be utilized because of such nonsmoothness, slowing down the uniform rate.*

Note that under Corollary RoC-NSP, none of the inequality constraints are binding. $\hat{g}(x_0, \cdot)$ at the nodes are thus asymptotically equivalent as the unconstrained pointwise estimator described in Section 1.4 under global identification pointwise at each node.

Let $\tilde{z}_1, ..., \tilde{z}_6$ be $(x_0, 0)$, $(x_{m1}, 0)$, $(x_{m1}, 1)$, $(x_0, 1)$, $(x_{m2}, 1)$ and $(x_{m2}, 0)$. Let $\delta = \mathbb{1}(Y \leq g_D^*(X, u_0))$, $\phi_{d1} = \frac{f_{Y|DXZ}(g_d^*(x_{m1}, u_0)|d, x_{m1}, 0)}{f_{Y|DXZ}(g_d^*(x_{m1}, u_0)|d, x_{m1}, 1)}$, and $\phi_{d2} = \frac{f_{Y|DXZ}(g_d^*(x_{m2}, u_0)|d, x_{m2}, 1)}{f_{Y|DXZ}(g_d^*(x_{m2}, u_0)|d, x_{m2}, 0)}$.

Theorem AsymDist-NSP. *Under all the conditions in Corollary RoC-NSP, if* $J^{3/2} r_n \to 0$,

$h_g^2 = o(1/nh_g)$ *and* $h_0 = o(1/nh_g)$*, then for any node* $u_0 \in \mathcal{U}_0$*,*

$$\sqrt{nh_g}(\hat{\boldsymbol{g}}(u_0) - \boldsymbol{g}^*(u_0)) \xrightarrow{d}$$
$$\mathcal{N}\big(0, (\Pi_{NSP}'\boldsymbol{W}_g(u_0)\Pi_{NSP})^{-1}\Pi_{NSP}'\boldsymbol{W}_g(u_0)\Sigma_{NSP}\boldsymbol{W}_g(u_0)'\Pi_{NSP}(\Pi_{NSP}'\boldsymbol{W}_g(u_0)\Pi_{NSP})^{-1}\big) \tag{1.6.10}$$

where $\Sigma_{NSP} = \kappa(\Sigma_{NSP,1} + \Sigma_{NSP,2} + \Sigma_{NSP,3})$ *and* $\Sigma_{NSP,d}$ $(d = 1, 2, 3)$ *equals*

$$\begin{pmatrix} \frac{p_d(\tilde{z}_1)^2\mathbb{V}_{\delta|DXZ}(d,\tilde{z}_1)}{f_{DXZ}(d,\tilde{z}_1)} & 0 & \frac{p_d(\tilde{z}_1)p_d(\tilde{z}_2)\mathbb{V}_{\delta|DXZ}(d,\tilde{z}_1)}{f_{DXZ}(d,\tilde{z}_1)} & 0 \\ 0 & \frac{p_d(\tilde{z}_4)^2\mathbb{V}_{\delta|DXZ}(d,\tilde{z}_4)}{f_{DXZ}(d,\tilde{z}_4)} & 0 & \frac{p_d(\tilde{z}_4)p_d(\tilde{z}_5)\mathbb{V}_{\delta|DXZ}(d,\tilde{z}_4)}{f_{DXZ}(d,\tilde{z}_4)} \\ \frac{p_d(\tilde{z}_1)p_d(\tilde{z}_2)\mathbb{V}_{\delta|DXZ}(d,\tilde{z}_1)}{f_{DXZ}(d,\tilde{z}_1)} & 0 & \sum_{k=1}^{3}\frac{\phi_{d1}^2 p_d(\tilde{z}_2)^2\mathbb{V}_{\delta|DXZ}(d,\tilde{z}_k)}{f_{DXZ}(d,\tilde{z}_k)} & 0 \\ 0 & \frac{p_d(\tilde{z}_4)p_d(\tilde{z}_5)\mathbb{V}_{\delta|DXZ}(d,\tilde{z}_4)}{f_{DXZ}(d,\tilde{z}_4)} & 0 & \sum_{k=4}^{6}\frac{\phi_{d2}^2 p_d(\tilde{z}_5)^2\mathbb{V}_{\delta|DXZ}(d,\tilde{z}_k)}{f_{DXZ}(d,\tilde{z}_k)} \end{pmatrix}$$

The asymptotic variance has similar form as in Theorem AsymDist-SP for $\hat{\boldsymbol{m}}(x_0)$. In particular, entries in $\Sigma_{NSP,d}$ are similar to those in $\Sigma_{SP,d}$ with only two distinctions: In $\Sigma_{NSP,d}$, the conditional variances are with respect to the indicator function δ instead of Y, and there are additional factors ϕ_{d1} and ϕ_{d2} in the last two diagonal elements. The first distinction is analogous to the comparison of the variance formulas for mean regression and quantile regression. The second one is also expected due to nonlinearity in φ_d.

Similar to previous sections, the optimal weighting matrix that achieves the smallest asymptotic variance under equation (1.6.10) is $\boldsymbol{W}_g(u_0) = \Sigma_{NSP}^{-1}$, and a consistent estimator can be obtained by plugging in $\hat{\boldsymbol{g}}(x_0, \cdot)$ into the CDF and density estimators.

Now let us discuss how to obtain the second-step estimator using the feasible optimal weighting matrix. Although we can plug the optimal weighting matrix at each node into $\hat{Q}_{NSP}$ and solve the joint minimization problem again, in terms of computation, joint minimization is less favorable than individual minimization due to its high dimensionality. We adopt it because only local identification holds pointwise at each u. However, this is no longer a problem once we obtain a uniformly consistent first-step estimator $\hat{\boldsymbol{g}}$ using any weighting matrix. Uniform consistency guarantees that at any interior u_0,

$\hat{\boldsymbol{g}}(x_0, u_0)$ is arbitrarily close to $\boldsymbol{g}^*(x_0, u_0)$. Therefore, if we focus on a new parameter space that is shrinking towards $\hat{\boldsymbol{g}}(u_0)$, identification at u_0 holds in that space for large enough n:

$$\hat{\boldsymbol{g}}^*(x_0, u_0) = \arg \min_{[\hat{\boldsymbol{g}}(x_0,u_0)-\frac{1}{J}, \hat{\boldsymbol{g}}(x_0,u_0)+\frac{1}{J}]} \hat{Q}^*_{NSP}(\boldsymbol{g}(x_0, u_0), u_0) \tag{1.6.11}$$

where $\hat{Q}^*_{NSP}(\boldsymbol{g}(u_0), u_0) = \big(\widehat{\Psi}(\boldsymbol{g}(u_0)) - \boldsymbol{u}_0\big)' \widehat{\Sigma}^{-1}_{NSP} \big(\widehat{\Psi}(\boldsymbol{g}(u_0)) - \boldsymbol{u}_0\big)$.

Theorem AsymDist-Op-NSP. *Under all the conditions in Theorem AsymDist-NSP,*

$$\sqrt{nh_g}(\hat{\boldsymbol{g}}^*(x_0, u_0) - \boldsymbol{g}^*(x_0, u_0)) \xrightarrow{d} \mathcal{N}(0, (\Pi'_{NSP}\Sigma^{-1}_{NSP}\Pi_{NSP})^{-1}) \tag{1.6.12}$$

Like $\hat{\boldsymbol{m}}^*(x_0)$, the asymptotic variance of $\hat{\boldsymbol{g}}^*(x_0, u_0)$ is straightforward to estimate. Bootstrap inference would also work. It would be computationally intensive if we computed $\hat{\boldsymbol{g}}^*$ in every bootstrap sample. Instead, we only solve the minimization problem once, then estimate the linear expansion of $\hat{Q}^*_{NSP}(\hat{\boldsymbol{g}}^*(x_0, u_0), u_0)$ in each bootstrap sample. The resulting distribution approximates that of $\hat{\boldsymbol{g}}^*(x_0, u_0) - \boldsymbol{g}^*(x_0, u_0)$.

Finally, an over-identification test can be constructed by $\mathcal{J}_{NSP} = nh_g \hat{Q}^*_{NSP}(\hat{\boldsymbol{g}}^*(u_0), u_0)$. Under the null that all the moment conditions hold jointly, $\mathcal{J}_{NSP} \xrightarrow{d} \chi^2_1$.

1.7 Monte Carlo Simulations

This section illustrates the finite sample performance of the estimator for two separable models similar to the return to education example. The first model we consider has a three-valued D and a binary Z. In the second model, D is binary too. It enables us to compare our approach and the standard IV approach. We find that the extra moment conditions do not increase finite sample bias by much but largely reduce variances.

1.7.1 A Three-Valued D

Let D follow the ordered choice model in Example OC. Let the outcome variable Y be determined by the following model:

$$Y = [\gamma_1 \mathbb{1}(D=1) + \gamma_2 \mathbb{1}(D=2) + \gamma_3 \mathbb{1}(D=3)] \cdot (X+1) + U$$

where $X \sim \text{Unif}[-3,3]$, $Z \sim Ber(0.5)$, $[U,V] \sim \mathcal{N}\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$ and $X \perp\!\!\!\perp Z \perp\!\!\!\perp (U,V)$.

For the parameters, we fix $(\gamma_1, \gamma_2, \gamma_3, \kappa_1, \kappa_2) = (1.5, 3, 3.5, -0.7, 0.1)$. The parameters (α, β) govern the strength of the instrument and the covariate. In this section we present the results for the cases $(\alpha, \beta) = (0.8, 0.4)$. The value is selected for two reasons: (a) all the propensity scores are far away from 0 so that in the simulated sample, there are sufficient observations to estimate each conditional expectation and propensity score; (b) X and Z have large effects on the propensity scores. Finally, we set $\rho = 0.5$ and $x_0 = 0$. Additional simulation results for small (α, β), different ρ and different x_0 are provided in Appendix A.4.

Table 1.6 contains the results for samples of size 1000, 2000 and 3000. The number of simulation replications is set at 500. In each replication, we estimate x_{m1} and x_{m2} using grid search with 500 grid nodes. The propensity scores are estimated using the biweight kernel. The bandwidth is equal to 1.6 times the Silverman's rule of thumb. The conditional expectations are estimated using the same kernel with a smaller bandwidth. Finally, we compute the actual coverage probabilities of the confidence intervals for $\boldsymbol{m}^*(x_0)$ based on both the asymptotic variance estimator (the top value in each cell in the last three columns) and 500 bootstrap samples (the bottom value in parentheses). The coverage probabilities of over-identification tests for (x_{m1}, x_{m2}) and for $\boldsymbol{m}^*(x_0)$ are also reported.

As is shown in Table 1.6, the variance of the estimator dominates in mean squared

Table 1.6: $x_0 = 0$. $\boldsymbol{m}^*(0) = (1.5, 3, 3.5)$.

	N	Average	Bias^2	Variance	MSE	90%	95%	99%
$\hat{m}_1(0)$	1000	1.49	2e-4	0.12	0.12	90.2% (86.8%)	95.4% (92.6%)	99% (97.6%)
	2000	1.51	3e-5	0.06	0.06	91.6% (88.4%)	96% (94%)	99% (99%)
	3000	1.49	4e-5	0.04	0.04	88.4% (88%)	94.4% (93.8%)	99% (97.6%)
$\hat{m}_2(0)$	1000	2.89	0.01	0.78	0.79	93.2% (88.8%)	96.2% (92.6%)	99% (96.6%)
	2000	2.88	0.01	0.37	0.39	89.6% (88.2%)	95% (93.6%)	99% (98.4%)
	3000	2.92	0.01	0.25	0.26	89.2% (88.6%)	95.6% (93.2%)	99.8% (98.4%)
$\hat{m}_3(0)$	1000	3.47	0.001	0.22	0.22	92.8% (88.2%)	97% (93.4%)	98.6% (96.8%)
	2000	3.49	2e-4	0.12	0.12	92.2% (90.4%)	97.2% (95.8%)	98.8% (98.8%)
	3000	3.49	1e-4	0.07	0.07	92.6% (89.6%)	97.2% (95%)	99.4% (98.6%)
$\mathcal{J}_x$	1000					90%	95%	99.2%
	2000					91.6%	95.8%	99.6%
	3000					92.6%	96.8%	99.2%
$\mathcal{J}_{SP}$	1000					91.6%	94.8%	98.2%
	2000					93.6%	96.4%	98.4%
	3000					91.6%	96.2%	98.8%

error (MSE) due to undersmoothing. The actual coverage probabilities are overall close to the nominal values. Bootstrap confidence intervals tend to undercover the true parameters while the asymptotic confidence intervals tend to be conservative.

1.7.2 A Binary D

We modify the data generating process in Section 1.7.1 to make D binary:

$$Y = [\gamma_1 \mathbb{1}(D = 0) + \gamma_2 \mathbb{1}(D = 1)] \cdot (X + 1) + U$$

$$D = \mathbb{1}(V \geq \kappa + \alpha Z + \beta X)$$

Table 1.7: Binary D

x_0		Matching	Bias^2	Variance	MSE	90%	95%	99%
0	$\hat{m}_0(x_0)$	No	8e-4	0.07	0.07	91%	96%	99.2%
		Yes	6e-4	0.03	0.03	88%	94.4%	99.2%
	$\hat{m}_1(x_0)$	No	0.001	0.03	0.04	90.8%	95.6%	99.8%
		Yes	0.002	0.02	0.02	90.4%	94.8%	99.4%
-0.3	$\hat{m}_0(x_0)$	No	0.001	0.09	0.09	92.6%	95.8%	99.6%
		Yes	7e-4	0.04	0.04	89.6%	95.2%	99%
	$\hat{m}_1(x_0)$	No	8e-4	0.03	0.03	90%	95.6%	99.6%
		Yes	0.001	0.02	0.02	90%	94.6%	99.2%
0.3	$\hat{m}_0(x_0)$	No	5e-4	0.05	0.05	91.4%	96%	99.4%
		Yes	2e-4	0.03	0.03	90.4%	94.6%	99.4%
	$\hat{m}_1(x_0)$	No	0.002	0.03	0.04	92%	97%	99.6%
		Yes	0.002	0.02	0.02	90.4%	94.8%	99.2%

The distribution of (X, Z, U, V) is the same as in Section 1.7.1. Similarly, we set $(\gamma_1, \gamma_2, \kappa, \alpha, \beta) = (1.5, 3, -0.7, 0.8, 0.4)$. The correlation coefficient of U and V is 0.5.

The results are presented in Table 1.7. The third column indicates whether the matching points are estimated and used. We can see that adding more moment conditions from the matching points does not have much impact on the bias, but reduces the variance of the estimator. Meanwhile, the coverage probabilities are close to the nominal ones in all cases.

1.8 Relation to the Existing Literature

We discuss the most relevant approaches to identify nonparametric models with endogeneity in this section. The discussion is based on whether a selection model is explicitly exploited for identification of the outcome function.

1.8.1 Triangular Models

Triangular models are widely employed in the control function approach (e.g. Newey, Powell and Vella (1999), Chesher (2003), Florens et al. (2008) and Imbens and Newey

(2009), etc.). This approach allows the outcome heterogeneity $\boldsymbol{U}$ to be multidimensional. On the other hand, D has to be continuous; these papers assume that V in the selection function is a scalar and $h(\boldsymbol{X}, Z, \cdot)$ is strictly increasing. Inverting $h(\boldsymbol{X}, Z, \cdot)$, the distribution of V can be traced out by $F_{D|\boldsymbol{XZ}}$. A "control variable" can be constructed, conditional on which endogeneity is eliminated.

Like this paper, D'Haultfœuille and Février (2015) and Torgovitsky (2015) study nonparametric identification in triangular models with a binary Z. As with the control function approach, they require D to be continuous and $h(\boldsymbol{X}, Z, \cdot)$ are strictly increasing. On the other hand, due to the small variation of the IV, they do not allow the unobservable in the outcome function to be multidimensional. This is more restrictive than the typical control function approach and the same as our paper. Gunsilius (2018) extends the model to allow for multidimensional heterogeneity, but the endogenous variable still has to be continuous.

Huang, Khalil and Yıldız (2019) consider a special additively separable triangular model where there are endogenous variables, denoted by D_1 and D_2, and a single IV Z. The triangular structure is with respect to one of the endogenous variables, for example D_1, in the sense that D_1 has a first stage as a function of Z and a scalar unobservable V. Specifically, the outcome variable Y is determined by $Y = m^*(D_1, D_2) + U$, and D_1 follows the equation $D_1 = h(Z, V)$. The benchmark case they focus on is that $m^*(D_1, D_2) = m_0^*(D_1) + \gamma D_2$. They follow the control function approach to construct a control variable for V, so $h(Z, \cdot)$ is assumed to be strictly increasing and D_1 needs to be continuously distributed (at least on a subset of its support).

One of the main contributions in this paper is that we allow for discrete D and dispense with monotonicity in the selection model. Although we are unable to trace out the entire distribution of $\boldsymbol{V}$ to construct a "control variable", the propensity scores provide useful information on the partitions of $S(\boldsymbol{V})$. Based on this, we find the matching points $\boldsymbol{x}_m$ that have the same level of endogeneity as $\boldsymbol{x}_0$. Endogeneity is eliminated by

comparing the distribution or the mean of their outcomes.

1.8.2 Single Equation Approaches

Single equation approaches refer to methods that achieve identification without relying on the structures of the first stage. Perhaps the most important example is the standard IV approach for nonparametric identification (e.g. Newey and Powell (2003), Chesher (2004), Das (2005), Matzkin (2003), Chernozhukov and Hansen (2005), Chernozhukov, Imbens and Newey (2007), Chen et al. (2014), etc.). As in this paper, the outcome function is usually assumed to be strictly increasing in the scalar unobservable. Typically, this approach requires Z to have large support.

Caetano and Escanciano (2018) develop a new identification strategy that achieves identification using small-support Z when D is multivalued. Their method does not rely on selection models. Similar to this paper, they utilize the variation in $\boldsymbol{X}$ for identification. But the strategy is different from this paper. Taking the nonseparable model as an example, their model essentially has a single index structure: $Y = g_D^*(U)$ and $U = \phi(\boldsymbol{X}, U_0)$. ϕ is a real-valued function and $g_D^*(\cdot)$ is strictly increasing a.s. Note that differently from our approach, they restrict the way in which the covariates enter the model. By contrast, we allow all the covariates to enter the model in arbitrary ways, but we need a selection model. Hence, their approach and ours are complementary.

1.9 Concluding Remarks

In this paper, we develop a novel approach to use covariates to identify separable and nonseparable models in a triangular system when the discrete endogenous variable takes on more values than the IV. This paper illustrates that information on endogenous selection has large identifying power. By tracing out the selection patterns across different values of the covariates and the IV, individuals that differ in observables but

have the same selection pattern may then be identified. Extrapolations can thus be made across them as they have the same degree of endogeneity, supplementing the insufficient information from the IV.

Moving forward, it would be of interest to extend the approach in this paper. In practice, the outcome variable is often limited, for example censored or truncated. Generalizing the approach to allow for such outcome variables would have a wide application. Another direction is to generalize the outcome function by allowing multidimensional heterogeneity, especially for the nonseparable model. For estimation, it would be interesting to investigate the optimal selection of matching points when they are uncountably infinite. It would also be useful to derive the optimal bandwidth choice for the multi-step local GMM or sieve estimation procedures proposed in this paper.

Chapter 2

Robust Principal Component Analysis with Non-Sparse Errors

Jushan Bai and Junlong Feng

2.1 Introduction

A low-rank component in high dimensional data sets is often the object of interest. In asset return analysis, for example, a low-rank matrix represents systematic risks (Ross, 1976). In psychology, the main personality traits form a low-rank matrix (e.g., Cattell (1978) and Goldberg (1990)). In background/foreground separation, by stacking the pixels of each frame of a video in a column vector, the static background is a rank one component in the resulting matrix because it stays unchanged across frames; see Bouwmans et al. (2017) for a survey. In gene expression prediction, the gene expression values may form a low-rank matrix because genes act in groups and at the expression levels they are interdependent (Kapur, Marwah and Alterovitz, 2016).

To fix ideas, in this paper, we assume the data matrix, Y, is of $N \times T$ dimension and consists of a low-rank component L_0, namely

$$Y = L_0 + Z_0 \tag{2.1.1}$$

where $rank(L_0) = r$ is small but unknown; L_0 can be random or deterministic. The magnitude of the its elements are allowed to diverge to infinity with N and T grow; Z_0 is a random error matrix with median zero entries that have positive densities around 0.

We propose to estimate L_0 using a variant of the Principal Component Pursuit (PCP), introduced and studied by Candès et al. (2011), Chandrasekaran et al. (2011) *etc*. We show the estimator is consistent for L_0 in the Frobenius norm under certain conditions. This is the first time that consistency is established with continuous and potentially fat-tailed random errors. Formally, the estimator $\hat{L}$ is as follows

$$\hat{L} = \arg\min_{L} \ ||L||_* + \lambda ||Y - L||_1, \qquad s.t. \ ||L||_\infty \leq \alpha \tag{2.1.2}$$

where $||\cdot||_*$ denotes the nuclear norm. $||\cdot||_1$ and $||\cdot||_\infty$ are the ℓ_1 norm and the ℓ_∞

norm of a matrix seen as a long vector. Both λ and α are N, T dependent. In particular, α can grow with N and T. We call $\hat{L}$ the *Bounded Principal Component Pursuit* (BPCP) as compared to the standard PCP, it has an additional constraint bounding the max entry of the estimator by α.

As a preview of how the estimator works, first note that the nuclear norm is the convex envelope of the rank of a matrix because by definition, the nuclear norm is the ℓ_1 norm of the vector of singular values while the rank is its ℓ_0 norm, that is, the number of nonzero elements in the vector. Thus, minimizing the nuclear norm is a convex surrogate of rank minimization.

The other term in the objective function is the ℓ_1 penalty for the residuals to induce robustness. The intuition is analogous to the LAD (least absolute deviation) estimator for linear regression; it is well-known that minimizing the sum of absolute deviations is robust to fat-tailed errors. It turns out for the BPCP estimator to work, the errors essentially only need to have zero median, just like the standard LAD estimator.

Finally, the constraint in (2.1.2) is needed for technical reasons. We allow α to be N, T dependent and can go to infinity as N and T increase. We will be precise about the allowed rate of such divergence. As it turns out, the restriction is actually very mild because in many stochastic models of L_0, $||L_0||_\infty$ diverges slower than the rate allowed with high probability. Therefore, imposing the constraint is without loss of generality in these models.

There are three main findings in this paper. First, we build connections between the median zero assumption, which is often made in statistics and econometrics, and the sparsity-type assumptions, more common and plausible in machine learning. We develop a novel theoretical argument, the Bernoulli device, to explore their relationship. The Bernoulli device is able to decompose a median zero matrix into a matrix whose norm is small enough to control the estimation error, and a matrix with a sufficient number of zeros. Second, although the fraction of zeros in the latter matrix is shrinking

to zero by construction, we show that a dual certificate similar to Candès et al. (2011) and Ganesh et al. (2010) can still be built based on it. Therefore that matrix can work as the sparse error matrix needed in the standard PCP literature. Finally, we provide an upper bound for the estimation error and consistency of our BPCP estimator follows.

This paper adds to the theory of PCP and some of its variants, developed in Candès et al. (2011), Chandrasekaran et al. (2011), Ganesh et al. (2010), Zhou et al. (2010) *etc*. In Candès and Recht (2009), it is assumed that Z_0 follows a Bernoulli model, i.e., each element in Z_0 is equal to 0 with probability $1-\rho$. They show that when $1-\rho$ is large enough, PCP can *exactly* recover both L_0 and Z_0 with high probability. The proof hinges on the existence of a matrix called *dual certificate*, which relies on the sparsity of Z_0. Ganesh et al. (2010) generalize the result by allowing for an arbitrarily small but fixed $1-\rho$ and show a dual certificate still exists. However, their results are not applicable to a continuously distributed error matrix because then for any entry $Z_{0,it}$, $P(Z_{0,it}=0)=0$ by definition, and thus $1-\rho=0$. Zhou et al. (2010) study a variant of PCP called the *Stable Principal Component Pursuit* (SPCP). In their model, Y is known to consist of a low rank, a sparse, and a dense component. They minimize a similar objective function over both L and Z with a constraint bounding the difference between the sum of them and Y in the Frobenius norm. They show that the Frobenius norm of the estimation error is bounded by the Frobenius norm of the dense component multiplied by $\max\{N,T\}$. This bound is evidently too large for consistency. Hsu, Kakade and Zhang (2011) change the objective function by adding a squared Frobenius norm penalty for the difference between Y and the sum of the low-rank and the sparse components. They prove nuclear norm consistency for the low-rank matrix provided that the sparse component has an increasing fraction of zeros and the dense errors are Gaussian (see the second example in their section D). All the existing work in the above literature require Z_0 to have a certain fraction of entries to be 0 with positive probability. In contrast, in this paper all entries in Z_0 can be nonzero almost surely. Agarwal, Negahban and Wainwright (2012)

study a broad class of models allowing Y to be determined by a general operator of L_0 and Z_0, which are not necessarily to be exactly low-rank or sparse. Similar to Hsu, Kakade and Zhang (2011), their objective function also has the additional Frobenius norm penalty as they allow for the existence of an additional noise component whose operator norm is not too large. Their results are more comparable with ours because they also allow for a Z_0 with all entries nonzero. However, to obtain consistency, Z_0 needs to be approximately sparse, i.e., the fraction of the entries that are large in magnitude needs to be shrinking to 0 and the sum of the absolute values of the rest entries is $o_p(NT)$. This condition rejects many random models for Z_0, especially if Z_0 has very fat tails. By contrast, this paper only focuses on the linear decomposition model (2.1.1) and under stronger assumptions including a probabilistic model for Z_0, consistency is established even if all entries of Z_0 are nonzero and most of them are large in magnitude; we do not put any restrictions on the entries' tail distributions, so long as they have zero median and positive densities around 0.

This paper also lies in the broader literature of estimating low-rank components in various settings. The following is only a small portion of many contributions in this literature. For low-rank matrix recovery, Tao and Yuan (2011), Xu, Caramanis and Sanghavi (2012), Wong and Lee (2017) and Brahma et al. (2018) study the case where both a sparse component and a dense noise component exist besides the low-rank component in the decomposition. Wright et al. (2013) study the noiseless case but Y is a compressive measurement of $L_0 + Z_0$. Method-wise, Xu, Caramanis and Sanghavi (2012) replaces the ℓ_1 norm in the objective function with the $\ell_{1,2}$ norm. Wong and Lee (2017) changes it to the Huber loss function. Brahma et al. (2018) allows other general forms of penalty, such as SCAD penalty. To achieve consistency, they all need the noise matrix to have small norms. Xie and Xing (2014) consider the Principal Component Analysis explicitly assuming Cauchy noise under the MLE framework and their minimization problem is nonconvex. Other related topics are low-rank matrix completion and multi-task regres-

sion with low-rank coefficient matrices. Examples are Cai, Candès and Shen (2010), Candès and Recht (2009), Candès and Tao (2010), Bach (2008), Negahban and Wainwright (2011) and Chao, Härdle and Yuan (2019). Besides using nuclear norm as a convex surrogate, another approach in the literature is to directly handle the rank constraint. Bai and Li (2012) impose a factor structure on the low-rank component and estimate it using MLE. Zhu et al. (2018) provide conditions under which a general minimization problem under rank constraint has no spurious local minima by factorization. Shalev-Shwartz, Gonen and Shamir (2011) propose an efficient greedy algorithm to handle the rank constraint. Chen, Raskutti and Yuan (2019) study low-rank tensor regression by applying a projected gradient descent method (Jain, Tewari and Kar (2014) and Jain, Rao and Dhillon (2016)), which also directly uses the rank constraint.

The rest of the paper is organized as follows. Section 2.2 introduces the Bernoulli device and its key properties. Section 2.3 extends the results in Ganesh et al. (2010) by showing that a dual certificate exists even if the fraction of zero entries is decreasing to 0 slowly. Section 2.4 presents the key condition for consistency derived from the optimality condition by using the dual certificate and exploiting the complementary structure of the two matrices decomposed from Z_0 by the Bernoulli device. Section 2.5 proves the main theorem on the estimation error bound and consistency. Simulation results are demonstrated in Section 2.6. Section 2.7 concludes. Proofs of some of the lemmas are contained in the Appendix.

Notation

Throughout, $||\cdot||_*$, $||\cdot||_1$, $||\cdot||_F$ and $||\cdot||_\infty$ denote the nuclear norm, the ℓ_1 norm, the Frobenius norm, and the max norm of a matrix. $||\cdot||$ denotes the Euclidean norm of a vector, or the operator norm of a matrix or an operator. For the same matrix, $||\cdot|| \leq ||\cdot||_F \leq ||\cdot||_*$ and $||\cdot||_F \leq ||\cdot||_1$. For two generic scalars a and b, denote $a \wedge b \equiv \min\{a, b\}$ and $a \vee b \equiv \max\{a, b\}$. For any positive sequences a and b, $a \asymp b$

means there exist $0 < c_1 \leq c_2 < \infty$ such that $c_1 a \leq b \leq c_2 a$. For any matrices A and B of the same size, $A \circ B$ is the component-wise product of A and B. For any two random objects X and Y, denote independence between X and Y by $X \perp\!\!\!\perp Y$. Finally, C, C', C_1 and C_2 denote generic positive constants that may be different in different uses.

2.2 A Bernoulli Device

The object of interest in this paper is

$$\frac{1}{NT}||\hat{L} - L_0||_F^2 \tag{2.2.1}$$

where $\hat{L}$ is defined in (2.1.2). To bound this quantity and obtain consistency, we follow the idea in Candès et al. (2011), Ganesh et al. (2010) and Zhou et al. (2010) to use a *dual certificate*, a matrix which will be defined in the next section, to derive the optimality condition for (2.1.2). This condition will then yield a bound for (2.2.1). The main theoretical challenge is that in the first place, the existence of a dual certificate hinges on the existence of zero entries (with positive probability) in the error matrix, a luxury we do not have in this paper. The key idea is to decompose Z_0 into $S_0 + D_0$ in such a way that i) a large enough fraction of entries in S_0 are 0 with positive probability to guarantee the existence of a dual certificate, and ii) that fraction cannot be too large, on the other hand, so that (2.2.1) can be bounded by a function of $||D_0||_1$ that converges to 0 in probability. We begin with constructing this decomposition using a Bernoulli device.

We first introduce the following assumption.

Assumption 2.2.1. *a)* $L_0 \perp\!\!\!\perp Z_0$. *b)* $Z_{0,it}$ *are independent and* $med(Z_{0,it}) = 0$; *c) The set of densities of* $Z_{0,it}$, $\{f_{Z_{0,it}}\}_{N,T}$ *are equicontinuous and uniformly bounded away from* 0 *at* 0.

Note that equicontinuity in c) can be replaced with continuity if we strengthen the independence condition in part b) to be i.i.d. Meanwhile, the independence condition

for $Z_{0,it}$ could be potentially replaced by certain notion of weak dependence by extending the concentration inequalities for random matrices used in this paper to non-independent cases.

Under Assumption 2.2.1, let $\{\delta\}$ be a positive sequence such that $\delta \to 0$ as $N, T \to \infty$. Let $(\underline{\gamma}_{it}, \bar{\gamma}_{it})$ be a pair of constants satisfying

$$P(Z_{0,it} \geq \bar{\gamma}_{it}) = P(Z_{0,it} \leq \underline{\gamma}_{it}) = \frac{1-\delta}{2} \tag{2.2.2}$$

Assumption 2.2.1 b) and c) guarantees the existence and uniqueness of such a pair for large enough N, T while δ approaches 0.

With $\underline{\gamma}_{it}$ and $\bar{\gamma}_{it}$, let M be an $N \times T$ matrix whose entries are defined by

$$M_{it} = \mathbb{1}(\underline{\gamma}_{it} < Z_{0,it} < \bar{\gamma}_{it}) \tag{2.2.3}$$

where $\mathbb{1}(\cdot)$ is the indicator function. Then let $D_0 = M \circ Z_0$ and $S_0 = Z_0 - M \circ Z_0$ and we have

$$Z_0 = S_0 + D_0 \tag{2.2.4}$$

Under Assumption 2.2.1, D_0 and S_0 have the following properties:

1. Both S_0 and D_0 contain 0 entries and their locations are complementary;

2. By construction, $P(S_{0,it} = 0) = \delta$ for any (i, t). Meanwhile, by Assumption 2.2.1 b) and c), for small enough δ, $P(Z_{0,it} \leq 0) - P(Z_{0,it} \leq \underline{\gamma}_{it}) = \frac{\delta}{2} = -f(\tilde{\gamma}_{it})\underline{\gamma}_{it}$ by the Mean Value Theorem where $\tilde{\gamma}_{it}$ lies between $\underline{\gamma}_{it}$ and 0. Since $f_{Z_{0,it}}(0)$ is uniformly bounded away from 0 over N, T for small enough δ, there exists uniform constants $C > 0$ such that $-\underline{\gamma}_{it} < C\delta$. Similar results hold for $\bar{\gamma}_{it}$. Therefore, there exist $C' > 0$ such that $|D_{0,it}| < C'\delta$ uniformly.

3. Let $E = \text{sign}(S_0)$, the the entries in E are i.i.d. with $P(E_{it} = 0) = \delta$ and $P(E_{it} = 1) = P(Z_{0,it} \geq \bar{\gamma}_{it}) = P(Z_{0,it} \leq \underline{\gamma}_{it}) = P(E_{it} = -1) = \frac{1-\delta}{2}$.

The Bernoulli device M thus delivers a pair (D_0, S_0) that achieves the two goals described in the beginning of this section. First, as will be seen, items 2 and 3 guarantee the existence of a dual certificate under certain conditions. Second, by Hoeffding's inequality, the order of $||D_0||_1$ is no greater than $C'NT\delta^2$ in high probability provided that δ converges to 0 at an appropriate rate. Note this holds regardless of how the distribution of the original error $Z_{0,it}$ behaves except the requirements for the zero-median and the positive and continuous density at 0.

2.3 Dual Certificate

In this section, we treat S_0 as the "sparse" error matrix and show that although δ decreases to 0, a dual certificate that is similar to Ganesh et al. (2010) exists. As mentioned in the introduction, Ganesh et al. (2010) show a dual certificate exists for any small yet fixed δ. We extend their results by carefully choosing the rate of δ, λ and other constants in their proof. We closely follow their construction of the dual certificate but for completeness, we record it here and shall indicate where the construction needs to be modified to handle a shrinking δ by construction.

First we need an identification condition to guarantee L_0 to be non-sparse so that it is distinguishable from S_0. We adopt the incoherence condition in Candès and Recht (2009), Candès et al. (2011), Ganesh et al. (2010) *etc*. Besides it, as there is an additional constraint in (2.1.2), we need a condition to guarantee L_0 to be a feasible solution. Let $U\Sigma V^*$ be a singular value decomposition of L_0, where U and V are $N \times r$ and $T \times r$ matrices of left and right singular vectors and Σ is an $r \times r$ diagonal matrix with singular values in descending order on its diagonal.

Assumption 2.3.1. *There exists a constant C such that with probability approaching 1,*

$$\max_i ||U^* e_i||^2 \leq C\frac{\mu r}{N}, \ \max_t ||V^* e_t||^2 \leq C\frac{\mu r}{T} \tag{2.3.1}$$

$$||L_0||_\infty \leq \alpha \tag{2.3.2}$$

where μ, r and α can be N, T dependent. $(e_i)_i$ and $(e_t)_t$ are canonical bases of N- and T-dimensional linear spaces.

Inequality (2.3.1) in Assumption 2.3.1 is the incoherence condition, stating that the singular vectors of L_0 are well-spread. A direct and useful consequence of (2.3.1) is that

$$||UV^*||_\infty \leq C\frac{\mu r}{\sqrt{NT}} \tag{2.3.3}$$

by noticing that $||UV^*||_\infty = \max_{it} |\sum_{k=1}^r U_{ik}V_{tk}| \leq \sqrt{\sum_{k=1}^r |U_{ik}|^2} \cdot \sqrt{\sum_{k=1}^r |V_{tk}|^2} \leq C\frac{\mu r}{\sqrt{NT}}$ where the first inequality follows from the Cauchy-Schwarz inequality and the second is from (2.3.1). Here μ characterizes how coherent the singular vectors are with the canonical bases. It can be N, T dependent and is allowed to diverge to ∞. Candès and Recht (2009) provide examples where $\mu = O(\log(N \vee T))$ and one of them is the *random orthogonal model* in which the columns in U and V are sampled uniformly among all families of r orthonormal vectors independently of each other. Fan, Wang and Zhong (2018) also give an example where $\mu = O\big(\sqrt{\log(N \vee T)}\big)$.

Inequality (2.3.2) is an inclusion assumption which implies L_0 is a feasible solution with probability approaching 1. It restricts the magnitude of the maximal entry in L_0. Again α is allowed to increase to ∞ with N and T. Note that (2.3.2) and (2.3.3) imply that L_0's largest singular value $\sigma_1 \leq \frac{\alpha}{C}\frac{\sqrt{NT}}{\mu r}$ because $L_0 = \sum_{k=1}^r u_k v_k^* \sigma_k$ while $UV^* = \sum_{k=1}^r u_k v_k^*$, where u_k and v_k are the kth column of U and V, respectively.

Before we define the dual certificate, it is useful to introduce some notations.

Let Φ be the linear space of matrices

$$\Phi \equiv \{UX^* + YV^*, X \in \mathbb{R}^{T\times r}, Y \in \mathbb{R}^{N\times r}\}$$

and let its orthogonal complement be $\Phi^\perp$. Denote the linear projection onto Φ and $\Phi^\perp$ by $\mathcal{P}_\Phi$ and $\mathcal{P}_{\Phi^\perp}$, respectively. Then it can be shown that for any $N \times T$ matrix R (e.g.Candès and Recht (2009)),

$$\mathcal{P}_\Phi R = UU^*R + RVV^* - UU^*RVV^*$$

and

$$\mathcal{P}_{\Phi^\perp} R = (I - UU^*)R(I - VV^*)$$

Let Ω be the support of S_0, i.e., $\Omega = \{(i,t) : S_{0,it} \neq 0\}$. With a slight abuse of notation, we also denote the linear space of matrices supported on Ω by Ω. The projection onto this space is denoted by $\mathcal{P}_\Omega$. Specifically, for an $N \times T$ matrix R,

$$(\mathcal{P}_\Omega R)_{i,t} = \mathbb{1}\big((i,t) \in \Omega\big) \cdot R_{it} \tag{2.3.4}$$

The complement of the support set Ω is denoted by Ω^c. Let the linear space of matrices supported on it be $\Omega^\perp$ and the projection onto the space be $\mathcal{P}_{\Omega^\perp}$, defined similarly as (2.3.4).

Finally, we characterize the subgradient of $||L||_*$ and $||S||_1$ using these notations. The subgradient of $||L||_*$ evaluated at L_0 is equal to $UV^* + W$ where $\mathcal{P}_\Phi W = 0$ and $||W|| \leq 1$. Meanwhile, recall that E denotes the sign of S_0, so the subgradient of $||S||_1$ at S_0 is equal to $E + F$, where $\mathcal{P}_\Omega F = 0$ and $||F||_\infty \leq 1$.

Now we are ready to define the dual certificate W as any $N \times T$ matrix satisfying the

following conditions:

$$\begin{cases} \mathcal{P}_{\Phi} W = 0 \\ ||W|| \leq \frac{1}{2} \\ ||\mathcal{P}_{\Omega}(UV^* + W - \lambda E)||_F \leq \frac{\lambda\delta}{16} \\ ||\mathcal{P}_{\Omega^{\perp}}(UV^* + W)||_{\infty} < \frac{\lambda}{2} \end{cases} \tag{2.3.5}$$

Note this definition is very similar to equation (6) in Ganesh et al. (2010). The only important difference occurs on the right hand side of the third inequality.

Now we present the construction of W that is similar to Ganesh et al. (2010) with necessary modifications to accommodate $\delta \to 0$.

Let $W = W_L + W_S$:

- Construction of W_L. As we have a more delicate random model and more structures regarding D_0 and S_0, we need more subtle argument to justify the construction of W_L. For any (i,t), write $D_{0,it}$ and $S_{0,it}$ as follows:

$$\begin{aligned} D_{0,it} =& \Delta_{1,it} \cdot \tilde{Z}_{1,it} \\ S_{0,it} =& (1 - \Delta_{1,it}) \cdot \left(\Delta_{2,it} \cdot \tilde{Z}_{2,it} + (1 - \Delta_{2,it}) \cdot \tilde{Z}_{3,it}\right) \\ Z_{0,it} =& D_{0,it} + S_{0,it} \end{aligned}$$

  where $\Delta_{1,it} = 1 - \prod_{j=1}^{j_0}(1 - \tilde{\Delta}_{j,it})$, $\tilde{\Delta}_{j,it} \overset{i.i.d.}{\sim} Ber(q)$ such that $1 - \delta = (1-q)^{j_0}$. $\Delta_{2,it} \overset{i.i.d.}{\sim} Ber(1/2)$. $\tilde{Z}_{1,it}$, $\tilde{Z}_{2,it}$ and $\tilde{Z}_{3,it}$ follow $Z_{0,it}$'s distribution truncated between $\underline{\gamma}_{it}$ and $\bar{\gamma}_{it}$, below $\underline{\gamma}_{it}$, and above $\bar{\gamma}_{it}$, respectively. All these Bernoulli random variables and $\tilde{Z}_{1,it}$, $\tilde{Z}_{2,it}$ and $\tilde{Z}_{3,it}$ are independent. It can be verified the distribution of $(D_{0,it}, S_{0,it})$ as well as $Z_{0,it}$ are the same as the original model, so the two models are equivalent.

  Now let $\Omega_j = \{(i,t) : \tilde{\Delta}_{j,it} = 1\}$. By construction, $\Omega^c = \cup_{j=1}^{j_0}\Omega_j$. Unlike Ganesh et al. (2010) in which $j_0 = 2\log(N)$, in this paper we need j_0 to be finite such that q converges to 0 at the same rate as δ. As shown in Lemma 2.3.2, $j_0 = 4$ is sufficient.

Now let $Q_0 = 0$, and

$$Q_j = Q_{j-1} + q^{-1}\mathcal{P}_{\Omega_j}\mathcal{P}_{\Phi}(UV^* - Q_{j-1}), j = 1, 2, ..., j_0 \tag{2.3.6}$$

Finally, let $W_L = \mathcal{P}_{\Phi^\perp} Q_{j_0}$. This construction is called the *golfing scheme*; it was first developed by Gross (2011) and Gross et al. (2010) for matrix completion problems and was later extended to matrix separation problems by Candès et al. (2011).

- Construction of W_S. Let $W_S = \lambda \mathcal{P}_{\Phi^\perp} \sum_{k\geq 0}(\mathcal{P}_\Omega \mathcal{P}_\Phi \mathcal{P}_\Omega)^k E$, provided that $||\mathcal{P}_\Omega \mathcal{P}_\Phi \mathcal{P}_\Omega|| = ||\mathcal{P}_\Omega \mathcal{P}_\Phi||^2 < 1$ to guarantee the Neumann series $\sum_{k\geq 0}(\mathcal{P}_\Omega \mathcal{P}_\Phi \mathcal{P}_\Omega)^k$ is well defined and is equal to $(\mathcal{P}_\Omega - \mathcal{P}_\Omega \mathcal{P}_\Phi \mathcal{P}_\Omega)^{-1}$. As pointed out by Ganesh et al. (2010), W_S can be viewed as constructed using least squares; it has the smallest Frobenius norm among matrices $\tilde{W}$ satisfying $\mathcal{P}_\Omega \tilde{W} = \lambda E$ and $\mathcal{P}_\Phi \tilde{W} = 0$.

The following lemmas then provide sufficient conditions for $W = W_L + W_S$ to satisfy (2.3.5). The proof of Lemma 2.3.1 is omitted because it is immediately implied by Theorem 2.6 in Candès et al. (2011), stated in the Appendix. The proof of Lemma 2.3.2 follow Lemma 3 and 4 in Ganesh et al. (2010) closely, but are tailored in a way to allow $\delta \to 0$. For completeness, they are contained in the Appendix.

Lemma 2.3.1. *Suppose* $\delta \geq C\frac{\mu r \log(N \vee T)}{\varepsilon^2 (N \wedge T)}$. *Then under Assumptions 2.2.1 and 2.3.1,* $||\mathcal{P}_\Omega \mathcal{P}_\Phi||^2 \leq 1 - \delta + \varepsilon\delta$ *with high probability.*

Lemma 2.3.2. *Suppose* $N \asymp T$, $\mu > \log N$, $\frac{\mu^{11/3} r^3}{N^{1/3}} = o(1)$, $\delta \asymp \frac{\mu r}{N^{1/3}}$, $\lambda \asymp \frac{\mu^{1/3}}{N^{2/3}}$, $\varepsilon \asymp \frac{\log N}{N^{1/3}}$. *If* $j_0 = 4$, *then under Assumptions 2.2.1 and 2.3.1, with high probability*

a). $||W_L|| \leq 1/4$,

b). $||\mathcal{P}_\Omega(UV^* + W_L)||_F < \frac{\lambda\delta}{16}$,

c). $||\mathcal{P}_{\Omega^\perp}(UV^* + W_L)||_\infty < \frac{\lambda}{4}$.

d). $||W_S|| < 1/4$,

e). $||\mathcal{P}_{\Omega^\perp} W_S||_\infty < \frac{\lambda}{4}$.

Remark 2.3.1. The condition $N \asymp T$ requires N and T to diverge to infinity at the same rate. Divergence of both N and T is a theoretical feature when modeling the high-dimensional data, yet the rate condition is indeed stronger than in the existing PCP literature as most of the work do not restrict the rate at all. Relaxation can be made to some extent by tuning the rate of other parameters like δ and λ.

Remark 2.3.2. The rates for δ, λ and ε in Lemma 2.3.2 are not unique. As will be seen in the next section, the convergence rate of $\frac{1}{NT}||\hat{L} - L_0||_F^2$ will be determined only by δ, α, r and μ, and the choice of the rates in Lemma 2.3.1 yields the fastest converging δ while W satisfies (2.3.5).

Remark 2.3.3. The rank is allowed to diverge as long as $r \leq CN^{1/9}\mu^{-11/9}(\log N)^{-1}$. Note this rate is smaller than the rate allowed in Ganesh et al. (2010), which is $CN\mu^{-1}(\log N)^{-2}$. The loss is unavoidable because from the condition in Lemma 2.3.1, instead of being constant, δ now decreases to 0 so the order allowed for r is smaller. The insight is that we are trading off between sparsity of the error matrix and sparsity of the vector of the singular values of L_0. It suggests when the random error matrix is continuous, the estimator may not perform well in finite sample for L_0 with relatively high rank.

2.4 Optimality Condition

In this section, using the dual certificate W and the complementarity of the support sets of S_0 and D_0, we derive the optimality condition for the minimization problem in (2.1.2). As it turns out, the condition induces an upper bound for (2.2.1) that facilitates the analysis of consistency.

For any feasible solution L to (2.1.2), let $Z \equiv Y - L$. Let $H \equiv L - L_0$ so $Z = Z_0 - H$. By the Bernoulli device, we have the following properties: i) if we consider the difference $H_S \equiv Z - S_0$, then by construction $H_S = D_0 - H$, and ii) $\mathcal{P}_{\Omega^\perp} D_0 = D_0$. Utilizing them, we have the following lemma.

Lemma 2.4.1. *Suppose* $||\mathcal{P}_\Omega \mathcal{P}_\Phi||^2 \leq 1 - \delta + \varepsilon\delta$, $\delta \to 0$, $\varepsilon \to 0$ *as* $N, T \to \infty$, *and the dual certificate* W *satisfying* (2.3.5) *exists. For any disturbance* $(H, -H)$ *at* (L_0, Z_0), *if*

$$||L_0 + H||_* + \lambda||Z_0 - H||_1 \leq ||L_0||_* + \lambda||Z_0||_1, \tag{2.4.1}$$

then for large enough N *and* T,

$$\frac{1}{\lambda}||\mathcal{P}_{\Phi^\perp} H||_* + ||\mathcal{P}_{\Omega^\perp} H||_1 \leq 8||D_0||_1 \tag{2.4.2}$$

Proof. Since $H_S + S_0 = Z = Z_0 - H$, $||L_0 + H||_* + \lambda||Z_0 - H||_1 = ||L_0 + H||_* + \lambda||S_0 + H_S||_1$. Let X_L and X_S be the subgradients of $||\cdot||_*$ and $||\cdot||_1$ at L_0 and S_0. We have the identities: $X_L = UV^* + W + \mathcal{P}_{\Phi^\perp}(X_L - UV^* - W)$, and $\lambda X_S = UV^* + W - \mathcal{P}_\Omega(UV^* + W - \lambda E) + \mathcal{P}_{\Omega^\perp}(\lambda X_S - UV^* - W)$.

By the definition of subgradient, we have

$$\begin{aligned}
&||L_0 + H||_* + \lambda||S_0 + H_S||_1 \\
\geq &||L_0||_* + \lambda||S_0||_1 + \langle X_L, H\rangle + \lambda\langle X_S, H_S\rangle \\
= &||L_0||_* + \lambda||S_0||_1 + \langle UV^* + W, H\rangle + \langle \mathcal{P}_{\Phi^\perp}(X_L - UV^* - W), H\rangle \\
&+ \langle UV^* + W - \mathcal{P}_\Omega(UV^* + W - \lambda E), H_S\rangle \\
&+ \langle \mathcal{P}_{\Omega^\perp}(\lambda X_S - UV^* - W), H_S\rangle \\
= &||L_0||_* + \lambda||S_0||_1 + \langle UV^* + W, D_0\rangle + \langle X_L - UV^* - W, \mathcal{P}_{\Phi^\perp}(H)\rangle \\
&- \langle \mathcal{P}_\Omega(UV^* + W - \lambda E), H_S\rangle + \langle \lambda X_S - UV^* - W, \mathcal{P}_{\Omega^\perp}(H_S)\rangle \\
\geq &||L_0||_* + \lambda||S_0||_1 + \langle UV^* + W, D_0\rangle + \langle X_L - UV^* - W, \mathcal{P}_{\Phi^\perp}(H)\rangle \\
&+ \langle \lambda X_S - UV^* - W, \mathcal{P}_{\Omega^\perp}(H_S)\rangle - \frac{1}{16}\lambda\delta||\mathcal{P}_\Omega H_S||_F
\end{aligned} \tag{2.4.3}$$

where the first inequality follows from the definition of subgradient. The following equality is obtained by substituting the two identities of X_L and X_S into the right hand

side. The second equality uses the relationship $H = D_0 - H_S$, as well as self-adjointness of $\mathcal{P}_\Omega$. The last inequality follows from (2.3.5).

We now manipulate the last four terms on the right hand side of the last inequality.

For the first term, since $D_0 = \mathcal{P}_{\Omega^\perp} D_0$ by construction,

$$\langle UV^* + W, D_0 \rangle = \langle UV^* + W, \mathcal{P}_{\Omega^\perp} D_0 \rangle \geq -||\mathcal{P}_{\Omega^\perp}(UV^* + W)||_\infty ||D_0||_1 \geq -\frac{\lambda}{2}||D_0||_1 \tag{2.4.4}$$

For the second term, by duality, let X_L be such that $\langle X_L, \mathcal{P}_{\Phi^\perp}(H) \rangle = ||\mathcal{P}_{\Phi^\perp}(H)||_*$. Also, note that $|\langle UV^* + W, \mathcal{P}_{\Phi^\perp}(H) \rangle| \leq ||\mathcal{P}_{\Phi^\perp}(UV^* + W)|| \cdot ||\mathcal{P}_{\Phi^\perp}(H)||_* \leq \frac{1}{2}||\mathcal{P}_{\Phi^\perp}(H)||_*$ since $\mathcal{P}_{\Phi^\perp}(UV^*) = 0$ and $\mathcal{P}_{\Phi^\perp} W = W$. Therefore, we have

$$\langle X_L - UV^* - W, \mathcal{P}_{\Phi^\perp}(H) \rangle \geq \frac{1}{2}||\mathcal{P}_{\Phi^\perp}(H)||_* \tag{2.4.5}$$

For the third term, let $X_S = \text{sign}(\mathcal{P}_{\Omega^\perp} H_S)$ so $\langle X_S, \mathcal{P}_{\Omega^\perp}(H_S) \rangle = ||\mathcal{P}_{\Omega^\perp}(H_S)||_1$. Then

$$\langle \lambda X_S - UV^* - W, \mathcal{P}_{\Omega^\perp}(H_S) \rangle \geq \left(\lambda - ||\mathcal{P}_{\Omega^\perp}(UV^* + W)||_\infty\right) ||\mathcal{P}_{\Omega^\perp}(H_S)||_1 \geq \frac{\lambda}{2}||\mathcal{P}_{\Omega^\perp}(H_S)||_1 \tag{2.4.6}$$

Finally for the last term, since $H_S = D_0 - H$ and $\mathcal{P}_\Omega D_0 = 0$, we have $||\mathcal{P}_\Omega H_S||_F = ||\mathcal{P}_\Omega H||_F$ and

$$\begin{aligned}
||\mathcal{P}_\Omega H||_F \leq & ||\mathcal{P}_\Omega \mathcal{P}_\Phi H||_F + ||\mathcal{P}_\Omega \mathcal{P}_{\Phi^\perp} H||_F \\
\leq & \sqrt{1 - \delta + \varepsilon\delta}||H||_F + ||\mathcal{P}_{\Phi^\perp} H||_F \\
\leq & \sqrt{1 - \delta + \varepsilon\delta}(||\mathcal{P}_\Omega H||_F + ||\mathcal{P}_{\Omega^\perp} H||_F) + ||\mathcal{P}_{\Phi^\perp} H||_F \\
\leq & \sqrt{1 - \delta + \varepsilon\delta}(||\mathcal{P}_\Omega H||_F + ||\mathcal{P}_{\Omega^\perp} H_S||_1 + ||\mathcal{P}_{\Omega^\perp} D_0||_1) + ||\mathcal{P}_{\Phi^\perp} H||_*
\end{aligned}$$

where the last inequality follows from the fact that $||\cdot||_F \leq ||\cdot||_1$ and $||\cdot||_F \leq ||\cdot||_*$. By

rearranging the terms and using $D_0 = \mathcal{P}_{\Omega^\perp} D_0$ again,

$$||\mathcal{P}_\Omega H||_F \leq \frac{\sqrt{1-\delta+\varepsilon\delta}}{1-\sqrt{1-\delta+\varepsilon\delta}}(||\mathcal{P}_{\Omega^\perp} H_S||_1 + ||D_0||_1) + \frac{1}{1-\sqrt{1-\delta+\varepsilon\delta}}||\mathcal{P}_{\Phi^\perp} H||_* \quad (2.4.7)$$

Combining the last inequality in (2.4.3) with (2.4.4), (2.4.5), (2.4.6) and (2.4.7), we have

$$\begin{aligned}
||L_0 + H||_* + \lambda||Z_0 - H||_1 \geq & ||L_0||_* + \lambda||S_0||_1 + (\frac{1}{2} - \frac{\lambda\delta}{16(1-\sqrt{1-\delta+\varepsilon\delta})})||\mathcal{P}_{\Phi^\perp} H||_* \\
& + \lambda(\frac{1}{2} - \frac{\delta\sqrt{1-\delta+\varepsilon\delta}}{16(1-\sqrt{1-\delta+\varepsilon\delta})})||\mathcal{P}_{\Omega^\perp} H_S||_1 \\
& - \frac{\lambda}{2}||D_0||_1 - \frac{\lambda\delta\sqrt{1-\delta+\varepsilon\delta}}{16(1-\sqrt{1-\delta+\varepsilon\delta})}||D_0||_1 \\
\geq & ||L_0||_* + \lambda||S_0||_1 + \frac{1}{4}||\mathcal{P}_{\Phi^\perp} H||_* + \frac{\lambda}{4}||\mathcal{P}_{\Omega^\perp} H_S||_1 - \frac{3}{4}\lambda||D_0||_1
\end{aligned}$$

where the second inequality holds because

$$\frac{\delta}{1-\sqrt{1-\delta+\varepsilon\delta}} = \frac{\delta(1+\sqrt{1-\delta+\varepsilon\delta})}{\delta(1-\varepsilon)} \leq 4 \quad (2.4.8)$$

and

$$\frac{\delta\sqrt{1-\delta+\varepsilon\delta}}{1-\sqrt{1-\delta+\varepsilon\delta}} = \frac{\sqrt{1-\delta+\varepsilon\delta}(1+\sqrt{1-\delta+\varepsilon\delta})}{1-\varepsilon} \leq 4 \quad (2.4.9)$$

for large enough N.

Then again, since $S_0 = Z_0 - D_0$ and $\mathcal{P}_{\Omega^\perp} H_S = \mathcal{P}_{\Omega^\perp}(D_0 - H) = D_0 - \mathcal{P}_{\Omega^\perp} H$,

$$\begin{aligned}
& ||L_0||_* + \lambda||S_0||_1 + \frac{1}{4}||\mathcal{P}_{\Phi^\perp} H||_* + \frac{\lambda}{4}||\mathcal{P}_{\Omega^\perp} H_S||_1 - \frac{3}{4}\lambda||D_0||_1 \\
\geq & ||L_0||_* + \lambda||Z_0||_1 + \frac{1}{4}||\mathcal{P}_{\Phi^\perp} H||_* + \frac{\lambda}{4}||\mathcal{P}_{\Omega^\perp} H||_1 - 2\lambda||D_0||_1
\end{aligned}$$

Therefore, by (2.4.1),

$$\begin{aligned} & ||\mathcal{P}_{\Phi^\perp} H||_* + \lambda ||\mathcal{P}_{\Omega^\perp} H||_1 \leq 8\lambda ||D_0||_1 \\ \implies & \frac{1}{\lambda} ||\mathcal{P}_{\Phi^\perp} H||_* + ||\mathcal{P}_{\Omega^\perp} H||_1 \leq 8 ||D_0||_1 \end{aligned}$$

□

Remark 2.4.1. Under the event that $||L_0||_\infty \leq \alpha$, L_0 is a feasible solution to the minimization problem in (2.1.2), so $\hat{L}$ yields a smaller objective function than L_0 by definition. Then equation (2.4.1) is satisfied so the bound given by Lemma 2.4.1 applies to the estimation error $\hat{L} - L_0$.

Remark 2.4.2. The basic idea of the proof is in line with Lemma 2.5 in Candès et al. (2011) and Lemma 5 in Zhou et al. (2010). However there are major differences which lead to a tighter bound than the one obtained in Zhou et al. (2010) for SPCP. By construction, the argument in $||\cdot||_*$ and $||\cdot||_1$ in our objective function in (2.1.2) always add up to Y, while when we take the subderivative for $||\cdot||_1$, we take it at S_0 instead of Z_0. This enables us to utilize both the dual certificate as well as the nice relationship between H_S and H and the relationship between the support sets of D_0 and S_0, all generated by the Bernoulli device. These special properties in turn yield a tighter bound in terms of $||D_0||_1$.

2.5 Main Results

From Lemma 2.4.1, we obtain bounds for $||\mathcal{P}_{\Phi^\perp} H||_*$ and $||\mathcal{P}_{\Omega^\perp} H||_1$. The next lemma bounds the corresponding norms of their complements.

Lemma 2.5.1. *Suppose* $||\mathcal{P}_\Omega \mathcal{P}_\Phi||^2 \leq 1 - \delta + \varepsilon\delta$, $\delta \to 0$ *and* $\varepsilon \to 0$ *as* $N, T \to \infty$, *then for*

large enough N and T,

$$||\mathcal{P}_\Phi H - \mathcal{P}_\Omega H||_F^2 \geq \frac{\delta}{4}(||\mathcal{P}_\Phi H||_F^2 + ||\mathcal{P}_\Omega H||_F^2) \tag{2.5.1}$$

Proof.

$$\begin{aligned}
||\mathcal{P}_\Phi H - \mathcal{P}_\Omega H||_F^2 =& ||\mathcal{P}_\Phi H||_F^2 + ||\mathcal{P}_\Omega H||_F^2 - 2\langle \mathcal{P}_\Phi H, \mathcal{P}_\Omega H\rangle \\
\geq& ||\mathcal{P}_\Phi H||_F^2 + ||\mathcal{P}_\Omega H||_F^2 - 2\sqrt{1-\delta+\varepsilon\delta}||\mathcal{P}_\Phi H||_F||\mathcal{P}_\Omega H||_F \\
\geq& (1-\sqrt{1-\delta+\varepsilon\delta})(||\mathcal{P}_\Phi H||_F^2 + ||\mathcal{P}_\Omega H||_F^2) \\
=& \frac{\delta(1-\varepsilon)}{1+\sqrt{1-\delta+\varepsilon\delta}}(||\mathcal{P}_\Phi H||_F^2 + ||\mathcal{P}_\Omega H||_F^2)
\end{aligned}$$

where the first inequality follows from $|\langle \mathcal{P}_\Phi H, \mathcal{P}_\Omega H\rangle| = |\langle \mathcal{P}_\Omega H, \mathcal{P}_\Omega \mathcal{P}_\Phi \mathcal{P}_\Phi H\rangle| \leq ||\mathcal{P}_\Omega \mathcal{P}_\Phi||$ $\cdot ||\mathcal{P}_\Omega H||_F \cdot ||\mathcal{P}_\Phi H||_F$. Then for large enough N and T, $1-\varepsilon > \frac{1}{2}$ and $1+\sqrt{1-\delta+\varepsilon\delta} < 2$. This completes the proof. □

Given Lemmas 2.4.1 and 2.5.1, now we are ready to state and prove the main result in the paper.

Theorem 2.5.1. *Under Assumptions 2.2.1, 2.3.1 and the conditions in Lemma 2.3.2, if $\alpha\mu r = o(N^{1/3})$, there exists $C > 0$ such that*

$$P\Big(\frac{1}{NT}||\hat{L} - L_0||_F^2 \leq C\big(\alpha \vee (\mu^{8/3} r^2)\big)\delta\Big) \to 1 \tag{2.5.2}$$

where $\big(\alpha \vee (\mu^{8/3} r^2)\big)\delta = \frac{(\alpha\mu r)\vee(\mu^{11/3} r^3)}{N^{1/3}} = o(1)$.

Proof. Under the event

$$\mathcal{E} \equiv \{||\mathcal{P}_\Omega \mathcal{P}_\Phi||^2 \leq 1-\delta+\varepsilon\delta\} \cap \{W \text{ satisfies (2.3.5)}\} \cap \{\text{conditions (2.3.1) } \textit{and} \text{ (2.3.2) holds}\},$$

let $H = \hat{L} - L_0$ and by Lemma 2.4.1,

$$||\mathcal{P}_{\Phi^\perp} H||_F^2 \leq 64\lambda^2 ||D_0||_1^2 \tag{2.5.3}$$

$$||\mathcal{P}_{\Omega^\perp} H||_F^2 \leq 16\alpha ||D_0||_1 \tag{2.5.4}$$

where (2.5.3) follows from $||\mathcal{P}_{\Phi^\perp} H||_F \leq ||\mathcal{P}_{\Phi^\perp} H||_*$ and (2.5.4) follows from $||\mathcal{P}_{\Omega^\perp} H||_F^2 \leq ||\mathcal{P}_{\Omega^\perp} H||_\infty \cdot ||\mathcal{P}_{\Omega^\perp} H||_1$. Notice under $\mathcal{E}$, $||L_0||_\infty \leq \alpha$, so $||\mathcal{P}_{\Omega^\perp} H||_\infty \leq ||H||_\infty \leq ||L_0||_\infty + ||\hat{L}||_\infty \leq 2\alpha$.

Since $H - H = \mathcal{P}_\Phi H + \mathcal{P}_{\Phi^\perp} H - \mathcal{P}_\Omega H - \mathcal{P}_{\Omega^\perp} H = 0$, we have $||\mathcal{P}_\Phi H - \mathcal{P}_\Omega H||_F = ||\mathcal{P}_{\Phi^\perp} H - \mathcal{P}_{\Omega^\perp} H||_F$. By Lemma 2.5.1,

$$\begin{aligned}
\frac{\delta}{4}(||\mathcal{P}_\Phi H||_F^2 + ||\mathcal{P}_\Omega H||_F^2) \leq & ||\mathcal{P}_\Phi H - \mathcal{P}_\Omega H||_F^2 \\
= & ||\mathcal{P}_{\Phi^\perp} H - \mathcal{P}_{\Omega^\perp} H||_F^2 \\
\leq & 4\big(||\mathcal{P}_{\Phi^\perp} H||_F^2 \vee ||\mathcal{P}_{\Omega^\perp} H||_F^2\big)
\end{aligned}$$

Therefore,

$$||H||_F^2 = \frac{1}{2}(||\mathcal{P}_{\Omega^\perp} H||_F^2 + ||\mathcal{P}_\Omega H||_F^2 + ||\mathcal{P}_{\Phi^\perp} H||_F^2 + ||\mathcal{P}_\Phi H||_F^2) \tag{2.5.5}$$

$$\leq \frac{C_1}{\delta}(\alpha ||D_0||_1 \vee \lambda^2 ||D_0||_1^2) \tag{2.5.6}$$

Next we derive the order of $||D_0||_1$. Note $|D_{0,it}| \leq C'\delta$ by construction. Therefore,

$$\begin{aligned}
P(||D_0||_1 - C'NT\delta^2 \geq C't\delta) \leq & P(\delta \sum_{it} M_{it} - NT\delta^2 \geq t\delta) \\
= & P(\sum_{it} M_{it} - NT\delta \geq t) \\
\leq & \exp\big(-\frac{2t^2}{NT}\big)
\end{aligned}$$

where M_{it} is defined in (2.2.3) and the last inequality follows from Hoeffding's inequality.

Let $t = N \log N$, then $\exp(-\frac{2t^2}{NT}) \to 0$. Meanwhile, since $NT\delta \asymp N^{5/3}\mu r$, it dominates t. Therefore, $||D_0||_1 \leq C_2 NT\delta^2$ with probability approaching 1.

Now conditional on the event $\bar{\mathcal{E}} \equiv \mathcal{E} \cap \{||D_0||_1 \leq C_2 NT\delta^2\}$, (2.5.6) implies

$$\begin{aligned} ||H||_F^2 \leq & \frac{C}{\delta}\big((\alpha NT\delta^2) \vee (\lambda^2 (NT)^2 \delta^4)\big) \\ \leq & C\delta\big((\alpha NT) \vee (\lambda^2 (NT)^2 \delta^2)\big) \\ \implies \frac{1}{NT}||H||_F^2 \leq & C\big(\alpha \vee (\mu^{8/3} r^2)\big)\delta \end{aligned}$$

This completes the proof as $\bar{\mathcal{E}}$ occurs with probability approaching 1. □

Remark 2.5.1. Theorem 2.5.1 indicates that the rate of convergence is determined by δ, r, μ and α. As the driving force of consistency, the faster δ converges to 0, the faster the convergence rate of $\frac{1}{NT}||\hat{L} - L_0||_F^2$ will be. However, the rate of δ is bounded because the fraction of 0 entries in S_0 cannot be too small for the existence of a dual certificate. The parameters r, μ and α slow the convergence down, but if we look at the relative bias, $\frac{||\hat{L}-L_0||_F^2}{||L_0||_F^2}$, the effects of these parameters may be mitigated or reversed. For example suppose L_0 follows the random orthogonal model introduced in Section 3 and all the r singular values are of same order. Then $\alpha \asymp \frac{\sigma_1 \mu r}{\sqrt{NT}}$ where σ_1 is the largest singular value, so the error bound is proportional to $(\sigma_1 \mu^2 r^2) \vee (\mu^{11/3} r^3)$. Meanwhile, the order of $||L_0||_F^2$ is proportional to $\sigma_1^2 \mu^2 r^2$. Therefore, the upper bound of the relative bias is proportional to $\frac{1}{\sigma_1} \vee \frac{\mu^{5/3} r}{\sigma_1^2}$, and we can see the effect of μ and r are smaller and the larger the singular values are, which, for fixed μ and r, implies a larger α, the smaller the relative bias is. This is because L_0 with larger singular values tend to dominate Z_0 in the decomposition, and thus recovering it from the observed data matrix Y is relatively easier.

2.6 Simulations

In this section, we present two simulation experiments to illustrate the effectiveness of the BPCP estimator. In the first experiment, we generate random L_0 and Z_0 with $r = 1, 3, 5$. In each case, we examine the performance of the estimator for both Gaussian and Cauchy error matrices by comparing $\frac{1}{NT}||\hat{L} - L_0||_F^2$ as well as the relative estimation error $||\hat{L} - L_0||_F^2/||L_0||_F^2$ in each case as we gradually increase the sample size. In the second experiment, we fix a picture as L_0, and superimpose it with Gaussian and Cauchy white noise. The purpose of this example is to visually show how the estimator performs when the error is continuously fat-tailed distributed.

To implement the estimator, note the minimization problem in (2.1.2) is equivalent as

$$\min_{L,Z} \quad ||L||_* + \lambda||Z||_1, \ s.t. \ L + Z = Y, \quad ||L||_\infty \leq \alpha \tag{2.6.1}$$

We first set α to be a large number and solve the problem without the inequality constraint, then verify whether the solution satisfies the inequality. For the first step, we adopt the Augmented Lagrangian Multiplier algorithm (ALM) studied in Lin, Chen and Ma (2010) and Yuan and Yang (2013). The algorithm solves the following problem

$$\min_{L,Z} \quad ||L||_* + \lambda||Z||_1 + \langle \Lambda, Y - L - Z\rangle + \frac{\nu}{2}||Y - L - Z||_F^2 \tag{2.6.2}$$

where Λ is the Lagrangian multiplier for the equality constraint and the last term is a penalty for deviating from the constraint. The algorithm solves the minimization problem iteratively and terminates if both $||Y - L_k - Z_k||_F$ and $\nu||Z_k - Z_{k-1}||_F$ are small enough, where the subscript k denotes the kth iteration. We set the stopping criteria to be $10^{-7}||Y||_F$ and 10^{-5} respectively. Following Yuan and Yang (2013) and Candès et al. (2011) we set $\nu = \frac{NT}{4||Y||_1}$. Finally, λ is set to be $0.7(\frac{\log(N\wedge T)}{NT})^{1/3}$ in the first experiment and $0.5(\frac{\log(N\wedge T)}{NT})^{1/3}$ in the second; both of them satisfy the conditions in the theory when μ

is of the order $\log(N)$ and $N \asymp T$.

Note that although the estimator in theory solves a one-step minimization problem, not needing to estimate the rank of L_0, it still requires an iterative procedure in practice. In each iteration, an SVD decomposition is performed. This could be potentially compute-intensive for very large N and/or T. In this respect, a good initial guess would be important to save computation time, although it does not affect the estimate obtained by the nature of convex programming.

2.6.1 Numerical Experiment

In this experiment, N and T take values from $\{200, 300, 400, 500\}$. The rank $r = 1, 3, 5$. In each case, we draw an $N \times r$ and an $r \times T$ matrix from $N(0, 1)$ and their product is L_0. For Gaussian error, entries of Z_0 are independently drawn from $N(0, 1)$ while for Cauchy error, they are drawn from the standard Cauchy distribution.

Table 2.1 shows the value $\frac{1}{NT}||\hat{L} - L_0||_F^2$ in each case. It decreases as the minimum of N and T increases. Also, consistent with Theorem 2.5.1, fixing N and T, the estimation error increases as the rank increases. However, the magnitude of the increase is smaller than indicated by Theorem 2.5.1, where the upper bound is proportional to r^3. This suggests our bound may be further improved and deriving the lower bound may also be useful. Also, the estimator works regardless whether the data matrix is square or not. When the matrix is not square, which dimension is larger should not affect the results by i.i.d. This is evident in both Table 2.1 and 2.2; the results in the rows of $N = 200, T = 500$ and $N = 300, T = 500$ are very close to those in the rows of $N = 500, T = 200$ and $N = 500, T = 300$, respectively. Another observation is that the estimation error under Cauchy errors is systematically bigger than under Gaussian errors. To explain this phenomenon, note that the standard Cauchy density is lower than the standard normal density around 0, so for any given δ that is small enough, $-\underline{\gamma}_{it}$ and $\bar{\gamma}_{it}$ introduced in Section 2.2 are bigger for the Cauchy error, resulting in bigger $||D_0||_1$. This difference is

negligible when $N, T \to \infty$ but leads to finite sample differences in the error bound.

Table 2.1: Average Estimation Error $\frac{1}{NT}||\hat{L} - L_0||_F^2$

	Gaussian Error			Cauchy Error		
	$r = 1$	$r = 3$	$r = 5$	$r = 1$	$r = 3$	$r = 5$
$N = T = 200$	0.0461	0.1349	0.2535	0.0812	0.3002	0.5491
$N = T = 300$	0.0322	0.0943	0.1692	0.0533	0.1999	0.3676
$N = T = 400$	0.0247	0.0736	0.1308	0.0438	0.1496	0.2764
$N = T = 500$	0.0209	0.0619	0.1078	0.0359	0.1198	0.2151
$N = 200, T = 500$	0.0313	0.0956	0.1641	0.0536	0.1899	0.3575
$N = 500, T = 200$	0.0333	0.0897	0.1645	0.0538	0.1881	0.3473
$N = 300, T = 500$	0.0270	0.0781	0.1358	0.0435	0.1531	0.2789
$N = 500, T = 300$	0.0265	0.0768	0.1339	0.0430	0.1570	0.2844

Table 2.2 shows the relative estimation error $\frac{||\hat{L}-L_0||_F^2}{||L||_F^2}$. We can see similar patterns as in Table 2.1. For instance, within every column, the relative error is also decreasing as the minimum of N and T increases. Under the same error distribution, the differences across ranks are now significantly smaller. This is because as the rank increases, $||L_0||_F^2$ also increases by construction. Specifically, since L_0 is the product of two independent matrices with i.i.d. standard normal entries, $||L_0||_F^2$ is of the order of NTr, so dividing it mitigates the effect of r.

Table 2.2: Relative Estimation Error $\frac{||\hat{L}-L_0||_F^2}{||L_0||_F^2}$

	Gaussian Error			Cauchy Error		
	$r=1$	$r=3$	$r=5$	$r=1$	$r=3$	$r=5$
$N=T=200$	0.0378	0.0411	0.0482	0.0679	0.1082	0.1139
$N=T=300$	0.0265	0.0303	0.0334	0.0459	0.0687	0.0729
$N=T=400$	0.0224	0.0243	0.0265	0.0383	0.0498	0.0538
$N=T=500$	0.0197	0.0203	0.0220	0.0310	0.0395	0.0416
$N=200, T=500$	0.0280	0.0301	0.0325	0.0432	0.0658	0.0707
$N=500, T=200$	0.0287	0.0286	0.0324	0.0482	0.0647	0.0707
$N=300, T=500$	0.0238	0.0260	0.0273	0.0365	0.0514	0.0548
$N=500, T=300$	0.0232	0.0243	0.0268	0.0380	0.0531	0.0555

2.6.2 Graphical Experiment

In this experiment, we superimpose white noises drawn from the standard normal or the standard Cauchy distribution on a picture. The picture we use is directly downloaded from a built-in example in MATLAB (the file name is eight.tif). The picture has resolution 242×308. We stack all columns into a long vector and duplicate it for 199 times, obtaining a 74536×200 matrix L_0. By construction, L_0 has rank 1 because all columns are equal. Then we draw a 74536×200 matrix with i.i.d. entries from either of the two distributions as Z_0 and add it to L_0. This procedure simulates 200 frames from a video of the static picture interfered by white noise.

Figure 2.1 and Figure 2.2 display the results. In each figure, the northwest (NW) is the original picture. There are four coins in it, two heads and two tails. The southwest (SW) shows one of the 200 frames after the the picture is superimposed with Gaussian noises (Figure 2.1) or Cauchy noises (Figure 2.2). It can be seen that the details of the coins

are no longer unrecognizable. The northeast (SE) quadrant shows the recovered picture. Differences in this picture between Figure 2.1 and Figure 2.2 are hardly to be seen, except the background in Figure 2.2 is slightly darker. The southeast quadrant shows the same frame of residuals, in which we cannot see contours of the coins, indicating it contains very few information about the original picture.

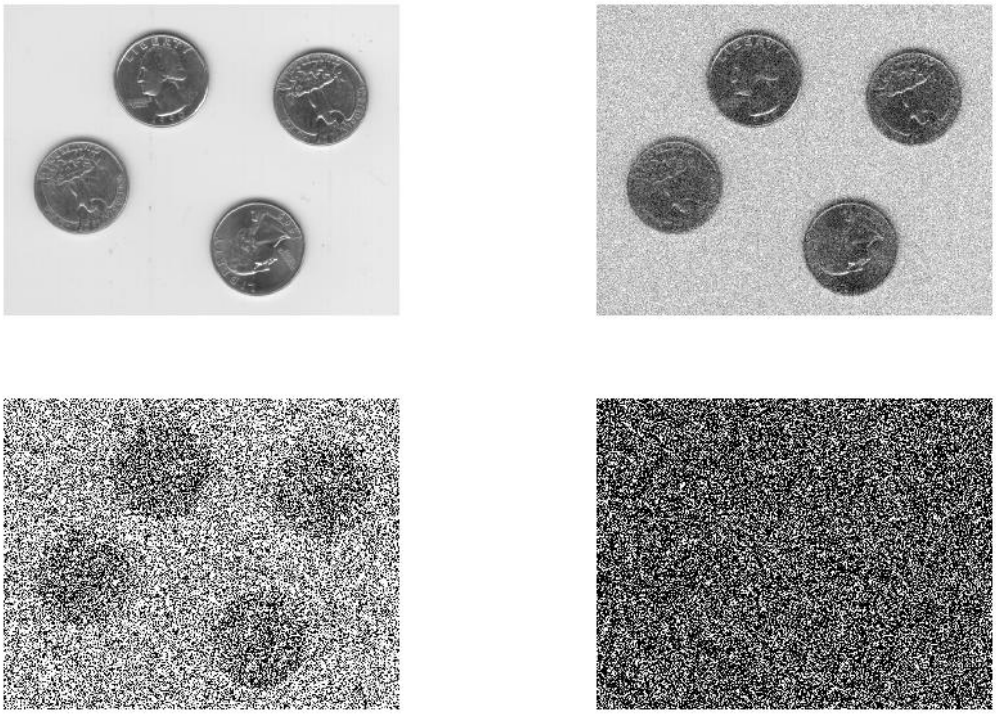

Figure 2.1: Gaussian Noise

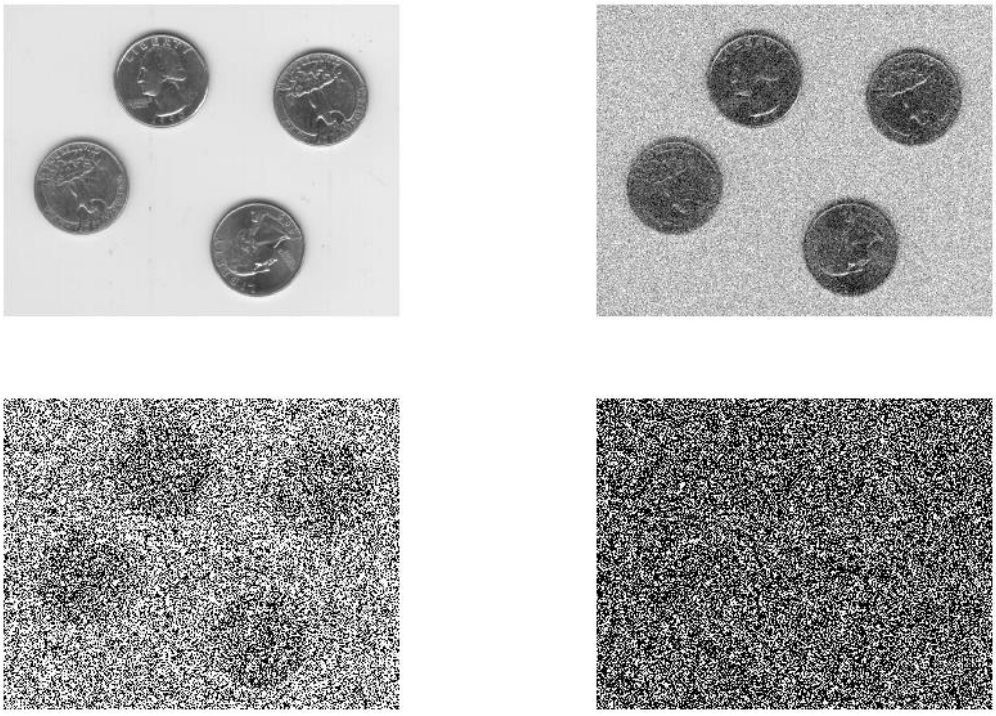

Figure 2.2: Cauchy Noise

2.7 Conclusion

This paper provides sufficient conditions under which the Bounded Principal Component Pursuit (BPCP) is consistent for the random or deterministic low-rank component, whose entries can go to infinity with N and T. By utilizing the Bernoulli device we developed, we show that the median zero assumption translates the non-sparse error matrix into the sum of a matrix with a sufficient number of zeros which guarantees the existence of a dual certificate, and a matrix of small enough norm that bounds the estimation error. Since no restrictions on errors' moments are imposed, the result in this paper implies that BPCP would work well for heavy- or fat-tailed errors, and thus more robust than the classical principal component analysis. The implication is confirmed by simulations.

This paper considers the case where all entries of Y are observed. It is also interesting

to apply our approach to the case where there are missing entries in Y. Also, this paper only shows consistency in the Frobenius norm. It would be interesting to see whether consistency holds componentwise or in the ℓ_∞ norm. These aspects will be studied in future work.

Chapter 3

Regularized Quantile Regression with Interactive Fixed Effects

Junlong Feng[†]

[†]I thank Jushan Bai, Roger Koenker, Sokbae (Simon) Lee, José Luis Montiel Olea, Roger Moon, Serena Ng, Bernard Salanié, Jörg Stoye and participants at the Columbia Econometrics Colloquium for helpful comments and discussions.

3.1 Introduction

Panel data models are widely applied in economics and finance. Allowing for rich heterogeneity, interactive fixed effects are important components in such models in many applications. Since fixed effects may not only impact the conditional mean of the outcome variable, but also have heterogeneous effects on its distribution, quantile regression would be handy in such cases. However, in contrast to the well-studied mean regression (e.g. Pesaran (2006), Bai (2009) and Moon and Weidner (2015)), quantile regression for panel data with interactive fixed effects only received attention recently (e.g. Harding and Lamarche (2014), Ando and Bai (2020), Belloni et al. (2019), Chen, Dolado and Gonzalo (2019) and Chen (2019)).

In this paper, I propose a new estimator for high-dimensional panel data models that employs a nuclear norm penalty to handle the interactive fixed effects. This estimator solves a convex problem, jointly obtaining consistent estimates of the slope coefficients and the interactive fixed effects, not needing to know the number of the fixed effects. For concreteness, consider the following linear conditional quantile model for an outcome variable Y_{it}: $q_{Y_{it}|X_{it}}(u) = X_{it}'\beta_0(u) + F_{0t}(u)'\Lambda_{0i}(u)$ where $F_{0t}(u)$ and $\Lambda_{0i}(u)$ contain $r(u)$ fixed effects that affect the u-th conditional quantile of Y_{it} and are treated as non-random parameters. Let $\rho_u(\cdot)$ denote the check function as in standard quantile regression. A natural strategy to estimate $(\beta_0(u), F_{0t}(u), \Lambda_{0i}(u))$, as in the related literature, is to solve the minimization problem: $\min_{(\beta,\{F_t\}_t,\{\Lambda_i\}_i)} \sum_{i,t} \rho_u(Y_{it} - X_{it}\beta - F_t'\Lambda_i)$. To implement, $r(u)$ needs to be known or pre-estimated. In the meantime, this objective function is non-convex in the parameters due to the interaction term, so we may only obtain a local minimum.

The estimator proposed in this paper avoids the estimation of $r(u)$ and the issue of nonconvexity. Note that the minimization problem above is equivalent as the following

constrained problem:

$$\min_{(\beta,L)} \sum_{i,t} \rho_u(Y_{it} - X_{it}'\beta - L_{it})$$
$$s.t. \ \text{rank}(L) \leq r(u)$$

In this problem, nonconvexity is due to the constraint as the rank of a matrix is nonconvex. Inspired by the seminal work of Candès and Recht (2009) and Candès et al. (2011), the estimator in this paper minimizes the convex surrogate objective function by replacing the nonconvex rank constraint by a convex penalty function, the nuclear norm of L, added to the objective function. Denoted by $||L||_*$, the nuclear norm is the sum of L's singular values. As the singular values are nonnegative by definition, the nuclear norm can be viewed as the ℓ_1 norm of the singular value vector. On the other hand, the rank of L is the number of its nonzero singular values. Therefore, the nuclear norm of a matrix is the tightest convex relaxation of its rank, analogous to the LASSO penalty in high dimensional regression. The minimizer of the regularized problem will thus presumably have similar properties as those of the unregularized constrained problem.

The benefits of the nuclear norm regularized quantile regression are two-fold. First, as the constraint is removed and L is treated as an entity, $r(u)$ is no longer needed, avoiding pre-estimation. The rank of the estimated common component is partly determined by the weight assigned to the penalty. In this paper, we also provide a consistent estimator for $r(u)$ as a by-product of the low-rank common component estimator. Second, since both the check function and the nuclear norm are convex in the parameters, there is no concern of local minimum.

The main contributions of this paper are as follows. I derive the uniform rate of convergence for the regularized estimator. The rate for the low-rank common component is nearly optimal but the rate for the slope coefficients is slower than the constrained problem due to the bias from regularization. To obtain the rate, the subgradient matrix of the

check function of the error processes plays a key role. I prove new results that establish a uniform upper bound for the operator norm of this matrix, and a uniform bound for its inner product with other matrices. Also, I develop a new theoretical argument under primitive conditions such that the conditional density of Y_{it} only needs to be away from o at the true parameters. This is weaker than the conditions in Ando and Bai (2020) and Chen, Dolado and Gonzalo (2019), where the density is assumed to be bounded away from o over compact subsets of the support. The required conditions are low-level and easier to interpret than the regularity conditions in Belloni et al. (2019). These results may have independent interest. Finally, I discuss post-regularization procedures where a consistent estimator for $r(u)$ is proposed.

This paper adds to the literature of quantile regression for panel data. Since Koenker (2004), panel data quantile regression began to draw increasing attention. Abrevaya and Dahl (2008), Lamarche (2010), Canay (2011), Kato, Galvao Jr and Montes-Rojas (2012), Galvao, Lamarche and Lima (2013) and Galvao and Kato (2016) study quantile regression with one-way or two-way fixed effects. Harding and Lamarche (2014) considers interactive fixed effects with endogenous regressors. They require the factors to be pre-estimated or known. Chen, Dolado and Gonzalo (2019) prospses a quantile factor model without regressors. They estimate the factors and the factor loadings via nonconvex minimization. Pre-estimation of the number of the factors is needed. Ando and Bai (2020) considers quantile regression with heterogeneous coefficients. They propose both a frequentist and a Bayesian estimation procedure. The number of the factors also needs to be estimated first, and the minimization problem is noncovex. Chen (2019) proposes a two-step estimator: by assuming the common factors are not quantile-dependent, they are estimated by the principal component analysis in the first step, and the regression coefficients and the individual fixed effects are estimated in the second step via smoothed quantile regression. Chao, Härdle and Yuan (2019) considers nuclear norm penalized multi-task quantile regression where multiple outcome variables are simultaneously con-

sidered and the matrix of the slope coefficients is low-rank. Finally, a recent independent work by Belloni et al. (2019) studies regularized quantile regression with both interactive fixed effects and high-dimensional regressors. To deal with the high-dimensional regressors, an ℓ_1 penalty is needed in addition to the nuclear penalty. Unlike this paper, they focus on pointwise convergence. As mentioned in the main contributions, establishing the uniform rate is challenging and requires new results in the random matrix theory.

Another literature this paper speaks to is the nuclear norm regularized estimation. This literature was initially motivated by low-rank matrix completion or recovery in compressed sensing and other applications in computer science, etc, for instance Candès and Recht (2009), Ganesh et al. (2010), Zhou et al. (2010), Candès et al. (2011), Hsu, Kakade and Zhang (2011), Negahban and Wainwright (2011), Agarwal, Negahban and Wainwright (2012) and Negahban et al. (2012) among others. The main parameter of interest in this literature is the low-rank matrix and they require the error terms to be either nonstochastic or to have finite second moments. Bai and Feng (2019) allows the stochastic error to be non-sparse and fat-tailed; the existence of its moments is not required. Nuclear norm regularized estimation and matrix-completion-related topics have also gained interest in econometrics recently. Bai and Ng (2019*a*) considers using factor analysis to impute missing data and counterfactuals. Bai and Ng (2019*b*) considers regularized estimation for approximate factor models with singular values thresholding. Athey et al. (2018), Moon and Weidner (2019)[1] and Chernozhukov et al. (2019) consider mean regression with interactive fixed effects.

The rest of the paper is organized as follows. Section 3.2 introduces the model and the estimator. Section 3.3 discusses the *restricted set*, an important theoretical device that is useful to establish the main results. The main results are presented in Section 3.4. Section 3.5 shows the results of a Monte Carlo simulation experiment. Section 3.6

[1]Moon and Weidner (2019) also briefly discuss nuclear norm regularized quantile regression with a single regressor as an extension. Using a different approach than this paper, they focus on pointwise convergence rate of the slope coefficient. In this paper, we obtain uniform rates for both the coefficients and the low-rank component.

concludes. The proofs of the results in Sections 3.3 and 3.4 and some technical lemmas are collected in Appendices C.1, C.2 and C.3 respectively.

Notation

Besides the nuclear norm $||\cdot||_*$, four additional matrix norms will be used in the paper. Let $||\cdot||$, $||\cdot||_F$, $||\cdot||_1$, and $||\cdot||_\infty$ denote the operator norm, the Frobenius norm, the ℓ_1 norm and the maximum norm. For two generic scalars, $a \vee b$ and $a \wedge b$ return the max and the min in a and b, respectively. For a generic random sample $W_{11}, ..., W_{NT}$ and a function f, denote the empirical process by $\mathbb{G}_{NT}(f) \equiv \mathbb{G}_{NT}(f(W_{it})) = \frac{1}{\sqrt{NT}} \sum_{i,t}(f(W_i) - \mathbb{E}(f(W_i)))$ where $\mathbb{E}$ denotes the expectation operator. Finally, I use the notion *"with high probability"* when an event occurs with probability arbitrarily close to 1 for large enough N and T.

3.2 The Model and the Estimator

I consider the following model for a panel dataset $(Y_{it}, X_{it} : i \in \{1, ..., N\}, t \in \{1, ..., T\})$:

$$Y_{it} = X_{it}'\beta_0(U_{it}) + \sum_{k=1}^{\bar{r}} \mathbb{1}_k(U_{it}) F_{kt}(U_{it}) \Lambda_{ki}(U_{it}), \tag{3.2.1}$$

where X_{it} is a $p \times 1$ vector of covariates, $U_{it} \sim \text{Unif}[0,1]$ is unobserved, $F_t(U_{it})$ and $\Lambda_i(U_{it})$ contain latent fixed effects with the total number $\bar{r}$ unknown. I allow the fixed effects to be heterogeneous in the realization of U_{it}. Specifically, for each realization $U_{it} = u$, the indicator function selects the effective fixed effects that impact Y_{it}. Hence, the effective number of the fixed effects is u-dependent; at $U_{it} = u$, $r(u) = \sum_{k=1}^{\bar{r}} \mathbb{1}_k(u)$. Similar setups can be found in Ando and Bai (2020) and Chen, Dolado and Gonzalo (2019).

The interactive fixed effects in model (3.2.1) form an $N \times T$ matrix of rank at most

equal to $\bar{r}$. To see this, let us rewrite equation (3.2.1) as follows:

$$\begin{aligned} Y_{it} =& X_{it}'\beta_0(u) + \underbrace{\sum_{k=1}^{\bar{r}} \mathbb{1}_k(u) F_{kt}(u)\Lambda_{ki}(u)}_{\equiv L_{0,it}(u)} \\ &+ \underbrace{X_{it}'(\beta_0(U_{it}) - \beta_0(u)) + \sum_{k=1}^{\bar{r}} \left(\mathbb{1}_k(U_{it})F_{kt}(U_{it})\Lambda_{ki}(U_{it}) - \mathbb{1}_k(u)F_{kt}(u)\Lambda_{ki}(u)\right)}_{\equiv V_{it}(u)} \end{aligned}$$

Or more compactly in matrix notation,

$$Y = \sum_{j=1}^{p} X_j \beta_{0,j}(u) + L_0(u) + V(u) \tag{3.2.2}$$

where $\text{rank}(L_0(u)) \leq r(u) \leq \bar{r}$ and X_j $(N \times T)$ is the j-th regressor. The number of the regressors p and the number of factors $\bar{r}$ are smaller than N and T and are assumed to be fixed for simplicity. $(\beta_0(u), L_0(u))$ are treated as non-random parameters to be estimated. Denote the conditional quantile of $V_{it}(u)$ at u by $q_{V_{it}(u)|X_{it}}(u)$. Then assuming that the function $X_{it}'\beta(\cdot) + L_{0,it}(\cdot)$ is strictly increasing almost surely, we have $q_{V_{it}(u)|X_{it}}(u) = 0$ almost surely.

The following examples provide models that admit the representations (3.2.1) and (3.2.2).

Example 3.2.1 (Location Shift Model). $Y_{it} = X_{it}'\beta^o + F_t^{o'}\Lambda_i^o + \epsilon_{it}$. *This is the model we discussed in Introduction. Let $q_\epsilon(\cdot)$ be the quantile function of ϵ, then $Y_{it} = X_{it}'\beta^o + F_t^{o'}\Lambda_i^o + q_\epsilon(U_{it})$. In this model, only the intercept is U_{it}-dependent.*

Example 3.2.2 (Location-Scale Model). $Y_{it} = X_{it}'\beta_0^a + F_t^{o'}\Lambda_t^a + (X_{it}'\beta^b + F_t^{o'}\Lambda_i^b)\epsilon_{it}$. *Again, let $q_\epsilon(\cdot)$ be the quantile function of ϵ, then we can rewrite the model as*

$$Y_{it} = X_{it}'[\beta_0^a + \beta_0^b q_\epsilon(U_{it})] + F_t^{o'}[\Lambda_i^a + \Lambda_i^b q_\epsilon(U_{it})]$$

where $x'\beta^b + F_t^{o'}\Lambda_i^b > 0$ *for all* i, t, *and all* x *in the support set of* X_{it}. *In this model, both the slope coefficients and the individual fixed effects are functions of* U_{it}. *If for some* $U_{it} = u$, *there are elements in* $\Lambda_i^a + \Lambda_i^b q_\varepsilon(u)$ *equal to* 0, *then the number of the effective fixed effects is smaller than that at other* u. *In this case,* $r(\cdot)$ *depends on* u.

Under representation (3.2.2), the conditional quantile restriction $q_{V_{it}(u)|X_{it}}(u) = 0$, and low-rankness of $L_0(u)$, I propose to estimate $(\beta_0(u), L_0(u))$ by the following regularized quantile regression:

$$(\hat{\beta}(u), \hat{L}(u)) \equiv \arg \min_{\beta \in \mathbb{R}^p, ||L||_\infty \leq \alpha} \frac{1}{NT} \boldsymbol{\rho}_u(Y - \sum_{j=1}^{p} X_j \beta_j - L) + \lambda ||L||_* \tag{3.2.3}$$

where for a generic matrix Z, $\boldsymbol{\rho}_u(Z) \equiv \sum_{i,t} \rho_u(Z_{it}) \equiv \sum_{i,t} Z_{it}(u - \mathbb{1}_{Z_{it} \leq 0})$ is the sum of the check function $\rho_u(\cdot)$ applied to each element in Z.

Without the penalty ($\lambda = 0$), the solution to the minimization problem is $(0, Y)$ because it sets the objective function equal to 0. To avoid this trivial and undesirable solution, the nuclear norm penalty as a convex surrogate of the rank constraint is added to the check functions. On the other hand, when λ is infinity, $\hat{L}(u)$ would be 0 to set the nuclear norm equal to o. We will specify the appropriate order of λ in the following sections to guarantee consistency of the estimator.

Finally, we allow the parameter space of the slope coefficients to be unbounded, as in standard quantile regression. On the other hand, we assume the magnitudes of the elements in $L_0(u)$ to be bounded by α, which is allowed to be N, T-dependent and can diverge to infinity at certain rate. We postpone to motivate this requirement in Section 3.4.

3.3 The Restricted Set

To establish consistency, intuitively we hope that $\hat{L}(u)$ lies near the space that contains $L_0(u)$. In particular, the nuclear norm penalty is effective if projecting $\hat{L}(u)$ onto the space of $L_0(u)$ yields a residual matrix that has small nuclear norm. Following Candès and Recht (2009), let $L_0(u) = R(u)\Sigma(u)S(u)'$ be a singular value decomposition of $L_0(u)$. Let $\Phi(u)$ be the space of matrices defined by $\Phi(u) \equiv \{M \in \mathbb{R}^{N\times T} : \exists A \in \mathbb{R}^{r(u)\times T} \text{ and } B \in \mathbb{R}^{N\times r(u)} \text{ s.t. } M = R(u)A + BS(u)'\}$. The linear projection of a generic $N \times T$ matrix W onto this space is

$$P_{\Phi(u)}W = R(u)R(u)'W + WS(u)S(u)' - R(u)R(u)'WS(u)S(u)',$$

and its orthogonal projection is $P_{\Phi^{\perp}(u)}W = \big(I_{N\times N} - R(u)R(u)'\big)W\big(I_{T\times T} - S(u)S(u)'\big)$. In principle, we hope $P_{\Phi(u)}\hat{L}(u)$ is sufficiently large in nuclear norm compared to $P_{\Phi^{\perp}(u)}\hat{L}(u)$.

To formalize the idea, let $\hat{\Delta}_\beta(u) = \hat{\beta}(u) - \beta_0(u)$ and $\hat{\Delta}_L(u) = \hat{L}(u) - L_0(u)$. For some positive constants C_1 and C_2, define the restricted set as follows:

$$\begin{aligned}\mathcal{R}_u \equiv \big\{(\Delta_\beta, \Delta_L) : &\big(\lambda - \frac{C_2\sqrt{N\vee T}}{NT}\big)||P_{\Phi^{\perp}(u)}\Delta_L||_* \\ &\leq C_1\sqrt{\frac{p\log(NT)}{NT}}||\Delta_\beta||_F + \big(\lambda + \frac{C_2\sqrt{N\vee T}}{NT}\big)||P_{\Phi(u)}\Delta_L||_*\big\} \qquad (3.3.1)\end{aligned}$$

For large enough λ, if $(\hat{\Delta}_\beta(u), \hat{\Delta}_L(u)) \in \mathcal{R}_u$, the estimation error $\hat{\Delta}_L(u)$ projected on the orthogonal space $\Phi^{\perp}(u)$ of $L_0(u)$ is indeed at most of the same order of that projected on to $\Phi(u)$. Similar notions of the restricted set can also be seen in such as Negahban and Wainwright (2011) for low-rank matrix recovery, Belloni and Chernozhukov (2011) for high-dimensional quantile regression, and Chernozhukov et al. (2019) and Moon and Weidner (2019) for mean regression with interactive fixed effects.

Under the following assumption, I show that $(\hat{\Delta}_\beta(u), \hat{\Delta}_L(u)) \in \mathcal{R}_u$ with high proba-

bility.

Assumption 3.3.1. *i)* $(X_{it}, U_{it} : i \in \{1, ..., N\}, t \in \{1, ..., T\})$ *are i.i.d., and all the regressors have finite variances; ii)* $X_{it} \perp\!\!\!\perp U_{it}$.

For simplicity, I only consider the i.i.d. case. The results can be extended to non-i.i.d. case, for example, where the covariates are stationary and for each regressor X_j, $\frac{1}{NT}\sum_{i,t} X_{j,it}^2$ converges in probability to a constant.

Lemma 3.3.1. *Let* $\mathcal{U}$ *be a compact subset in the interior of* $[0, 1]$. *Under Assumption 3.3.1,* $(\hat{\Delta}_\beta(u), \hat{\Delta}_L(u)) \in \mathcal{R}_u$ *uniformly in* $u \in \mathcal{U}$ *with high probability.*

Proof. See Appendix C.1. □

From now on, I set $\lambda = \frac{2C_2\sqrt{N \vee T}}{NT}$. Then the restricted set can be simplified as

$$\mathcal{R}_u \equiv \left\{(\Delta_\beta, \Delta_L) : ||P_{\Phi^\perp(u)}\Delta_L||_* \leq \frac{C_1\sqrt{p\log(NT)(N \wedge T)}}{C_2}||\Delta_\beta||_F + 3||P_{\Phi(u)}\Delta_L||_*\right\} \tag{3.3.2}$$

The key property delivered by the restricted set and Lemma 3.3.1 is that the nuclear norm and the Frobenius norm of the estimation error for the low-rank component can be of the same order. To see this, by the definition of $P_{\Phi(u)}$, $\text{rank}(P_{\Phi(u)}A) \leq 3r(u) \leq 3\bar{r}$ for a generic $N \times T$ matrix A. Since $||A||_F \leq ||A||_* \leq \sqrt{\text{rank}(A)}||A||_F$, under the event that $(\hat{\Delta}_\beta(u), \hat{\Delta}_L(u)) \in \mathcal{R}_u$, we have

$$\begin{aligned}
||\hat{\Delta}_L(u)||_F \leq ||\hat{\Delta}_L(u)||_* =& ||P_{\Phi(u)}\hat{\Delta}_L(u)||_* + ||P_{\Phi^\perp(u)}\hat{\Delta}_L(u)||_* \\
\leq& 4||P_{\Phi(u)}\hat{\Delta}_L(u)||_* + \frac{C_1\sqrt{p\log(NT)(N \wedge T)}}{C_2}||\Delta_\beta||_F \\
\leq& 4\sqrt{3r(u)}||P_{\Phi(u)}\hat{\Delta}_L(u)||_F + \frac{C_1\sqrt{p\log(NT)(N \wedge T)}}{C_2}||\Delta_\beta||_F \\
\leq& 4\sqrt{3\bar{r}}||\hat{\Delta}_L(u)||_F + \frac{C_1\sqrt{p\log(NT)(N \wedge T)}}{C_2}||\Delta_\beta||_F
\end{aligned}$$

Hence, if the first term on the right hand side of the last inequality dominates the second, $||\hat{\Delta}_L(u)||_F$ and $||\hat{\Delta}_L(u)||_*$ are in the same order; in contrast, in general, the nuclear norm

could be $\sqrt{N \wedge T}$ times of the Frobenius norm. This property will be repeatedly used to prove the main result in the following section.

3.4 The Main Results

In this section, I provide the uniform rate of convergence for $\hat{\beta}(u)$ and $\hat{L}(u)$. I will also briefly discuss post-regularization procedures to obtain consistent estimators for the individual- and the time-fixed effects, the number of the fixed effects, and a debiased slope estimator. Let us begin by introducing the following assumptions.

Assumption 3.4.1 (Conditional Density). *There exist positive constants $\underline{f}$ and $\bar{f}'$ such that i) the conditional density of the unobservable $V_{it}(u)$ in equation (3.2.2) satisfies $f_{V_{it}(u)|X_{it}}(0) \geq \underline{f} > 0$ uniformly in $u \in \mathcal{U}$ almost surely, and ii) the derivative of the density is uniformly bounded in absolute value by $\bar{f}'$.*

Assumption 3.4.2 (Bounds on Magnitude). $\max_{j=1,\ldots,p} ||X_j||_\infty = o_p\left(\frac{\sqrt{N \wedge T}}{\alpha \sqrt{\log(NT)}}\right)$. $||L_0(u)||_\infty \leq \alpha$.

Assumption 3.4.3 (Smoothness). *For any $u' \neq u \in \mathcal{U}$, there exist $\zeta_1, \zeta_2 > 0$ such that*

$$\begin{aligned} ||\beta_0(u') - \beta_0(u)||_F \leq & \zeta_1 |u' - u|, \\ \frac{1}{\sqrt{NT}} ||L_0(u') - L_0(u)||_F \leq & \zeta_2 |u' - u| \end{aligned}$$

Assumption 3.4.4 (Identification). *i) $\mathbb{E}(X_{it} X_{it}')$ is invertible; ii) Denote $(P_{\Phi(u)} \boldsymbol{X})_{it} = ((P_{\Phi(u)} X_1)_{it}, \ldots, (P_{\Phi(u)} X_p)_{it})'$ and $(P_{\Phi^\perp(u)} \boldsymbol{X})_{it} = ((P_{\Phi^\perp(u)} X_1)_{it}, \ldots, (P_{\Phi^\perp(u)} X_p)_{it})'$. Assume the following holds uniformly in $u \in \mathcal{U}$:*

$$\mathbb{E} \sum_{i,t} \left((P_{\Phi^\perp(u)} \boldsymbol{X})_{it} (P_{\Phi^\perp(u)} \boldsymbol{X})_{it}' - (\log(NT) \bar{r}) (P_{\Phi(u)} \boldsymbol{X})_{it} (P_{\Phi(u)} \boldsymbol{X})_{it}' \right) \text{ is positive definite.}$$

Assumptions 3.4.1 and 3.4.2 guarantee that the objective function can be bounded

from below by a quadratic function. Assumption 3.4.1 is standard in quantile regression. However, with interactive fixed effects, a stronger assumption is often made in the literature to overcome the theoretical difficulty caused by estimating the $N \times T$ matrix $L_0(u)$; unlike Assumption 3.4.1, the conditional density is often required to be bounded away from o on compact intervals around o (e.g. Ando and Bai (2020) and Chen, Dolado and Gonzalo (2019)). In this paper, I develop a new argument under which the conditional density being bounded away from 0 at $V_{it}(u) = 0$ suffices. To apply the argument, I need to control the magnitude $L_{0,it}(u)$ and X_{it} by Assumption 3.4.2. As I allow α to be a polynomial of $\log(NT)$, these restrictions are mild in practice.

To illustrate the role Assumption 3.4.2 plays, let us consider a simple case with no covariates. Consider the expectation of the difference in the check functions evaluated at the true parameter $L_0(u)$ and $L_0(u) + \Delta_L$ for some deviation $\Delta_L \neq 0$. We hope the difference can be bounded from below by a positive function. By Knight's identity (Knight, 1998),

$$\begin{aligned}
\mathbb{E}\big(\boldsymbol{\rho}_u(V_{it}(u) + \Delta_{L,it}) - \boldsymbol{\rho}_u(V_{it}(u))\big) &= \sum_{i,t} \mathbb{E} \int_0^{\Delta_{L,it}} (\mathbb{1}_{V_{it}(u) \leq s} - \mathbb{1}_{V_{it}(u) \leq 0}) ds \\
&= \sum_{i,t} \int_0^{\Delta_{L,it}} (F_{V_{it}(u)}(s) - F_{V_{it}(u)}(0)) ds \\
&= \sum_{i,t} \int_0^{\Delta_{L,it}} (s f_{V_{it}(u)}(0) + \frac{s^2}{2} f'_{V_{it}(u)}(\tilde{s})) ds \\
&\geq \frac{1}{2} \underline{f} \sum_{i,t} \Delta_{L,it}^2 - \frac{1}{6} \bar{f}' \sum_{i,t} |\Delta_{L,it}|^3
\end{aligned}$$

The main challenge here is that $\sum_{i,t} |\Delta_{L,it}|^3$ may be of higher order than $\sum_{i,t} \Delta_{L,it}^2$, even if $\frac{1}{NT} \sum_{i,t} \Delta_{L,it}^2$ is converging. If that is the case, the right hand side of the last inequality is negative.

To avoid this undesirable case, the existing literature either modifies the Taylor expansion step (the third equality) to expand $F_{V_{it}(u)}(s)$ only to the first order at the mean value ($f_{V_{it}(u)}(\tilde{s})$), and assume $f_{V_{it}(u)}(\cdot) \geq \underline{f} > 0$ on a large set around 0, or directly as-

sumes that the order of $\sum_{i,t} |\Delta_{L,it}|^3$ is not too large (Belloni et al., 2019). In contrast, I develop a new argument that maintains Assumption 3.4.1 under the low-level primitive conditions in Assumption 3.4.2. The argument is based on two observations. First, the integral $\int_0^{\Delta_{L,it}} (\mathbb{1}_{V_{it}(u)\leq s} - \mathbb{1}_{V_{it}(u)\leq 0}) ds$ is decreasing in $|\Delta_{L,it}|$ (see Lemma C.2.2). Second, if $|\Delta_{L,it}| \leq 1$, then $\sum_{i,t} |\Delta_{L,it}|^3 < \sum_{i,t} \Delta_{L,it}^2$. Though we do not know whether $|\Delta_{L,it}|$ is smaller than 1 for each i and t, as long as it is bounded (justified by Assumption 3.4.2 and the inequality constraint in the definition of the estimator (3.2.3)), we can divide it by the bound in the upper limit of the integral. The resulting integral is then smaller than the integral under investigation by monotonicity and is positive. The details can be found in Lemma C.2.1 in Appendix C.2.

Assumption 3.4.3 is needed for uniformity. The first part is the same as in Belloni and Chernozhukov (2011). The second part is the counterpart for $L_0(\cdot)$. Note that the condition rules out the case where $r(u)$ changes on $\mathcal{U}$. To see this, suppose there exists u_0 in the interior of $\mathcal{U}$ such that $r(u) < r(u')$ for any $u < u_0 \leq u'$. Then $\frac{1}{NT}||L_0(u') - L_0(u)||_F$ can be bounded away from o by a constant uniformly in $|u' - u|$. However, if $r(u)$ only has finite number of jump-points in $[0, 1]$, uniformity on the union of compact interior subsets between jump-points can still be obtained by applying Theorem 3.4.1 to each of these subsets.

Assumption 3.4.4 is used to bound the estimation errors $\hat{\Delta}_\beta(u)$ and $\hat{\Delta}_L(u)$ separately. Without it, I can only obtain the rate for a weighted sum of them. This is a sufficient condition in our context for the widely assumed "restricted strong convexity" or the "restricted identifiability" condition in the related literature (e.g. Negahban and Wainwright (2011, 2012), Agarwal, Negahban and Wainwright (2012), Negahban et al. (2012), Belloni and Chernozhukov (2011), Belloni et al. (2019), Chernozhukov et al. (2019), Moon and Weidner (2019), etc.). Assumption 3.4.4 says that the covariates need to be sufficiently far from the space of $L_0(u)$ uniformly in $u \in \mathcal{U}$. Otherwise, for instance, suppose for some j, $X_j = L_0(u)$, that is $X_j = P_{\Phi(u)} X_j$, then $\beta_{0,j}(u)$ is not identified due to perfect

collinearity.

Under these assumptions, we have the main theorem of this paper.

Theorem 3.4.1. *Under Assumptions 3.3.1-3.4.4,*

$$\sup_{u\in\mathcal{U}} \; ||\hat{\Delta}_\beta(u)||_F^2 + \frac{1}{NT}||\hat{\Delta}_L(u)||_F^2 = O_p\left(\frac{\alpha^4 \bar{f}'^4 \log(NT)}{\underline{f}^8}\left(\frac{p\log(NT)}{NT} \vee \frac{\bar{r}}{N\wedge T}\right)\right) \quad (3.4.1)$$

Proof. See Appendix C.2. □

From Theorem 3.4.1, the rate of convergence depends on the rank of $L_0(u)$ (captured by $\bar{r}$), the number of regressors (p), and the magnitude of elements in $L_0(u)$ (α). Note that in the parentheses, $\frac{p}{NT}$ would be the rate of convergence of the standard quantile regression estimator if $L_0(u)$ was not in the model. Meanwhile, $\frac{\bar{r}}{N\wedge T}$ is the minimax optimal rate of convergence for the low-rank matrix estimator using nuclear norm regularization (see Agarwal, Negahban and Wainwright (2012) and Negahban and Wainwright (2012) for instance), and is identical to the case for mean regression (Athey et al. (2018), Moon and Weidner (2019) and Chernozhukov et al. (2019)).

From Theorem 3.4.1 and the definition of the restricted set, it is straightforward to show the following corollary.

Corollary 1. *Under Assumptions 3.3.1-3.4.4,*

$$\sup_{u\in\mathcal{U}} \; \frac{1}{N\wedge T}||\hat{\Delta}_L(u)||_* = O_p\left(\frac{\alpha^2 \bar{f}'^2 \sqrt{\log(NT)}}{\underline{f}^4}\left(\frac{\sqrt{p\log(NT)}}{N\wedge T} \vee \frac{\sqrt{\bar{r}(N\vee T)}}{N\wedge T}\right)\right) \quad (3.4.2)$$

To see why this is true, note that being in the restricted set, the rate in Theorem 3.4.1 implies that $||P_{\Phi^\perp(u)}\hat{\Delta}_L(u)||_*$ is bounded by $||P_{\Phi(u)}\hat{\Delta}_L(u)||_*$, which is further bounded by $\sqrt{3\bar{r}}||P_{\Phi(u)}\hat{\Delta}_L(u)||_F$ as shown in Section 3.3. Low-rankness of $L_0(u)$ plays a key role here because only then the estimation error in the nuclear norm is of the same order of that in the Frobenius norm. Corollary 1 implies that if the panel data matrix is not too

"tall" or "fat", i.e., the orders N and T are not too different, the average of the singular values of $L_0(u)$ can be uniformly estimated.

From Theorem 3.4.1 and Corollary 1, we can consistently estimate the rank of $L_0(u)$ by thresholding, similar to Moon and Weidner (2019) and Chernozhukov et al. (2019). To sketch the idea, denote the singular values of $\hat{L}(u)$ and $L_0(u)$ by $\hat{\sigma}_1(u), \cdots, \hat{\sigma}_{N\wedge T}(u)$ and $\sigma_1(u), \cdots, \sigma_{N\wedge T}(u)$ in descending order. Given that both p and $\bar{r}$ are fixed, by Weyl's theorem and Theorem 3.4.1, $\max_k |\hat{\sigma}_k(u) - \sigma_k(u)|$ is bounded by $O(\sqrt{N \vee T})$ up to a multiplicative polynomial of $\log(NT)$. Since $\sigma_k(u) = 0$ for all $k > r(u)$, the singular values of $\hat{L}(u)$ are well separated if $F_t(u)$ are strong factors such that the largest $r(u)$ singular values of $L_0(u)$ are $O(N \wedge T)$: $\hat{\sigma}_k(u)$ is either the order of $(N \wedge T) + \sqrt{N \vee T}$ (for $k \leq r(u)$), or $\sqrt{N \vee T}$ (for $k > r(u)$). Therefore, by setting a threshold of any order in between, a consistent estimator $\hat{r}(u)$ can be obtained by simply counting the number of the estimated singular values that are above the threshold.

To correct the bias that regularization brings in, achieve the optimal rate for $\hat{\beta}(u)$ (i.e., $O(\frac{1}{\sqrt{NT}})$), and establish the inference theory, I conjecture that one can follow similar post-regularization procedures proposed in Moon and Weidner (2019) and Chernozhukov et al. (2019) to obtain an asymptotically normal estimator for $\beta_0(u)$. Specifically, construct $\hat{F}_t$ and $\hat{\Lambda}_i$ by singular value decomposition with only those corresponding to the largest $\hat{r}(u)$ singular values kept. Then iteratively minimize the standard quantile regression objective function without penalty by setting $\hat{\beta}(u)$ as the initial guess. The iterations are similar to Ando and Bai (2020) but should only take a small number of steps and by consistency of $\hat{\beta}(u)$, the resulting iterative estimator would be close to $\beta_0(u)$ although the minimization problem is now globally nonconvex.

3.5 Monte Carlo Simulations

In this section, I illustrate the finite sample performance of the estimator using a simulation study. I consider the following data generating process adapted from Ando and Bai (2020):

$$Y_{it} = X_{it}'\beta_0(U_{it}) + \sum_{k=1}^{5} \mathbb{1}_k(U_{it}) F_{kt}\Lambda_{ki}(U_{it}) + V_{it}(U_{it})$$

where $U_{it} \sim \text{Unif}[0,1]$. X_{it} contains the following four regressors:

$$X_{1,it} = W_{1,it} + 0.02F_{1,t}^2 + 0.02\zeta_{1,i}^2, \quad X_{2,it} = W_{2,it}$$
$$X_{3,it} = W_{3,it} - 0.01F_{3,t}^2 + 0.02\zeta_{3,i}^2, \quad X_{4,it} = W_{4,it} - 0.01F_{4t}^2 + 0.03\zeta_{4,i}^2,$$

where all $W_{k,it}$, $\zeta_{k,i}$ and the factors F_{kt} are independently drawn from $\text{Unif}[0,2]$. The factor loadings satisfy $\Lambda_{ki}(U_{it}) = \zeta_{k,i} + 0.1U_{it}$. For $\beta_0(U_{it})$, $\beta_{0,1}(U_{it}) = \beta_{0,3}(U_{it}) = \beta_{0,4}(U_{it}) = -1 + 0.1U_{it}$ and $\beta_{0,2}(U_{it}) = 1 + 0.1U_{it}$. Finally, the indicator function satisfies:

$$\mathbb{1}_k(\cdot) = 1 \text{ for } k = 1,2,3, \quad \mathbb{1}_4(u) = \mathbb{1}(0.3 < u \leq 0.7), \quad \mathbb{1}_5(u) = \mathbb{1}(u > 0.7),$$

that is, the first three factors always affect Y_{it}, while the fourth and the fifth factor only affect Y_{it} when U_{it} is between 0.3 and 0.7 and above 0.7, respectively. Finally, $V_{it}(U_{it}) = G^{-1}(U_{it})$ where G is the cumulative distribution function of standard normal or student-t distribution with degree of freedom 8.

For implementation, I set $\lambda = \frac{1}{\sqrt{N}}$; the order is suggested by the theory. I then iteratively update $\beta(u)$ and $L(u)$ until convergence. I update $\beta(u)$ by pooled quantile regression subtracting $L(u)$ from Y. For $L(u)$, I adapt the *alternating directions* method proposed in Lin, Chen and Ma (2010) and Yuan and Yang (2013), which is also used in Candès et al. (2011). In our notation, the algorithm was originally designed for $u = 0.5$.

Table 3.1: Average Bias^2, Variance and RMSE of $\hat{\beta}(u)$ and $\hat{\beta}^{pooled}$

	A. Normal Error				
	Bias^2		Variance		$\|\|\hat{\Delta}_L(u)\|\|_F^2/\|\|L_0(u)\|\|_F^2$
	Penalized	Pooled	Penalized	Pooled	Penalized
$u = 0.2$	$2.82 \cdot 10^{-4}$	$2.6 \cdot 10^{-3}$	$1.25 \cdot 10^{-4}$	$5.48 \cdot 10^{-4}$	0.07
$u = 0.5$	$7.40 \cdot 10^{-6}$	$2 \cdot 10^{-3}$	$1.46 \cdot 10^{-4}$	$5.72 \cdot 10^{-4}$	0.07
$u = 0.8$	$2.22 \cdot 10^{-5}$	$2.1 \cdot 10^{-3}$	$5.99 \cdot 10^{-4}$	$9.62 \cdot 10^{-4}$	0.06
	B. Student-t Error				
	Bias^2		Variance		$\|\|\hat{\Delta}_L(u)\|\|_F^2/\|\|L_0(u)\|\|_F^2$
	Penalized	Pooled	Penalized	Pooled	Penalized
$u = 0.2$	$2.82 \cdot 10^{-4}$	$2.6 \cdot 10^{-3}$	$1.45 \cdot 10^{-4}$	$5.40 \cdot 10^{-4}$	0.08
$u = 0.5$	$4.54 \cdot 10^{-6}$	$2 \cdot 10^{-3}$	$1.65 \cdot 10^{-4}$	$7.53 \cdot 10^{-4}$	0.07
$u = 0.8$	$2.06 \cdot 10^{-5}$	$2 \cdot 10^{-3}$	$7.17 \cdot 10^{-4}$	$1.1 \cdot 10^{-3}$	0.06

I modify it to accommodate any other $u \in (0, 1)$.

Table 3.1 presents the results based on 100 simulation replications. The bias squared and the variances are the average for the four components in $\hat{\beta}(u)$. Column Penalized contains the results using our estimator. Column Pooled are results obtained from quantile regression ignoring the interactive fixed effects. As the regularized estimator is biased by the penalty, the pooled regression provides a reference to compare the bias from regularization with the bias from endogeneity. As seen from the results, the squared bias ignoring the interactive fixed effects is 10-1000 times than that of our estimator. The performance of the estimated low-rank component is good too. The sum of squared estimation error is between 6% and 8% of the magnitude of the true low-rank component.

3.6 Concluding Remarks

In this paper, I propose nuclear norm regularized quantile regression for panel data models with interactive fixed effects. I derive uniform rates of convergence for both the slope coefficients and the low-rank component of the interactive fixed effects. The rate for the latter is nearly optimal.

The results can be extended to models with heterogeneous effects. To see it, note that in the rate, the number of coefficients can be as large as $O(N \vee T)$, only slowing down convergence by a factor of $\log(NT)$. Therefore, almost the same rate could be maintained if p is fixed but $\beta_0(u)$ is i- or t-dependent. Similarly, for homogeneous effect models, the number of the regressors p can be potentially allowed to diverge to infinity at the rate of $(N \vee T)$. Finally, as the penalty introduces bias, post-regularization estimation procedures described in the paper may be needed to obtain the asymptotic distribution and establish inference theory. This is left for future work.

Epilogue

In the dissertation, I show how appropriately imposed structures help elicit information from limited or noisy data. The research questions answered by the three chapters are only a few examples where difficulties in identification or estimation arise in such data. In practice, there could be many other cases. For instance, extrapolation in regression discontinuity design is needed when the data are limited, and matrix completion/prediction of missing values with noises falls into the second scenario. Moving forward, I would like to explore further into these areas of research to find out structures and develop new econometric tools to address such issues.

Bibliography

Abrevaya, Jason, and Christian M. Dahl. 2008. "The effects of birth inputs on birthweight: evidence from quantile estimation on panel data." *Journal of Business & Economic Statistics*, 26(4): 379–397.

Agarwal, Alekh, Sahand Negahban, and Martin J. Wainwright. 2012. "Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions." *The Annals of Statistics*, 40(2): 1171–1197.

Ambrosetti, Antonio, and Giovanni Prodi. 1995. *A Primer of Nonlinear Analysis.* Vol. 34 of *Cambridge Studies in Advanced Mathematics*. Cambridge: Cambridge University Press.

Ando, Tomohiro, and Jushan Bai. 2020. "Quantile co-movement in financial markets: A panel quantile model with unobserved heterogeneity." *Journal of the American Statistical Association*, 115(529): 266–279.

Angrist, Joshua D., and Guido W. Imbens. 1995. "Two-stage least squares estimation of average causal effects in models with variable treatment intensity." *Journal of the American statistical Association*, 90(430): 431–442.

Athey, Susan, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. 2018. "Matrix completion methods for causal panel data models." *arXiv preprint arXiv:1710.10251*.

Bach, Francis R. 2008. "Consistency of trace norm minimization." *Journal of Machine Learning Research*, 9(Jun): 1019–1048.

Bai, Jushan. 2009. "Panel data models with interactive fixed effects." *Econometrica*, 77(4): 1229–1279.

Bai, Jushan, and Junlong Feng. 2019. "Robust principal component analysis with non-sparse errors." *arXiv preprint arXiv:1902.08735*.

Bai, Jushan, and Kunpeng Li. 2012. "Statistical analysis of factor models of high dimension." *The Annals of Statistics*, 40(1): 436–465.

Bai, Jushan, and Serena Ng. 2019*a*. "Matrix completion, counterfactuals, and factor analysis of missing data." *arXiv preprint arXiv:1910.06677*.

Bai, Jushan, and Serena Ng. 2019*b*. "Rank regularized estimation of approximate factor models." *Journal of Econometrics*, 212(1): 78–96.

Belloni, Alexandre, and Victor Chernozhukov. 2011. "ℓ1-penalized quantile regression in high-dimensional sparse models." *The Annals of Statistics*, 39(1): 82–130.

Belloni, Alexandre, Mingli Chen, Oscar Hernan Madrid Padilla, and Zixuan (Kevin) Wang. 2019. "High dimensional latent panel quantile regression with an application to asset pricing." *arXiv preprint arXiv:1912.02151*.

Bouwmans, Thierry, Andrews Sobral, Sajid Javed, Soon Ki Jung, and El-Hadi Zahzah. 2017. "Decomposition into low-rank plus additive matrices for background/foreground separation: A review for a comparative evaluation with a large-scale dataset." *Computer Science Review*, 23: 1–71.

Brahma, Pratik Prabhanjan, Yiyuan She, Shijie Li, Jiade Li, and Dapeng Wu. 2018. "Reinforced robust principal component pursuit." *IEEE Transactions on Neural Networks and Learning Systems*, 29(5): 1525–1538.

Caetano, Carolina, and Juan C. Escanciano. 2018. "Identifying multiple marginal effects with a single instrument." *Working Paper*.

Cai, Jian-Feng, Emmanuel J. Candès, and Zuowei Shen. 2010. "A singular value thresholding algorithm for matrix completion." *SIAM Journal on Optimization*, 20(4): 1956–1982.

Canay, Ivan A. 2011. "A simple approach to quantile regression for panel data." *The Econometrics Journal*, 14(3): 368–386.

Candès, Emmanuel J., and Benjamin Recht. 2009. "Exact matrix completion via convex optimization." *Foundations of Computational mathematics*, 9(6): 717.

Candès, Emmanuel J., and Terence Tao. 2010. "The power of convex relaxation: Near-optimal matrix completion." *IEEE Transactions on Information Theory*, 56(5): 2053–2080.

Candès, Emmanuel J., Xiaodong Li, Yi Ma, and John Wright. 2011. "Robust principal component analysis?" *Journal of the ACM (JACM)*, 58(3): 11.

Card, David. 1995. "Using geographic variation in college proximity to estimate the return to schooling." In *Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp*. ed. Louis N. Christofides, E. Kenneth Grant, and Roebert Swidinsky, 201-222. Toronto: University of Toronto Press.

Carneiro, Pedro, James J. Heckman, and Edward J. Vytlacil. 2011. "Estimating marginal returns to education." *American Economic Review*, 101(6): 2754–81.

Cattell, R. B. 1978. *The Use of Factor Analysis in Behavioral and Life Sciences.* New York: Plenum.

Chandrasekaran, Venkat, Sujay Sanghavi, Pablo A. Parrilo, and Alan S. Willsky. 2011. "Rank-sparsity incoherence for matrix decomposition." *SIAM Journal on Optimization*, 21(2): 572–596.

Chao, Shih-Kang, Wolfgang K. Härdle, and Ming Yuan. 2019. "Factorisable multitask quantile regression." *arXiv preprint arXiv:1507.03833*.

Chen, Han, Garvesh Raskutti, and Ming Yuan. 2019. "Non-convex projected gradient descent for generalized low-rank tensor regression." *The Journal of Machine Learning Research*, 20(1): 172–208.

Chen, Liang. 2019. "Two-step estimation of quantile panel data models with interactive fixed effects." *Working Paper*.

Chen, Liang, Juan J. Dolado, and Jesus Gonzalo. 2019. "Quantile factor models." *arXiv preprint arXiv:1911.02173*.

Chen, Xiaohong. 2007. "Large sample sieve estimation of semi-nonparametric models." In *Handbook of Econometrics, Vol. 6B*. ed. James J. Heckman and Edward E. Leamer, 5549–5632. Amsterdam: Elsevier.

Chen, Xiaohong, and Demian Pouzo. 2012. "Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals." *Econometrica*, 80(1): 277–321.

Chen, Xiaohong, and Demian Pouzo. 2015. "Sieve Wald and QLR inferences on semi/nonparametric conditional moment models." *Econometrica*, 83(3): 1013–1079.

Chen, Xiaohong, Victor Chernozhukov, Sokbae Lee, and Whitney K. Newey. 2014. "Local identification of nonparametric and semiparametric models." *Econometrica*, 82(2): 785–809.

Chernozhukov, Victor, and Christian Hansen. 2005. "An IV model of quantile treatment effects." *Econometrica*, 73(1): 245–261.

Chernozhukov, Victor, Christian Hansen, Yuan Liao, and Yinchu Zhu. 2019. "Inference for heterogeneous effects using low-rank estimations." *arXiv preprint arXiv:1812.08089*.

Chernozhukov, Victor, Guido W. Imbens, and Whitney K. Newey. 2007. "Instrumental variable estimation of nonseparable models." *Journal of Econometrics*, 139(1): 4–14.

Chernozhukov, Victor, Han Hong, and Elie Tamer. 2007. "Estimation and confidence regions for parameter sets in econometric models." *Econometrica*, 75(5): 1243–1284.

Chesher, Andrew. 2003. "Identification in nonseparable models." *Econometrica*, 71(5): 1405–1441.

Chesher, Andrew. 2004. "Identification in additive error models with discrete endogenous variables." *cemmap Working Paper, CWP11/04*.

Cunha, Flavio, James J. Heckman, and Salvador Navarro. 2007. "The identification and economic content of ordered choice models with stochastic thresholds." *International Economic Review*, 48(4): 1273–1309.

Das, Mitali. 2005. "Instrumental variables estimators of nonparametric models with discrete endogenous regressors." *Journal of Econometrics*, 124(2): 335–361.

De Marco, Giuseppe, Gianluca Gorni, and Gaetano Zampieri. 2014. "Global inversion of functions: An introduction." *arXiv preprint arXiv:1410.7902*.

D'Haultfœuille, Xavier, and Philippe Février. 2015. "Identification of nonseparable triangular models with discrete instruments." *Econometrica*, 83(3): 1199–1210.

Fan, Jianqing, Weichen Wang, and Yiqiao Zhong. 2018. "An ℓ_∞ eigenvector perturbation bound and its application to robust covariance estimation." *Journal of Machine Learning Research*, 18: 1–42.

Feng, Qian, Quang Vuong, and Haiqing Xu. 2020. "Estimation of Heterogeneous Individual Treatment Effects With Endogenous Treatments." *Journal of the American Statistical Association*, 115(529): 231–240.

Florens, Jean-Pierre, James J. Heckman, Costas Meghir, and Edward Vytlacil. 2008. "Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects." *Econometrica*, 76(5): 1191–1206.

Galvao, Antonio F., and Kengo Kato. 2016. "Smoothed quantile regression for panel data." *Journal of Econometrics*, 193(1): 92–112.

Galvao, Antonio F., Carlos Lamarche, and Luiz Renato Lima. 2013. "Estimation of censored quantile regression for panel data with fixed effects." *Journal of the American Statistical Association*, 108(503): 1075–1089.

Ganesh, Arvind, John Wright, Xiaodong Li, Emmanuel J. Candes, and Yi Ma. 2010. "Dense error correction for low-rank matrices via principal component pursuit." In *2010 IEEE International Symposium on Information Theory*. 1513–1517.

Goldberg, Lewis R. 1990. "An alternative "description of personality": The big-five factor structure." *Journal of Personality and Social Psychology*, 59(6): 1216.

Gross, David. 2011. "Recovering low-rank matrices from few coefficients in any basis." *IEEE Transactions on Information Theory*, 57(3): 1548–1566.

Gross, David, Yi-Kai Liu, Steven T. Flammia, Stephen Becker, and Jens Eisert. 2010. "Quantum state tomography via compressed sensing." *Physical Review Letters*, 105(15): 150401.

Guerre, Emmanuel, Isabelle Perrigne, and Quang Vuong. 2000. "Optimal nonparametric estimation of first-price auctions." *Econometrica*, 68(3): 525–574.

Gunsilius, Florian. 2018. "Point-identification in multivariate nonseparable triangular models." *arXiv preprint arXiv:1806.09680*.

Hansen, Bruce E. 2004. "Nonparametric estimation of smooth conditional distributions." *Working Paper, Department of Economics, University of Wisconsin, Madison*.

Harding, Matthew, and Carlos Lamarche. 2014. "Estimating and testing a quantile regression model with interactive effects." *Journal of Econometrics*, 178: 101–113.

Härdle, Wolfgang, J. S. Marron, and Matt Wand. 1990. "Bandwidth choice for density derivatives." *Journal of the Royal Statistical Society: Series B (Methodological)*, 52(1): 223–232.

Heckman, James J, and Edward J Vytlacil. 1999. "Local instrumental variables and latent variable models for identifying and bounding treatment effects." *Proceedings of the national Academy of Sciences*, 96(8): 4730–4734.

Heckman, James J, and Edward Vytlacil. 2001. "Policy-relevant treatment effects." *American Economic Review*, 91(2): 107–111.

Heckman, James J., and Edward Vytlacil. 2005. "Structural equations, treatment effects, and econometric policy evaluation." *Econometrica*, 73(3): 669–738.

Heckman, James J, Sergio Urzua, and Edward Vytlacil. 2006. "Understanding instrumental variables in models with essential heterogeneity." *The Review of Economics and Statistics*, 88(3): 389–432.

Heckman, James J., Sergio Urzua, and Edward Vytlacil. 2008. "Instrumental variables in models with multiple outcomes: The general unordered case." *Annales d'Economie et de Statistique*, 91/92: 151–174.

Hsu, Daniel, Sham M. Kakade, and Tong Zhang. 2011. "Robust matrix decomposition with sparse corruptions." *IEEE Transactions on Information Theory*, 57(11): 7221–7234.

Huang, Liquan, Umair Khalil, and Neşe Yıldız. 2019. "Identification and estimation of a triangular model with multiple endogenous variables and insufficiently many instrumental variables." *Journal of Econometrics*, 208(2): 346–366.

Ichimura, Hidehiko, and Christopher Taber. 2000. "Direct estimation of policy impacts." *NBER Technical Working Paper No. 254*.

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and estimation of local average treatment effects." *Econometrica*, 62(2): 467–475.

Imbens, Guido W., and Whitney K. Newey. 2009. "Identification and estimation of triangular simultaneous equations models without additivity." *Econometrica*, 77(5): 1481–1512.

Jain, Prateek, Ambuj Tewari, and Purushottam Kar. 2014. "On iterative hard thresholding methods for high-dimensional m-estimation." 685–693.

Jain, Prateek, Nikhil Rao, and Inderjit S Dhillon. 2016. "Structured sparse regression via greedy hard thresholding." 1516–1524.

Kapur, Arnav, Kshitij Marwah, and Gil Alterovitz. 2016. "Gene expression prediction using low-rank matrix completion." *BMC Bioinformatics*, 17(1): 243.

Kato, Kengo, Antonio F. Galvao Jr, and Gabriel V. Montes-Rojas. 2012. "Asymptotics for panel quantile regression models with individual effects." *Journal of Econometrics*, 170(1): 76–91.

Kline, Patrick, and Christopher R. Walters. 2016. "Evaluating public programs with close substitutes: The case of Head Start." *The Quarterly Journal of Economics*, 131(4): 1795–1848.

Knight, Keith. 1998. "Limiting distributions for L_1 regression estimators under general conditions." *Annals of Statistics*, 26(2): 755–770.

Koenker, Roger. 2004. "Quantile regression for longitudinal data." *Journal of Multivariate Analysis*, 91(1): 74–89.

Lamarche, Carlos. 2010. "Robust penalized quantile regression estimation for panel data." *Journal of Econometrics*, 157(2): 396–408.

Ledoux, Michel, and Michel Talagrand. 1991. *Probability in Banach Spaces: Isoperimetry and Processes.* Ergebnisse der Mathematik und ihrer Grenzgebiete, Springer-Verlag.

Lee, Sokbae, and Bernard Salanié. 2018. "Identifying effects of multivalued treatments." *Econometrica*, 86(6): 1939–1963.

Lewbel, Arthur. 2007. "A local generalized method of moments estimator." *Economics Letters*, 94(1): 124–128.

Lin, Zhouchen, Minming Chen, and Yi Ma. 2010. "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices." *arXiv preprint arXiv:1009.5055*.

Li, Qi, and Jeffrey S. Racine. 2008. "Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data." *Journal of Business & Economic Statistics*, 26(4): 423–434.

Mack, Yue-pok, and Bernard W. Silverman. 1982. "Weak and strong uniform consistency of kernel regression estimates." *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 61(3): 405–415.

Masry, Elias. 1996. "Multivariate local polynomial regression for time series: uniform strong consistency and rates." *Journal of Time Series Analysis*, 17(6): 571–599.

Matzkin, Rosa L. 1993. "Nonparametric identification and estimation of polychotomous choice models." *Journal of Econometrics*, 58(1-2): 137–168.

Matzkin, Rosa L. 2003. "Nonparametric estimation of nonadditive random functions." *Econometrica*, 71(5): 1339–1375.

Moon, Hyungsik Roger, and Martin Weidner. 2015. "Linear regression for panel with unknown number of factors as interactive fixed effects." *Econometrica*, 83(4): 1543–1579.

Moon, Hyungsik Roger, and Martin Weidner. 2019. "Nuclear norm regularized estimation of panel regression models." *arXiv preprint arXiv:1810.10987*.

Negahban, Sahand, and Martin J. Wainwright. 2011. "Estimation of (near) low-rank matrices with noise and high-dimensional scaling." *The Annals of Statistics*, 39(2): 1069–1097.

Negahban, Sahand, and Martin J. Wainwright. 2012. "Restricted strong convexity and weighted matrix completion: Optimal bounds with noise." *Journal of Machine Learning Research*, 13(May): 1665–1697.

Negahban, Sahand N., Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. 2012. "A unified framework for high-dimensional analysis of *M*-estimators with decomposable regularizers." *Statistical Science*, 27(4): 538–557.

Newey, Whitney K., and Daniel. McFadden. 1994. "Large sample estimation and hypothesis testing." In *Handbook of Econometrics, Vol. 4*. ed. Robert F. Engle and Daniel L. McFadden, 2111–2245. Amsterdam: Elsevier.

Newey, Whitney K., and James L. Powell. 2003. "Instrumental variable estimation of nonparametric models." *Econometrica*, 71(5): 1565–1578.

Newey, Whitney K., James L. Powell, and Francis Vella. 1999. "Nonparametric estimation of triangular simultaneous equations models." *Econometrica*, 67(3): 565–603.

Ortega, James M., and Werner C. Rheinboldt. 1970. *Iterative Solution of Nonlinear Equations in Several Variables.* New York: Academic Press.

Pesaran, M. Hashem. 2006. "Estimation and inference in large heterogeneous panels with a multifactor error structure." *Econometrica*, 74(4): 967–1012.

Ross, Stephen A. 1976. "The arbitrage theory of capital asset pricing." *Journal of Economic Theory*, 13(3): 341–360.

Shalev-Shwartz, Shai, Alon Gonen, and Ohad Shamir. 2011. "Large-scale convex minimization with a low-rank constraint." 329–336.

Silverman, Bernard W. 1986. *Density Estimation for Statistics and Data Analysis.* London: Chapman & Hall.

Tao, Min, and Xiaoming Yuan. 2011. "Recovering low-rank and sparse components of matrices from incomplete and noisy observations." *SIAM Journal on Optimization*, 21(1): 57–81.

Tao, Terence. 2012. *Topics in Random Matrix Theory.* Vol. 132 of *Graduate Studies in Mathematics*, American Mathematical Society.

Torgovitsky, Alexander. 2015. "Identification of nonseparable models using instruments with small support." *Econometrica*, 83(3): 1185–1197.

Torgovitsky, Alexander. 2017. "Minimum distance from independence estimation of nonseparable instrumental variables models." *Journal of Econometrics*, 199(1): 35–48.

United States Department of Health and Human Services. Administration for Children and Families. Office of Planning, Research and Evaluation. 2018-02-08. "Head Start Impact Study (HSIS), 2002-2006 [United States]." Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/ICPSR29462.v7.

van der Vaart, Aad W., and Jon A. Wellner. 1996. *Weak Convergence and Empirical Processes.* Springer.

Vuong, Quang, and Haiqing Xu. 2017. "Counterfactual mapping and individual treatment effects in nonseparable models with binary endogeneity." *Quantitative Economics*, 8(2): 589–610.

Vytlacil, Edward, and Neşe Yıldız. 2007. "Dummy endogenous variables in weakly separable models." *Econometrica*, 75(3): 757–779.

Wong, Raymond K. W., and Thomas Lee. 2017. "Matrix completion with noisy entries and outliers." *The Journal of Machine Learning Research*, 18(1): 5404–5428.

Wright, John, Arvind Ganesh, Kerui Min, and Yi Ma. 2013. "Compressive principal component pursuit." *Information and Inference: A Journal of the IMA*, 2(1): 32–68.

Xie, Pengtao, and Eric Xing. 2014. "Cauchy principal component analysis." *arXiv preprint arXiv:1412.6506*.

Xu, Huan, Constantine Caramanis, and Sujay Sanghavi. 2012. "Robust PCA via outlier pursuit." *IEEE Transactions on Information Theory*, 5(58): 3047–3064.

Yuan, Xiaoming, and Junfeng Yang. 2013. "Sparse and low-rank matrix decomposition via alternating direction method." *Pacific Journal of Optimization*, 9(1): 167.

Zhou, Zihan, Xiaodong Li, John Wright, Emmanuel Candes, and Yi Ma. 2010. "Stable principal component pursuit." In *2010 IEEE International Symposium on Information Theory*. 1518–1522.

Zhu, Zhihui, Qiuwei Li, Gongguo Tang, and Michael B Wakin. 2018. "Global optimality in low-rank matrix optimization." *IEEE Transactions on Signal Processing*, 66(13): 3614–3628.

Appendix A

Appendix to Chapter 1

A.1 General Cases

A.1.1 $|S(D)| > |S(Z)|$ and Multiple Endogenous Variables

Let us first discuss the general case of $|S(D)| > |S(Z)|$. For a given $\boldsymbol{x}_0$, at least $|S(D)| - |S(Z)|$ m-connected points are needed for identification. The size difference is not as formidable as it appears: When $|S(Z)|$ increases, the size of the m-connected points may increase at a faster rate. Recall Figure 1.1, each value of $z \in S(Z)$ induces an arm to grow m-connected points. Then, for instance, if $|S(Z)| = 3$, two matching points may be obtained by solving the following two equations:

$$\boldsymbol{p}(\boldsymbol{x}, z') = \boldsymbol{p}(\boldsymbol{x}_0, z) \text{ and } \boldsymbol{p}(\boldsymbol{x}, z'') = \boldsymbol{p}(\boldsymbol{x}_0, z)$$

Since z takes on 3 values, up to 6 matching points may be obtained even if each propensity score matching equation has only one solution. By induction, the number of matching points can be as many as $|S(Z)| \cdot (|S(Z)| - 1)$. With the variation in Z itself, a discrete IV taking on $|S(Z)|$ values may be able to identify a nonparametric model with an endogenous D with $|S(D)| = |S(Z)|^2$, instead of $|S(Z)|$ using the standard IV approach.

For the m-connected points, the number can be even larger.

Example A.1.1. *Consider an ordered choice model where* $Z \in \{0,1,2\}$. *Equivalently, define* $Z_1 \geq Z_2 \in \{0,1\}$ *such that* $Z = 0$ *if and only if* $Z_1 = Z_2 = 0$, $Z = 1$ *if and only if* $Z_1 = 1$ *and* $Z_2 = 0$, *and* $Z = 2$ *if and only if* $Z_1 = Z_2 = 1$. *Let the linear single index be* $X\beta + Z_1\alpha_1 + Z_2\alpha_2$, *then we have the following six matching points:*

$$
\begin{aligned}
(z=0):& \beta x_{m1} + \alpha_1 \cdot 1 + \alpha_2 \cdot 0 = x_0\beta + \alpha_1 \cdot 0 + \alpha_2 \cdot 0 \implies x_{m1} = x_0 - \frac{\alpha_1}{\beta} \\
& \beta x_{m2} + \alpha_1 \cdot 1 + \alpha_2 \cdot 1 = x_0\beta + \alpha_1 \cdot 0 + \alpha_2 \cdot 0 \implies x_{m2} = x_0 - \frac{\alpha_1 + \alpha_2}{\beta} \\
(z=1):& \beta x_{m3} + \alpha_1 \cdot 0 + \alpha_2 \cdot 0 = x_0\beta + \alpha_1 \cdot 1 + \alpha_2 \cdot 0 \implies x_{m3} = x_0 + \frac{\alpha_1}{\beta} \\
& \beta x_{m4} + \alpha_1 \cdot 1 + \alpha_2 \cdot 1 = x_0\beta + \alpha_1 \cdot 1 + \alpha_2 \cdot 0 \implies x_{m4} = x_0 - \frac{\alpha_2}{\beta} \\
(z=2):& \beta x_{m5} + \alpha_1 \cdot 0 + \alpha_2 \cdot 0 = x_0\beta + \alpha_1 \cdot 1 + \alpha_2 \cdot 1 \implies x_{m5} = x_0 + \frac{\alpha_1 + \alpha_2}{\beta} \\
& \beta x_{m6} + \alpha_1 \cdot 1 + \alpha_2 \cdot 0 = x_0\beta + \alpha_1 \cdot 1 + \alpha_2 \cdot 1 \implies x_{m6} = x_0 + \frac{\alpha_2}{\beta}
\end{aligned}
$$

Remark A.1.1. Note that the nonlinearity of Z's effect is needed to generate six matching points. If $\alpha_1 = \alpha_2$, then $x_{m1} = x_{m4}$ and $x_{m3} = x_{m6}$; only four matching points are generated.

We can further generalize our approach to the case of multiple discrete endogenous variables. Suppose there are M discrete endogenous variables $D_1, ..., D_M$ in a model. It is equivalent to recode them as one single endogenous variable D_0. For instance, if $S(D_1, ..., D_M) = S(D_1) \times \cdots \times S(D_M)$, then let $S(D_0) = \{1, 2, ..., \Pi_{m=1}^{M} |S(D)_m|\}$. By construction, there exists a one-to-one mapping from $(D1, ..., D_M)$ to D_0, so the two models are equivalent.

The matching points can still be found by propensity score matching. Yet it is worth noting that since each D_m may be determined by different mechanisms, they may be affected by different components in $\boldsymbol{X}$. Therefore, the dimension of $\boldsymbol{X}$ needed tends to

be larger than the case of a single endogenous variable. An example illustrating PSC for multiple Ds can be found in Appendix A.3 in the supplement.

A.1.2 Multiple Solutions to $\mathbf{p}(x, z') = \mathbf{p}(x_0, z)$

In this section we adapt the estimator $(\hat{x}_{m1}, \hat{x}_{m2})$ to the general case where the solution to $\boldsymbol{p}(x, z') = \boldsymbol{p}(x_0, z)$ is not unique.

Recall equation (1.4.1) in Section 1.4, we define the estimator to be any $\hat{\boldsymbol{x}}_m \equiv (\hat{x}_{m1}, \hat{x}_{m2})$ that satisfies the following inequality:

$$\hat{Q}_x(\hat{\boldsymbol{x}}_m) \leq \inf_{S_0^2(X)} \hat{Q}_x(\boldsymbol{x}) + a_n$$

Let $\mathcal{X}_m$ be the set of all solutions to $Q(\boldsymbol{x}) = 0$. The estimator is consistent when $a_n = 0$ if $\mathcal{X}_m$ is a singleton. In general, we need $a_n > 0$ to consistently estimate $\mathcal{X}_m$ in Hausdorff distance[1].

Theorem Cons-MP-Set. *Let* $a_n = C\frac{(\log(n))^2}{nh_x}$ *for* $C > 0$. *Under Assumptions Reg-K and Reg-MP,* $\rho_H(\hat{\mathcal{X}}_m, \mathcal{X}_m) = O_p(\frac{\log(n)}{\sqrt{nh_x}})$.

Remark A.1.2. The rate of convergence is slower than that in the case of unique solution $(\frac{1}{\sqrt{nh_x}})$. This is because of the bias introduced by a_n; the convergence of the boundaries of $\hat{\mathcal{X}}_m$ is determined by the rate of a_n, and as a_n converges to o slower than $\hat{Q}_x$, the overall rate is slowed down.

Once $\hat{\mathcal{X}}_m$ is obtained, one can select an element in it as the estimator of a matching point and use it to estimate $\boldsymbol{m}^*(\boldsymbol{x}_0)$ and $\boldsymbol{g}^*(\boldsymbol{x}_0, \cdot)$. However, in order to conduct the overidentification test, $\boldsymbol{x}_m$ has to be locally unique, i.e., an isolated solution, so that the Jacobian of $Q_x(\boldsymbol{x}_m)$ is full rank and the asymptotic distribution in Theorem AsymDist-

[1]The Hausdorff distance ρ_H between two generic subsets A and B of a metric space endowed with metric ρ is defined as $\rho_H(A, B) = \max\{\sup_{a\in A} \inf_{b\in B} \rho(a, b), \sup_{b\in B} \inf_{a\in A} \rho(a, b)\}$. Intuitively, if the Hausdorff distance between two sets are small, for any point in either of the set, there exists a point close to it from the closure of the other set.

MP holds. Therefore, we need to (a) find the isolated $\boldsymbol{x}_m$ and (b) reestimate the isolated matching point to obtain the optimal rate and the asymptotic distribution. We impose the following assumption.

Assumption ISO. *Let $\mathcal{X}_{ISO} \subseteq \mathcal{X}_m$ be the set of all isolated solutions. Suppose the following hold:*

(a) All the isolated solutions are in the interior of $S_0^2(X)$.

(b) There exists $\nu > 0$ such that:

$$\inf_{\boldsymbol{x}' \in \mathcal{X}_m \setminus \mathcal{X}_{ISO}, \boldsymbol{x} \in \mathcal{X}_m} |\boldsymbol{x} - \boldsymbol{x}'| > \nu > 0.$$

(c) The Jacobian Π_x defined in Theorem AsymDist-MP is full rank at every isolated matching point.

Assumption ISO guarantees that each isolated solution is well separated from any other solutions. Under it, we now propose the following post-estimation procedure to estimate the isolated matching points:

- Step 0. Obtain $\hat{\mathcal{X}}_m$ by equation (1.4.1) with $a_n = C\frac{(\log(n))^2}{nh_x}$.
- Step 1 (Isolated solution selection). Cover $\hat{\mathcal{X}}_m$ with squares $\{I_k\}$ of side $b_n = \log(n)\sqrt{a_n}$. Select I_k if it fully covers a cluster of solutions.
- Step 2 (Reestimation). Minimize $\hat{Q}_x(\boldsymbol{x})$ on each selected I_k.

The rationale behind the procedure is as follows. From Theorem Cons-MP-Set and Assumption ISO, $\hat{\mathcal{X}}_m$ consists of isolated clusters of solutions with probability approaching one, and the diameter of a cluster converging in probability to an isolated solutions shrinks to zero at the rate of $\sqrt{a_n}$. Therefore, each of these cluster can be contained in one square I_k with probability approaching one since $b_n > \sqrt{a_n}$. For the reestimation step, again as $b_n/\sqrt{a_n} \to \infty$, by the rate in Theorem Cons-MP-Set, the true matching point is

in the interior of the square with probability approaching 1. Therefore, the minimizer of $\hat{Q}_x$ satisfies the first order condition. By Assumption ISO, Theorem AsymDist-MP thus holds at this estimated matching point.

A.2 Proofs of Results in Sections 1.2 and 1.3

Proof of Theorem MEQ. We first show that $\boldsymbol{p}(\boldsymbol{x}_0, z) = \boldsymbol{p}(\boldsymbol{x}_m, z')$. For any $d \in S(D)$,

$$\begin{aligned}
p_d(\boldsymbol{x}_0, z) &= \int_{h_d(\boldsymbol{x}_0, z, \boldsymbol{v})=1} d\mathbb{P}(\boldsymbol{V} = \boldsymbol{v} | \boldsymbol{X} = \boldsymbol{x}_0, Z = z) \\
&= \int_{h_d(\boldsymbol{x}_0, z, \boldsymbol{v})=1} d\mathbb{P}(\boldsymbol{V} = \boldsymbol{v} | \boldsymbol{X} = \boldsymbol{x}_0) \\
&= \int_{h_d(\boldsymbol{x}_m, z', \boldsymbol{v})=1} d\mathbb{P}(\boldsymbol{V} = \boldsymbol{v} | \boldsymbol{X} = \boldsymbol{x}_m) \\
&= \int_{h_d(\boldsymbol{x}_m, z', \boldsymbol{v})=1} d\mathbb{P}(\boldsymbol{V} = \boldsymbol{v} | \boldsymbol{X} = \boldsymbol{x}_m, Z = z') \\
&= p_d(\boldsymbol{x}_m, z')
\end{aligned}$$

where the second inequality is from Assumption E-SP or E-NSP and the third inequality is from Definition MP

Next we prove equation (1.2.6). For all $d \in S(D)$,

$$\begin{aligned}
\mathbb{E}_{Y|DXZ}(d, \boldsymbol{x}_0, z) - m_d^*(\boldsymbol{x}_0, z) =& \mathbb{E}_{U|DXZ}(d, \boldsymbol{x}_0, z) \\
=& \mathbb{E}_{U|VXZ}(h_d(\boldsymbol{x}_0, z, \boldsymbol{V}) = 1, \boldsymbol{x}_0, z) \\
=& \frac{\int_{h_d(\boldsymbol{x}_0, z, \boldsymbol{v})=1} \mathbb{E}_{U|VXZ}(\boldsymbol{v}, \boldsymbol{x}_0, z) d\mathbb{P}(\boldsymbol{V} = \boldsymbol{v} | \boldsymbol{X} = \boldsymbol{x}_0, Z = z)}{p_d(\boldsymbol{x}_0, z)} \\
=& \frac{\int_{h_d(\boldsymbol{x}_m, z', \boldsymbol{v})=1} \mathbb{E}_{U|VX}(\boldsymbol{v}, \boldsymbol{x}_0) d\mathbb{P}(\boldsymbol{V} = \boldsymbol{v} | \boldsymbol{X} = \boldsymbol{x}_0)}{p_d(\boldsymbol{x}_m, z')} \\
=& \frac{\int_{h_d(\boldsymbol{x}_m, z', \boldsymbol{v})=1} \mathbb{E}_{U|VX}(\boldsymbol{v}, \boldsymbol{x}_m) d\mathbb{P}(\boldsymbol{V} = \boldsymbol{v} | \boldsymbol{X} = \boldsymbol{x}_m)}{p_d(\boldsymbol{x}_m, z')} \\
=& \frac{\int_{h_d(\boldsymbol{x}_m, z', \boldsymbol{v})=1} \mathbb{E}_{U|VXZ}(\boldsymbol{v}, \boldsymbol{x}_m, z') d\mathbb{P}(\boldsymbol{V} = \boldsymbol{v} | \boldsymbol{X} = \boldsymbol{x}_m, z')}{p_d(\boldsymbol{x}_m, z')} \\
=& \mathbb{E}_{U|VXZ}(h_d(\boldsymbol{x}_m, z', \boldsymbol{V}) = 1, \boldsymbol{x}_m, z') \\
=& \mathbb{E}_{U|DXZ}(d, \boldsymbol{x}_m, z') \\
=& \mathbb{E}_{Y|DXZ}(d, \boldsymbol{x}_m, z') - m_d^*(\boldsymbol{x}_m, z')
\end{aligned}$$

where the fourth inequality follows Assumption E-SP. The fifth inequality is from Definition MP. For the sixth equality, Definition MP implies the exogeneity assumption also holds at $\boldsymbol{x}_m$.

Finally we show equation (1.2.7). For all $d \in S(D)$,

$$\begin{aligned}
F_{Y|DXZ}(g_d^*(\boldsymbol{x}_0, u) | d, \boldsymbol{x}_0, z) =& F_{U|DXZ}(u | d, \boldsymbol{x}_0, z) \\
=& \frac{\int_{h_d(\boldsymbol{x}_0, z, \boldsymbol{v})=1} F_{U|VXZ}(u | \boldsymbol{v}, \boldsymbol{x}_0, z) d\mathbb{P}(\boldsymbol{V} = \boldsymbol{v} | \boldsymbol{X} = \boldsymbol{x}_0, Z = z)}{p_d(\boldsymbol{x}_0, z)} \\
=& \frac{\int_{h_d(\boldsymbol{x}_m, z', \boldsymbol{v})=1} F_{U|VXZ}(u | \boldsymbol{v}, \boldsymbol{x}_m, z') d\mathbb{P}(\boldsymbol{V} = \boldsymbol{v} | \boldsymbol{X} = \boldsymbol{x}_m, Z = z')}{p_d(\boldsymbol{x}_m, z')} \\
=& F_{U|VXZ}(u | h_d(\boldsymbol{x}_m, z', \boldsymbol{V}) = 1, \boldsymbol{x}_m, z') \\
=& F_{U|DXZ}(u | d, \boldsymbol{x}_m, z') \\
=& F_{Y|DXZ}(g_d^*(\boldsymbol{x}_m, u) | d, \boldsymbol{x}_m, z')
\end{aligned}$$

where the first equality follows from Assumptions FS and CM. The third and the fourth equalities are from Definitions MP and Assumption E-NSP. For the third and the last equality, by Definition of MP, the exogeneity and the full support assumptions also hold at $\boldsymbol{x}_m$. □

Proof of Theorem ID-OC. Under the model setup, inequality (1.3.3) holds if

$$\begin{aligned}&\big(F_{V_1}(x_0\beta) - F_{V_1}(x_0\beta - \alpha)\big)\big(F_{V_2}(x_0\beta) - F_{V_2}(x_0\beta + \alpha)\big)\\ \neq&\big(F_{V_1}(x_0\beta) - F_{V_1}(x_0\beta + \alpha)\big)\big(F_{V_2}(x_0\beta) - F_{V_2}(x_0\beta - \alpha)\big)\end{aligned} \tag{A.2.1}$$

Without loss of generality, assume $\alpha > 0$. Then there are four cases.

- $x_0\beta - \alpha < x_0\beta < x_0\beta + \alpha \leq c$. Then $F_{V_2}(x_0\beta - \alpha) = F_{V_2}(x_0\beta) = F_{V_2}(x_0\beta + \alpha) = 0$. Inequality (A.2.1) does not hold.
- $x_0\beta - \alpha < x_0\beta \leq c < x_0\beta + \alpha$. The left hand side is negative but the right hand side is zero because $F_{V_2}(x_0\beta - \alpha) = F_{V_2}(x_0\beta) = 0$. Inequality (A.2.1) holds.
- $x_0\beta - \alpha < c < x_0\beta < x_0\beta + \alpha$. The left hand side is still negative but the right hand side is again zero because $F_{V_1}(x_0\beta) = F_{V_1}(x_0\beta + \alpha) = 1$.
- $c \leq x_0\beta - \alpha < x_0\beta < x_0\beta + \alpha$. Both sides are zero because $F_{V_1}(x_0\beta) = F_{V_1}(x_0\beta - \alpha) = F_{V_1}(x_0\beta + \alpha) = 1$.

Therefore, inequality (A.2.1) holds if there are two elements from $\{(x_0, 0), (x_0, 1), (x_m, 0)\}$ making the single indices located left and right to c respectively. □

Proof of Lemma UNQ. For each l, denote the lower and the upper bounds of S_l by $\underline{y}_l$ and $\bar{y}_l$. By monotonicity, for each component y_l^*, $y_l^*(0) = \underline{y}_l$ and $y_l^*(1) = \bar{y}_l$.

Suppose $\boldsymbol{y}^*$ is not the unique solution path in $\mathcal{Y}^*$, then any other solution path $\tilde{\boldsymbol{y}} \equiv$

$(\tilde{y}_1, ..., \tilde{y}_l, ..., \tilde{y}_L)$ must also satisfy

$$\tilde{y}_l(0) = \underline{y}_l \text{ and } \tilde{y}_l(1) = \bar{y}_l, \forall l.$$

Therefore, the set $\{u' : \tilde{\boldsymbol{y}}(u) = \boldsymbol{y}^*(u), 0 \leq u \leq u'\}$ is nonempty and its supremum, denoted by $\bar{u}$, is well-defined. Then there are the following two cases depending on whether $\tilde{\boldsymbol{y}}$ is continuous at $\bar{u}$.

Case 1. $\tilde{\boldsymbol{y}}(\cdot)$ is continuous at $\bar{u}$. By continuity of $\boldsymbol{y}^*$ and $\tilde{\boldsymbol{y}}$, $\bar{u} \in \{u' : \tilde{\boldsymbol{y}}(u) = \boldsymbol{y}^*(u), 0 \leq u \leq u'\}$.

If $\bar{u} = 1$, we are done.

If $\bar{u} < 1$, by monotonicity, $\tilde{\boldsymbol{y}}(u)$ has at most countable discontinuities. Thus, there exists an interval $(\bar{u}, \bar{u}')$ where $\bar{u}' < 1$ such that on the interval, $\tilde{\boldsymbol{y}}(u)$ is continuous and $\tilde{\boldsymbol{y}}(u) \neq \boldsymbol{y}^*(u)$ for $u \in (\bar{u}, \bar{u}')$.

Since the Jacobian of $\nabla \boldsymbol{M}(\boldsymbol{y}^*(u), u)$ is full-rank and continuous in $(\boldsymbol{y}^*(u), u)$, there exists a neighborhood of $(\boldsymbol{y}^*(u), u)$ on which $\boldsymbol{M}(\cdot, \cdot)$ is one-to-one. Meanwhile, by continuity of $\tilde{\boldsymbol{y}}$ and $\boldsymbol{y}^*$, there exists u'' and $\tilde{\boldsymbol{y}}(u'')$ in that neighborhood. Then $\tilde{\boldsymbol{y}}(u'') \neq \boldsymbol{y}^*(\boldsymbol{x}_0, u'')$ but $\boldsymbol{M}(\tilde{\boldsymbol{u}}(u''), u'') = \boldsymbol{M}(\boldsymbol{y}^*(u''), u'')$, a contradiction. A similar argument can also be found in Ortega and Rheinboldt (1970), pp. 133-134, Ambrosetti and Prodi (1995), pp. 48-49, and De Marco, Gorni and Zampieri (2014), as an intermediate step to show variants of the Hadamard Theorem.

Case 2. $\tilde{\boldsymbol{y}}(\cdot)$ is not continuous at $\bar{u}$. Since any solution at $u = 1$ is equal to $\boldsymbol{y}^*(1)$, we only consider the case where $\bar{u} < 1$.

By monotonicity, there is at least one l such that $\lim_{u \searrow \bar{u}} \tilde{y}_l(u) > \lim_{u \nearrow \bar{u}} \tilde{y}_l(u)$, i.e., $\tilde{y}_l(\cdot)$ jumps up at $\bar{u}$. However, as $\boldsymbol{M}(\cdot, u)$ is continuous and strictly increasing, to make $\lim_{u \searrow \bar{u}} \boldsymbol{M}(\tilde{\boldsymbol{y}}(u), u) = \lim_{u \nearrow \bar{u}} \boldsymbol{M}(\tilde{\boldsymbol{y}}(u), u) = \boldsymbol{0}$, there must exist $l' \neq l$ such that $\lim_{u \searrow \bar{u}} \tilde{y}_{l'}(u) < \lim_{u \nearrow \bar{u}} \tilde{y}_{l'}(u)$. A contradiction with that $\tilde{y}_{l'}(\cdot)$ is increasing.

Therefore, $\boldsymbol{y}^*$ is the unique solution path in $\mathcal{Y}$ to $\boldsymbol{M}(\cdot, u) = \boldsymbol{0}$. □

Proof of Theorem ID-NSP. Let $\mathcal{G}^* \subset \mathcal{G}$ contain functions whose ranges are contained in $\prod_{d=1}^{3} S(Y|d, \boldsymbol{x}_0)$. Under Assumptions CM and Assumption FS, conditions in Lemma UNQ are satisifed. Hence $g^*(\boldsymbol{x}_0, \cdot)$ is the unique solution path in $\mathcal{G}^*$.

Now we consider the uniqueness in $\mathcal{G}$. Suppose there exists another solution path $\breve{g} : [0,1] \in \mathcal{G}$. Construct $g^\dagger$ as follows: For each $d \in S(D)$, let

$$g_d^\dagger(u) = \begin{cases} \underline{y}_{d\boldsymbol{x}_0}, & \text{if } \breve{g}_d(u) < \underline{y}_{d\boldsymbol{x}_0} \\ \breve{g}_d(u), & \text{if } \breve{g}_d(u) \in S(Y|d, \boldsymbol{x}_0) \\ \bar{y}_{d\boldsymbol{x}_0}, & \text{if } \breve{g}_d(u) > \bar{y}_{d\boldsymbol{x}_0} \end{cases}$$

Clearly, $g^\dagger \in \mathcal{G}^*$. By the uniqueness of $g^*(\boldsymbol{x}_0, \cdot)$ in $\mathcal{G}^*$, $g^\dagger(\cdot) = g^*(\boldsymbol{x}_0, \cdot)$, so $\breve{g}(u) = g^*(\boldsymbol{x}_0, u)$ for all $u \in (0,1)$. For $u = 0, 1$, $\breve{g}_d(u)$ can take any value smaller than $\underline{y}_{d\boldsymbol{x}_0}$ or greater than $\bar{y}_{d\boldsymbol{x}_0}$ respectively for all d. □

A.3 Examples for Propensity Score Coherence

Example OC Cont'd 3. *In this example we generalize the ordered choice model in Example OC Cont'd 2.*

Let $h_1(\boldsymbol{X}, Z, V) = \mathbb{1}(V_1 \leq \gamma_1(\boldsymbol{X}, Z))$, $h_3(\boldsymbol{X}, Z, V) = \mathbb{1}(V_2 > \gamma_2(\boldsymbol{X}, Z))$, *and* $h_2 = 1 - h_1 - h_3$ *where* $V_1 < V_2$ *a.s. are two scalar random variables that are continuously distributed on* $\mathbb{R}$. *Also, assume* $\gamma_1(\boldsymbol{X}, Z) < \gamma_2(\boldsymbol{X}, Z)$ *a.s. This model nests parametric ordered choice models and also some nonparametric models with more complicated structures, for instance the general ordered choice model in Cunha, Heckman and Navarro (2007).*

The matching points are identified by propensity score matching (1.3.1): By strict monotonicity of F_{V_1} *and* F_{V_2}, *it is straightforward that for any* $(\boldsymbol{x}, z)$ *and* $(\boldsymbol{x}', z')$, $p_1(\boldsymbol{x}, z) = p_1(\boldsymbol{x}', z')$ *implies* $\gamma_1(\boldsymbol{x}, z) = \gamma_1(\boldsymbol{x}', z')$, *and* $p_3(\boldsymbol{x}, z) = p_3(\boldsymbol{x}', z')$ *implies* $\gamma_2(\boldsymbol{x}, z) = \gamma_2(\boldsymbol{x}', z')$.

Example MC (Multinomial Choice)**.** *In this model we consider a nonparametric multinomial choice model. Variants of it are considered in Matzkin (1993), Heckman, Urzua and Vyt-*

lacil (2008), and Lee and Salanié (2018). Let $R_d(\boldsymbol{X}, Z) + \tilde{V}_d$ be the indirect utility of choosing treatment d where $\tilde{V}_d$ is an unobservable continuous random variable. Alternative d is selected if $R_d(\boldsymbol{X}, Z) + \tilde{V}_d > R_{-d}(\boldsymbol{X}, Z) + \tilde{V}_{-d}$ where the subscript $-d$ refers to any selection other than d. Reparameterize the model by letting $V_1 = \tilde{V}_2 - \tilde{V}_1$, $V_2 = \tilde{V}_3 - \tilde{V}_1$, $V_3 = \tilde{V}_3 - \tilde{V}_2$, $\gamma_1(\boldsymbol{X}, Z) = R_1(\boldsymbol{X}, Z) - R_2(\boldsymbol{X}, Z)$ and $\gamma_2(\boldsymbol{X}, Z) = R_1(\boldsymbol{X}, Z) - R_3(\boldsymbol{X}, Z)$. The model can be rewritten as

$$
\begin{aligned}
D = 1 &\iff V_1 < \gamma_1(\boldsymbol{X}, Z), V_2 < \gamma_2(\boldsymbol{X}, Z) \\
D = 2 &\iff V_1 > \gamma_1(\boldsymbol{X}, Z), V_3 < \gamma_2(\boldsymbol{X}, Z) - \gamma_1(\boldsymbol{X}, Z) \\
D = 3 &\iff V_2 > \gamma_2(\boldsymbol{X}, Z), V_3 < \gamma_2(\boldsymbol{X}, Z) - \gamma_1(\boldsymbol{X}, Z)
\end{aligned}
$$

If $0 < p_d(\boldsymbol{x}_0, z) < 1$ for all d, it can be verified that the solutions to equation (1.3.1) are matching points of $\boldsymbol{x}_0$.

Example Two Endogenous Variables. *In this example, we have two dummy endogenous variables. We show that PSC still holds in this model. Let $D_1, D_2 \in \{0, 1\}$ be two endogenous variables. Suppose they are determined by the following model:*

$$
\begin{aligned}
D_1 &= \mathbb{1}(\gamma_1(X, Z) \leq V_1) \\
D_2 &= \mathbb{1}(\gamma_2(X, Z) \leq V_2)
\end{aligned}
$$

where V_1 and V_2 are unobservables continuously distributed on $\mathbb{R}^2$. Let $D_0 = 1, 2, 3, 4$, corresponding to $(D_1, D_2) = (0, 0), (0, 1), (1, 0), (1, 1)$ respectively. Rewrite the model as

$$
\begin{aligned}
h_1(X, Z, V_1, V_2) &= \big(1 - \mathbb{1}(\gamma_1(X, Z) \leq V_1)\big) \cdot \big(1 - \mathbb{1}(\gamma_2(X, Z) \leq V_2)\big) \\
h_2(X, Z, V_1, V_2) &= \big(1 - \mathbb{1}(\gamma_1(X, Z) \leq V_1)\big) \cdot \mathbb{1}(\gamma_2(X, Z) \leq V_2) \\
h_3(X, Z, V_1, V_2) &= \mathbb{1}(\gamma_1(X, Z) \leq V_1) \cdot \big(1 - \mathbb{1}(\gamma_2(X, Z) \leq V_2)\big) \\
h_4(X, Z, V_1, V_2) &= \mathbb{1}(\gamma_1(X, Z) \leq V_1) \cdot \mathbb{1}(\gamma_2(X, Z) \leq V_2)
\end{aligned}
$$

It is clear that $\sum_d h_d(X,Z) = 1$. *Let us now verify that PSC holds for this model. Suppose* $\boldsymbol{p}(x_0,z) = \boldsymbol{p}(x_m,z')$. *Then*

$$\mathbb{P}\big(V_1 < \gamma_1(x_0,z), V_2 < \gamma_2(x_0,z)\big) = \mathbb{P}\big(V_1 < \gamma_1(x_m,z'), V_2 < \gamma_2(x_m,z')\big)$$

Suppose $\gamma_1(x_0,z) \neq \gamma_1(x_m,z')$. *Without loss of generality, let* $\gamma_1(x_0,z) < \gamma_1(x_m,z')$. *Then* $\gamma_2(x_0,z) > \gamma_2(x_m,z')$. *Consequently,*

$$\mathbb{P}\big(V_1 \geq \gamma_1(x_0,z), V_2 < \gamma_2(x_0,z)\big) > \mathbb{P}\big(V_1 \geq \gamma_1(x_m,z'), V_2 < \gamma_2(x_m,z')\big)$$

But this implies $p_3(x_0,z) > p_3(x_m,z')$, *a contradiction.*

A.4 Additional Simulation Results

In this section, we present the simulation results under different parameters in the model in Section 1.7.1. Table A.1 and Table A.2 present the results for $x_0 = \pm 0.3$ with parameters the same as in Table 1.6. We can see the results are similar to those in Table 1.6. Table A.3 shows the results for weaker IV: $(\alpha,\beta) = (0.16, 0.08)$. These parameters are one fifth of the benchmark one, while the smallest eigenvalue of the matrix $\Pi_{SP}'\Pi_{SP}$ is $1/200$ of the original. The variances blow up, but the biases are still very small like in the benchmark case. Table A.4 shows the results for different correlation coefficients $\rho = 0.3, 0.7$. The results are again very similar to the benchmark case $\rho = 0.5$.

Table A.1: $x_0 = -0.3, \boldsymbol{m}^*(-0.3) = (1.05, 2.1, 2.45)$

	N	Average	Bias^2	Variance	MSE	90%	95%	99%
$\hat{m}_1(-0.3)$	1000	1.02	0.001	0.14	0.14	93% (87.2%)	97% (92%)	99.2% (98%)
	2000	1.04	2e-4	0.07	0.07	91% (88.6%)	96.4% (94.8%)	99.2% (98.6%)
	3000	1.04	1e-4	0.05	0.05	89.8% (88.6%)	94.4% (93.2%)	99% (98.8%)
$\hat{m}_2(-0.3)$	1000	2.09	1e-4	0.85	0.85	93.8% (88%)	96.8% (94.8%)	99% (98.2%)
	2000	2.04	0.003	0.34	0.35	92.8% (90.2%)	96.8% (95.4%)	99.6% (99.4%)
	3000	2.04	0.004	0.25	0.26	90% (88.2%)	95.8% (93.2%)	99.4% (97.8%)
$\hat{m}_3(-0.3)$	1000	2.36	0.01	0.21	0.22	91.4% (88.6%)	96% (93.4%)	98.4% (97.4%)
	2000	2.41	0.002	0.10	0.10	93.2% (90.6%)	96.6% (95.4%)	98.6% (99.4%)
	3000	2.43	3e-4	0.07	0.07	89.2% (88.4%)	94.4% (94.4%)	99.2% (98.8%)
$\mathcal{J}_x$	1000					88.6%	94.6%	99%
	2000					90.4%	95.2%	99.2%
	3000					91.2%	97%	99.4%
$\mathcal{J}_{SP}$	1000					93.4%	96.6%	99.2%
	2000					92.2%	96.4%	99.2%
	3000					91%	94.2%	99.2%

Table A.2: $x_0 = 0.3, \boldsymbol{m}^*(0.3) = (1.95, 3.9, 4.55)$

	N	Average	Bias^2	Variance	MSE	90%	95%	99%
$\hat{m}_1(0.3)$	1000	1.95	2e-5	0.11	0.11	89.4% (85%)	93.4% (90.4%)	97.2% (96.2%)
	2000	1.98	8e-4	0.05	0.05	90.2% (89.8%)	95% (94.2%)	98.6% (98.2%)
	3000	1.95	2e-6	0.04	0.04	89% (88.8%)	94.2% (92.8%)	99% (98.6%)
$\hat{m}_2(0.3)$	1000	3.73	0.03	0.87	0.90	90.2% (85%)	93.8% (90.2%)	97% (95.6%)
	2000	3.72	0.03	0.38	0.41	92% (86.8%)	95.2% (94%)	99% (98.8%)
	3000	3.81	0.01	0.27	0.28	89.6% (89.4%)	95.6% (95.6%)	99.6% (98.4%)
$\hat{m}_3(0.3)$	1000	4.55	1e-5	0.27	0.27	92.4% (86.2%)	95.4% (91%)	97.6% (95.4%)
	2000	4.56	2e-4	0.13	0.13	93.8% (89%)	97.4% (94.2%)	99.2% (97.8%)
	3000	4.53	3e-4	0.08	0.08	93% (90.6%)	97.8% (95.6%)	99.2% (98.8%)
$\mathcal{J}_x$	1000					88.6%	94.6%	99%
	2000					88.8%	95.2%	99.2%
	3000					92%	95.6%	99%
$\mathcal{J}_{SP}$	1000					90.2%	94.2%	97.4%
	2000					91.4%	96.8%	99%
	3000					90.8%	96.2%	99.2%

Table A.3: Different Strengths of (Z, X)

(α, β)	min. eig.		Average	Bias^2	Variance	MSE	90%	95%	99%
(0.8,0.4)	0.02	$\hat{m}_1(0)$	1.51	3e-5	0.06	0.06	91.6% (88.4%)	96% (94%)	99% (99%)
		$\hat{m}_2(0)$	2.88	0.01	0.37	0.39	89.6% (88.2%)	95% (93.6%)	99% (98.4%)
		$\hat{m}_3(0)$	3.49	2e-4	0.12	0.12	92.2% (90.4%)	97.2% (95.8%)	98.8% (98.8%)
(0.16,0.08)	1e-4	$\hat{m}_1(0)$	1.48	3e-4	13.10	13.10	93% (90.6%)	96.4% (95.6%)	99.2% (99%)
		$\hat{m}_2(0)$	3.11	0.01	50.82	50.83	93.2% (89.2%)	95.6% (94.6%)	99.4% (98%)
		$\hat{m}_3(0)$	3.44	0.004	11.04	11.05	91.8% (90.6%)	95.2% (96%)	99.2% (99%)

Table A.4: Different Degree of Endogeneity

	ρ	Average	Bias^2	Variance	MSE	90%	95%	99%
$\hat{m}_1(0)$	0.3	1.50	2e-5	0.06	0.06	92.4% (89.8%)	96.8% (95.4%)	99.6% (99%)
	0.5	1.51	3e-5	0.06	0.06	91.6% (88.4%)	96% (94%)	99% (99%)
	0.7	1.51	4e-5	0.05	0.05	91.2% (89.2%)	95.4% (93.8%)	99.6% (97.6%)
$\hat{m}_2(0)$	0.3	2.88	0.01	0.38	0.39	91.8% (86.4%)	96.4% (93.4%)	99.4% (97.4%)
	0.5	2.88	0.01	0.37	0.39	89.6% (88.2%)	95% (93.6%)	99% (98.4%)
	0.7	2.88	0.01	0.35	0.37	91% (87.6%)	96.2% (93.6%)	99% (98%)
$\hat{m}_3(0)$	0.3	3.49	5e-5	0.12	0.12	93% (90.8%)	96.8% (95.2%)	99.2% (98.6%)
	0.5	3.49	2e-4	0.12	0.12	92.2% (90.4%)	97.2% (95.8%)	98.8% (98.8%)
	0.7	3.49	2e-4	0.11	0.11	92% (89.8%)	96.8% (95.6%)	99.4% (98.8%)

A.5 Proofs of Results in Section 1.6

In Section A.5.1 we prove some of the asymptotic results in Section 1.6. As the results of the matching point (the unique solution case) and of the separable model are standard, we only include proofs for the multiple matching points case (Theorem Cons-MP-Set) and the results of the nonseparable model estimator. Section A.5.2 contains used in

Section A.5.1.

A.5.1 Asymptotic Properties

Let us first introduce the following lemmas. It will be needed in deriving the asymptotic properties for all the estimators.

Lemma A.5.1. *Suppose* $h_0/h_g \to 0$. *Let* $S_0(Y|d,x)$ *be an interior set of* $S(Y|d,x)$. *Under Assumptions Reg-K, Reg-L, Reg-MP, Reg-SP, and Reg-NSP,*

$$\sup_{x\in S_0(X)} \left|\hat{p}_d(x,z) - p_d(x,z))\right| = O_p(\sqrt{\frac{\log(n)}{nh_x}} + h_x^2) \tag{A.5.1}$$

$$\sup_{x\in S_0(X)} \left|\hat{\mathbb{E}}_{Y|DXZ}(d,x,z) - \mathbb{E}_{Y|XZ}(d,x,z)\right| = O_p(\sqrt{\frac{\log(n)}{nh_m}} + h_m^2) \tag{A.5.2}$$

$$\sup_{\substack{y\in S(Y|d)\\ x\in S_0(X)}} \left|\hat{F}_{Y|DXZ}(y|d,x,z) - F_{Y|DXZ}(y|d,x,z)\right| = O_p(\sqrt{\frac{\log(n)}{nh_g}} + h_g) \tag{A.5.3}$$

$$\sup_{x\in S_0(X)} \left|\partial_x\hat{p}_d(x,z) - \partial_x p_d(x,z)\right| = o_p(1) \tag{A.5.4}$$

$$\sup_{x\in S_0(X)} \left|\partial_x\hat{\mathbb{E}}_{Y|DXZ}(d,x,z) - \partial_x\mathbb{E}_{Y|XZ}(d,x,z)\right| = o_p(1) \tag{A.5.5}$$

$$\sup_{\substack{y\in S_0(Y|d,x)\\ x\in S_0(X)}} \left|\partial_y\hat{F}_{Y|DXZ}(y|d,x,z) - f_{Y|DXZ}(y|d,x,z)\right| = o_p(1) \tag{A.5.6}$$

$$\sup_{\substack{y\in S_0(Y|d,x)\\ x\in S_0(X)}} \left|\partial_x\hat{F}_{Y|DXZ}(y|d,x,z) - \partial_x F_{Y|DXZ}(y|d,x,z)\right| = o_p(1) \tag{A.5.7}$$

The first three results are needed for consistency and the rate of convergence for each estimator proposed in the paper. The last four are needed to derive the rate of convergence as well and the asymptotic distributions. Tthese results are standard except equation (A.5.3). The bias term is $O_p(h_g)$ instead of the standard $O(h_g^2)$. This is due to the nonsmoothness of $F_{Y|DXZ}(\cdot|d,x,z)$ at the boundaries. We prove it in Appendix A.5.2.

The Matching Points

Proof of Theorem Cons-MP-Set. As the estimator is identical to Chernozhukov, Hong and Tamer (2007), we prove the theorem by verifying the conditions needed for their Theorem 3.1. Specifically, let $\tilde{Q}_x(\boldsymbol{x}) \equiv \hat{Q}_x(\boldsymbol{x}) - \inf_{\boldsymbol{x}\in S_0^2(X)} \hat{Q}_x(\boldsymbol{x})$. We need to verify that (a) $\sup_{\boldsymbol{x}\in S_0^2(X)} |\tilde{Q}_x(\boldsymbol{x}) - Q_x(\boldsymbol{x})| = O_p(\frac{\log(n)}{nh_x})$ and (b) there exist positive (δ, κ) such that for any $\varepsilon \in (0,1)$ there are $(\kappa_\varepsilon, n_\varepsilon)$ such that for all $n > n_\varepsilon$, $\tilde{Q}_x(\boldsymbol{x}) \geq \kappa[\rho(\boldsymbol{x}, \mathcal{X}_m) \wedge \delta]^2$ uniformly on $\Delta \equiv \{\boldsymbol{x} \in S_0^2(X) : \rho(\boldsymbol{x}, \mathcal{X}_m) \geq \sqrt{\frac{\kappa_\varepsilon \log(n)}{nh_x}}\}$ with probability at least $1-\varepsilon$.

We first derive the uniform rate of convergence for $\tilde{Q}_x(\boldsymbol{x})$. Let $\boldsymbol{x}_{mn}$ be such that $\hat{Q}_x(\boldsymbol{x}_{mn}) = \inf_{S_0^2(X)} \hat{Q}_x(\boldsymbol{x})$. Then

$$\begin{aligned}
\sup_{\boldsymbol{x}\in S_0^2(X)} \left|\hat{Q}_x(\boldsymbol{x}) - \hat{Q}_x(\boldsymbol{x}_{mn}) - Q_x(\boldsymbol{x})\right| \leq & \sup_{\boldsymbol{x}\in S_0^2(X)} \left|\hat{Q}_x(\boldsymbol{x}) - Q_x(\boldsymbol{x})\right| + \left|\hat{Q}_x(\boldsymbol{x}_{mn}) - Q_x(\boldsymbol{x}_m)\right| \\
\leq & \sup_{\boldsymbol{x}\in S_0^2(X)} \left|\hat{Q}_x(\boldsymbol{x}) - Q_x(\boldsymbol{x})\right| + \hat{Q}_x(\boldsymbol{x}_m) - Q_x(\boldsymbol{x}_m) \\
\leq & 2 \sup_{\boldsymbol{x}\in S_0^2(X)} \left|\hat{Q}_x(\boldsymbol{x}) - Q_x(\boldsymbol{x})\right| \\
= & O_p(\frac{\log(n)}{nh_x})
\end{aligned}$$

where we used the definition of $\boldsymbol{x}_{mn}$, nonnegativity of $\hat{Q}_x$ and $Q_x(\boldsymbol{x}_m) = 0$. The last inequality is a consequence of Lemma A.5.1 and the choice of h_m.

For (b), note that by Assumption Reg-MP, there exists $C > 0$ such that $Q_x(\boldsymbol{x}) \geq C\kappa[\rho(\boldsymbol{x}, \mathcal{X}_m) \wedge \delta]^2$ uniformly on Δ by continuity of Q_x and compactness of $S_0^2(\boldsymbol{X}) \cap \Delta$.

$$\begin{aligned}
\inf_{\boldsymbol{x}\in\Delta} \tilde{Q}_x(\boldsymbol{x}) = & \inf_{\boldsymbol{x}\in\Delta} |\tilde{Q}_x(\boldsymbol{x}) - Q_x(\boldsymbol{x}) + Q_x(\boldsymbol{x})| \\
\geq & \inf_{\boldsymbol{x}\in\Delta} |Q_x(\boldsymbol{x})| - \sup_{\boldsymbol{x}\in\Delta} |\tilde{Q}_x(\boldsymbol{x}) - Q_x(\boldsymbol{x})| \\
\geq & C\kappa[\rho(\boldsymbol{x}, \mathcal{X}_m) \wedge \delta]^2 - O_p(\frac{\log(n)}{nh_x})
\end{aligned}$$

Therefore, we can choose $(\kappa_\varepsilon, n_\varepsilon)$ large enough so that the desired inequality holds uni-

formly on Δ. □

The Nonseparable Model-NSP

Proof of Theorem ID-Sup. Denote $\mathcal{G}_0^- = \mathcal{G}_0 \backslash \{g : \sup_{u \in \mathcal{U}_0} |g(u) - g^*(x_0, u)| \geq \delta\}$. Suppose inequality (1.6.7) does not hold. Then there exists a sequence $g_k \in \mathcal{G}_0^-$ such that

$$\lim_{k \to \infty} \Big(\int_0^1 Q_{NSP}(g_k(u), u) du - \int_0^1 Q_{NSP}(g^*(x_0, u), u) du \Big) = 0$$

As the sequence $\{g_k\}$ are uniformly bounded monotonic functions on a compact interval, by Helly's Selection Theorem there exists a pointwise convergent subsequence $\tilde{g}_{k_l}$. Denote its limit by $\tilde{g}$. Note the equation above also holds for this subsequence. Then by the Dominated Convergence Theorem, we can change the order of the limit and the integral operators:

$$\begin{aligned} &\int_0^1 \lim_{k_l \to \infty} Q_{NSP}(g_{k_l}(u), u) du = \int_0^1 Q_{NSP}(g^*(x_0, u), u) du = 0 \\ \Longrightarrow \quad &\int_0^1 Q_{NSP}(\tilde{g}(u), u) du = \int_0^1 Q_{NSP}(g^*(x_0, u), u) du = 0 \end{aligned}$$

where the last equation follows from continuity of Q_{NSP}.

As $\mathcal{G}_0$ is closed, $\tilde{g} \in \mathcal{G}_0$. By Theorem ID-NSP, $\tilde{g}(\cdot) = g^*(x_0, \cdot)$ on $\mathcal{U}_0$. Since pointwise convergence of a sequence of monotonic functions on a compact domain implies uniform convergence if the limiting function is continuous, by continuity of $g^*(x_0, u)$,

$$\lim_{k_l \to \infty} \sup_{u \in \mathcal{U}_0} |g^*(x_0, u) - g_{k_l}(u)| \to 0,$$

contradicting $g_{k_l} \in \mathcal{G}_0^-$. □

From the last step in the proof, $\mathcal{U}_0$ in the theorem can be replaced by $[0, 1]$ if uniqueness of $g^*(x_0, \cdot)$ holds on the entire interval $[0, 1]$. As discussed in Section 1.3.3, this

would be the case when the codomain in $\mathcal{G}_0$ was set to be $\prod_d S(Y|d, x_0)$ instead of $\prod_d S(Y|d)$.

Next, to prove Theorem Cons-NSP, we need the following lemmas. Let x_m be a generic matching point.

Lemma A.5.2. *Let $r_n = O_p(\sqrt{\frac{\log(n)}{nh_g}} + h_g)$. Suppose $|\hat{x}_m - x_m| = O_p(a_n)$. Under the conditions in Theorem Cons-NSP and Lemma A.5.1, we have the following for all $d \in S(D)$:*

$$\sup_{y \in [\underline{y}_d, \bar{y}_d]} \left| \hat{F}_{Y|DXZ}(\hat{\varphi}_d(y; \hat{x}_m, z')|d, \hat{x}_m, z) - F_{Y|DXZ}(\varphi_d(y; x_m, z')|d, x_m, z) \right| = O_p(r_n + a_n)$$

The main challenge to show this lemma is that the supremum is taken on $S(Y|d) \equiv [\underline{y}_d, \bar{y}_d]$, which contains the support set $S(Y|d, x_0)$ and $S(Y|d, x_m)$ as two subsets. By definition, $\varphi_d(y; x_m, z')$ is not unique when y is outside $S(Y|d, x_0)$, so it is not valid to show uniform consistent of $\hat{\varphi}_d$. The key lies in the fact that when $\varphi_d(y; x_m, z')$ is not unique, $F_{Y|DXZ}(\varphi_d(y; x_m, z')|d, x_m, z)$ is still unique (equal to 0 or 1). So instead of showing it by establishing uniform convergence for $\hat{F}_{Y|DXZ}$ and $\hat{\varphi}_d$ separately, we treat it as one object.

Under Lemma A.5.2, it is straightforward that $\sup_{\boldsymbol{y} \in \prod_d S(Y|d)} |\hat{Q}_{NSP}(\boldsymbol{y}, u) - Q_{NSP}(\boldsymbol{y}, u)| = o_p(1)$. Then we have the following lemmas.

Lemma A.5.3. *Under all the conditions in Lemma A.5.2, the following holds:*

$$\sup_{g \in \hat{\mathcal{G}}} \left| \frac{1}{J} \sum_{j=1}^{J} \hat{Q}_{NSP}(\boldsymbol{g}(u_j), u_j) - \int_0^1 Q_{NSP}(\boldsymbol{g}(u), u) du \right| = o_p(1)$$

Lemma A.5.4. *There exists $\boldsymbol{g}_0 \in \hat{\mathcal{G}}$ such that $|\boldsymbol{g}_0(u) - \boldsymbol{g}^*(x_0, u)| = o(1)$ for all $u \in (0, 1)$.*

Proof of Theorem Cons-NSP. Suppose there exists $\delta > 0$ such that

$$\sup_{u \in \mathcal{U}_0} |\hat{\boldsymbol{g}}(x_0, u) - \boldsymbol{g}^*(x_0, u)| > \delta.$$

By construction, $\hat{g} \in \mathcal{G}_0$. By Theorem ID-Sup, there exists $\varepsilon > 0$ such that

$$\left(\int_0^1 Q_{NSP}(\hat{g}(x_0,u),u)du - \int_0^1 Q_{NSP}(g^*(x_0,u),u)du \right) > \varepsilon$$

For simplicity, denote the sample objective function by $\hat{\mathcal{L}}$. By Lemma A.5.3 and the rate of λ, $\sup_{g\in\hat{\mathcal{G}}} |\hat{\mathcal{L}}(g(u)) - \int_0^1 Q_{NSP}(g(u),u)du| = o_p(1)$. Then

$$\begin{aligned}
&\left(\int_0^1 Q_{NSP}(\hat{g}(x_0,u),u)du - \int_0^1 Q_{NSP}(g^*(x_0,u),u)du \right) \\
\leq &\hat{\mathcal{L}}(\hat{g}(x_0,u)) - \int_0^1 Q_{NSP}(g^*(x_0,u),u)du \\
&+ \int_0^1 Q_{NSP}(\hat{g}(x_0,u),u)du - \hat{\mathcal{L}}(\hat{g}(x_0,u)) \\
\leq &\hat{\mathcal{L}}(g_0(u)) - \int_0^1 Q_{NSP}(g_0(u),u)du \\
&+ \int_0^1 Q_{NSP}(g_0(u),u)du - \int_0^1 Q_{NSP}(g^*(x_0,u),u)du \\
&+ \sup_{g\in\hat{\mathcal{G}}} |\hat{\mathcal{L}}(g(u)) - \int_0^1 Q_{NSP}(g(u),u)du| \\
\leq &2 \sup_{g\in\hat{\mathcal{G}}} |\hat{\mathcal{L}}(g(u)) - \int_0^1 Q_{NSP}(g(u),u)du| + \int_0^1 (Q_{NSP}(g_0(u),u) - Q_{NSP}(g^*(x_0,u),u))du \\
= &o_p(1)
\end{aligned}$$

where g_0 is defined in Lemma A.5.4. The last inequality follows from Lemma A.5.4 and the dominated convergence theorem. □

Proof of Theorem RoC-NSP. It is straightforward that

$$\max_{u_j\in\mathcal{U}_0} |\hat{g}(x_0,u_j) - g^*(x_0,u_j)| \leq \sqrt{\sum_{u_j\in\mathcal{U}_0} (|\hat{g}(x_0,u_j) - g^*(x_0,u_j)|)'(|\hat{g}(x_0,u_j) - g^*(x_0,u_j)|)}$$

We derive the rate of convergence for the right hand side.

By Theorem Cons-NSP, $\hat{g}(\cdot)$ on $\mathcal{U}_0$ in the interior of $S(Y|d,x_0)$ with probability ap-

proaching one. Under this event, $\Psi(\cdot)$ is differentiable at $\hat{\boldsymbol{g}}(x_0, u_j)$ and thus

$$\Psi(\hat{\boldsymbol{g}}(x_0, u_j)) - \Psi(\boldsymbol{g}^*(x_0, u_j)) = \tilde{\Pi}_{NSP}(u_j) \cdot (\hat{\boldsymbol{g}}(x_0, u_j) - \boldsymbol{g}^*(x_0, u_j)) \tag{A.5.8}$$

where $\tilde{\Pi}_{NSP}(u_j)$ is the Jacobian evaluated at the mean value. Again, by uniform convergence and the full rank condition in Theorem ID-NSP, $\tilde{\Pi}_{NSP}(u_j)$ is full rank uniformly in $u_j \in \mathcal{U}_0$. Then,

$$\hat{\boldsymbol{g}}(x_0, u_j) - \boldsymbol{g}^*(x_0, u_j) = \left(\tilde{\Pi}'_{NSP}(u_j)\tilde{\Pi}_{NSP}(u_j)\right)^{-1}\tilde{\Pi}'_{NSP}(u_j) \cdot \left(\Psi(\hat{\boldsymbol{g}}(x_0, u_j)) - \Psi(\boldsymbol{g}^*(x_0, u_j)\right)$$

Therefore, by boundedness of all the conditional densities, there exists a constant $C > 0$ such that

$$\begin{aligned}
&\sum_{u_j \in \mathcal{U}_0} (\hat{\boldsymbol{g}}(x_0, u_j) - \boldsymbol{g}^*(x_0, u_j))'(\hat{\boldsymbol{g}}(x_0, u_j) - \boldsymbol{g}^*(x_0, u_j)) \\
\leq & C \sum_{u_j \in \mathcal{U}_0} \left(\Psi(\hat{\boldsymbol{g}}(x_0, u_j)) - \Psi(\boldsymbol{g}^*(x_0, u_j))\right)' W_g(u_j) \left(\Psi(\hat{\boldsymbol{g}}(x_0, u_j)) - \Psi(\boldsymbol{g}^*(x_0, u_j))\right)
\end{aligned} \tag{A.5.9}$$

Add and subtract $\hat{\Psi}(\hat{\boldsymbol{g}}(x_0, u_j))$, the right hands side of inequality (A.5.9) can be expanded to the sum of the following three terms for some $C', C'', C''' > 0$:

$$\begin{aligned}
(A): \quad & \sum_{u_j \in \mathcal{U}_0} \left(\widehat{\Psi}(\hat{\boldsymbol{g}}(x_0, u_j)) - \Psi(\hat{\boldsymbol{g}}(x_0, u_j))\right)' W_g(u_j) \left(\widehat{\Psi}(\hat{\boldsymbol{g}}(x_0, u_j)) - \Psi(\hat{\boldsymbol{g}}(x_0, u_j))\right) \\
\leq & C' J \sup_{\boldsymbol{y} \in S(Y|d)^3} \left(\widehat{\Psi}(\boldsymbol{y}) - \Psi(\boldsymbol{y})\right)' \left(\widehat{\Psi}(\boldsymbol{y}) - \Psi(\boldsymbol{y})\right) \\
= & O(J r_n^2)
\end{aligned} \tag{A.5.10}$$

where the last inequality is from Lemma A.5.2 and the rate condition $h_g/h_x \to 0$.

$$
\begin{aligned}
(B): \quad & \sum_{u_j \in \mathcal{U}_0} \big(\widehat{\Psi}(\hat{\boldsymbol{g}}(x_0,u_j)) - \Psi(\boldsymbol{g}^*(x_0,u_j))\big)' W_g(u_j) \big(\widehat{\Psi}(\hat{\boldsymbol{g}}(x_0,u_j)) - \Psi(\boldsymbol{g}^*(x_0,u_j))\big) \\
&= \sum_{u_j \in \mathcal{U}_0} \big(\widehat{\Psi}(\hat{\boldsymbol{g}}(x_0,u_j)) - \boldsymbol{u}_j\big)' W_g(u_j) \big(\widehat{\Psi}(\hat{\boldsymbol{g}}(x_0,u_j)) - \boldsymbol{u}_j\big) \\
&\le \sum_{j=1}^{J} \big(\widehat{\Psi}(\hat{\boldsymbol{g}}(x_0,u_j)) - \boldsymbol{u}_j\big)' W_g(u_j) \big(\widehat{\Psi}(\hat{\boldsymbol{g}}(x_0,u_j)) - \boldsymbol{u}_j\big) \\
&\quad + \lambda \sum_{j=2}^{J} \big(\hat{\boldsymbol{g}}(x_0,u_j) - \hat{\boldsymbol{g}}(x_0,u_{j-1})\big)' \big(\hat{\boldsymbol{g}}(x_0,u_j) - \hat{\boldsymbol{g}}(x_0,u_{j-1})\big) \\
&\le \sum_{j=1}^{J} \big(\widehat{\Psi}(\boldsymbol{g}_0(u_j)) - \boldsymbol{u}_j\big)' W_g(u_j) \big(\widehat{\Psi}(\boldsymbol{g}_0(u_j)) - \boldsymbol{u}_j\big) \\
&\quad + \lambda \sum_{j=2}^{J} \big(\boldsymbol{g}_0(u_j)) - \boldsymbol{g}_0(u_{j-1}))\big)' \big(\boldsymbol{g}_0(u_j)) - \boldsymbol{g}_0(u_{j-1}))\big)
\end{aligned}
$$

where the first inequality is due to non-negativity of the penalty. The second inequality is by the definition of the estimator. $\boldsymbol{g}_0$ is the same as in Lemma A.5.4. In particular, let $\boldsymbol{g}_0(u_j) = \boldsymbol{g}^*(x_0,u_j)$. Then the right hand side of the last inequality is

$$
C'' J \sup_{\boldsymbol{y} \in S(Y|d)^3} \big(\widehat{\Psi}(\boldsymbol{y}) - \Psi(\boldsymbol{y})\big)' \big(\widehat{\Psi}(\boldsymbol{y}) - \Psi(\boldsymbol{y})\big) + \lambda/J = O(J r_n^2) \tag{A.5.11}
$$

$$
\begin{aligned}
(C): \quad & 2\sum_{u_j\in\mathcal{U}_0}\big(\widehat{\Psi}(\hat{\boldsymbol{g}}(x_0,u_j))-\Psi(\hat{\boldsymbol{g}}(x_0,u_j))\big)'W_g(u_j)\big(\widehat{\Psi}(\hat{\boldsymbol{g}}(x_0,u_j))-\Psi(\boldsymbol{g}^*(x_0,u_j))\big)\\
=&2\sum_{u_j\in\mathcal{U}_0}\big(\widehat{\Psi}(\hat{\boldsymbol{g}}(x_0,u_j))-\Psi(\hat{\boldsymbol{g}}(x_0,u_j))\big)'W_g(u_j)\big(\widehat{\Psi}(\hat{\boldsymbol{g}}(x_0,u_j))-\Psi(\hat{\boldsymbol{g}}(x_0,u_j))\big)\\
&+2\sum_{u_j\in\mathcal{U}_0}\big(\widehat{\Psi}(\hat{\boldsymbol{g}}(x_0,u_j))-\Psi(\hat{\boldsymbol{g}}(x_0,u_j))\big)'W_g(u_j)\big(\Psi(\hat{\boldsymbol{g}}(x_0,u_j))-\Psi(\boldsymbol{g}^*(x_0,u_j))\big)\\
=&2\cdot(A)+2\sum_{u_j\in\mathcal{U}_0}\big(\widehat{\Psi}(\hat{\boldsymbol{g}}(x_0,u_j))-\Psi(\hat{\boldsymbol{g}}(x_0,u_j))\big)'W_g(u_j)\tilde{\Pi}_{NSP}\cdot\big(\hat{\boldsymbol{g}}(x_0,u_j)-\boldsymbol{g}^*(x_0,u_j)\big)\\
\leq&2\cdot(A)+2\sqrt{\sum_{j=1}^{J}C'''\cdot(A)}\cdot\sqrt{\sum_{u_j\in\mathcal{U}_0}(\hat{\boldsymbol{g}}(x_0,u_j)-\boldsymbol{g}^*(x_0,u_j))'(\hat{\boldsymbol{g}}(x_0,u_j)-\boldsymbol{g}^*(x_0,u_j))}\\
=&O_p(Jr_n^2)+O_p(\sqrt{J}r_n)\sqrt{\sum_{u_j\in\mathcal{U}_0}(\hat{\boldsymbol{g}}(x_0,u_j)-\boldsymbol{g}^*(x_0,u_j))'(\hat{\boldsymbol{g}}(x_0,u_j)-\boldsymbol{g}^*(x_0,u_j))}
\end{aligned}
\tag{A.5.12}
$$

where the inequality follows from the Cauchy-Schwartz inequality.

Combining equations (A.5.8), (A.5.9), (A.5.10), (A.5.11) and (A.5.12), we have the desired result. □

Proof of Theorem AsymDist-NSP. Fix $u_0 \in \mathcal{U}_0$. By Corollary RoC-NSP, $\hat{\boldsymbol{g}}(x_0,u_0)$ is in the interior so satisfies the first order condition. Under the rates of the bandwidths, the estimated propensity scores, the matching points and the penalty converge faster than $1/\sqrt{nh_g}$. By Lemma A.5.1, under some manipulation, we obtain the following expansion for $\hat{\boldsymbol{g}}(x_0,u_0)-\boldsymbol{g}^*(x_0,u_0)$:

$$
\begin{aligned}
\hat{\boldsymbol{g}}(x_0,u_0)-\boldsymbol{g}^*(x_0,u_0)=&-\big(\Pi_{NSP}'\boldsymbol{W}_g(u_0)\Pi_{NSP}\big)^{-1}\cdot\Pi_{NSP}'\boldsymbol{W}_g(u_0)\cdot\Big[\big(\hat{\Psi}_{NSP}(\boldsymbol{g}^*(\boldsymbol{x}_0,u_0))-\boldsymbol{u}\big)\\
&+\begin{pmatrix}0\\0\\\sum_{d=1}^{3}\phi_{d1}\big(\hat{F}_{Y|DXZ}(g_d^*(x_0,u_0)|d,x_0,0)-\hat{F}_{Y|DXZ}(g_d^*(x_{m1},u_0)|d,x_{m1},1)\big)\\\sum_{d=1}^{3}\phi_{d2}\big(\hat{F}_{Y|DXZ}(g_d^*(x_0,u_0)|d,x_0,1)-\hat{F}_{Y|DXZ}(g_d^*(x_{m2},u_0)|d,x_{m2},0)\big)\end{pmatrix}+o_p\Big(\frac{1}{\sqrt{Nh_g}}\Big)\Big]
\end{aligned}
$$

where Π_{NSP}, ϕ_{d1}, and ϕ_{d2} are as defined in Section 1.6.3. Recall that $\Psi_{NSP}\big(g^*(\boldsymbol{x}_0, u_0)\big) = \boldsymbol{u}$, $F_{Y|DXZ}(g_d^*(x_0, u_0)|d, x_0, 0) = F_{Y|DXZ}(g_d^*(x_{m1}, u_0)|d, x_{m1}, 1)$ and $F_{Y|DXZ}(g_d^*(x_0, u_0)|d, x_0, 1) = F_{Y|DXZ}(g_d^*(x_{m2}, u_0)|d, x_{m2}, 0)$. Also, by Lemma A.5.1, the denominator of $\hat{F}_{YDXZ}(y, d, x, z)$ converges in probability to $f_{Y|DXZ}(d, x, z)$. Let

$$\mathbb{G}_d(y, x, z) \equiv \hat{F}_{YDXZ}(y, d, x, z) - F_{Y|DXZ}(y|d, x, z)\hat{f}_{DXZ}(d, x, z).$$

Then the asymptotic distribution is determined by the following vector:

$$\begin{pmatrix} \sum_{d=1}^{3} \dfrac{\mathbb{G}_d(g_d^*(x_0, u_0), x_0, 0)}{f_{DXZ}(d, x_0, 0)} \\ \\ \sum_{d=1}^{3} \dfrac{\mathbb{G}_d(g_d^*(x_0, u_0), x_0, 1)}{f_{DXZ}(d, x_0, 1)} \\ \\ \sum_{d=1}^{3} \Big[\dfrac{\mathbb{G}_d(g_d^*(x_{m1}, u_0), x_{m1}, 0)}{f_{DXZ}(d, x_{m1}, 0)} + \phi_{d1}\dfrac{\mathbb{G}_d(g_d^*(x_0, u_0), x_0, 0)}{f_{DXZ}(d, x_0, 0)} - \phi_{d1}\dfrac{\mathbb{G}_d(g_d^*(x_{m1}, u_0), x_{m1}, 1)}{f_{DXZ}(d, x_{m1}, 1)}\Big] \\ \\ \sum_{d=1}^{3} \Big[\dfrac{\mathbb{G}_d(g_d^*(x_{m2}, u_0), x_{m2}, 1)}{f_{DXZ}(d, x_{m2}, 1)} + \phi_{d1}\dfrac{\mathbb{G}_d(g_d^*(x_0, u_0), x_0, 1)}{f_{DXZ}(d, x_0, 1)} - \phi_{d1}\dfrac{\mathbb{G}_d(g_d^*(x_{m2}, u_0), x_{m2}, 0)}{f_{DXZ}(d, x_{m2}, 0)}\Big] \end{pmatrix}$$

The variance of each $\mathbb{G}_d$ follows Theorem 2.2 in Li and Racine (2008):

$$\mathbb{V}(\mathbb{G}_d(y, x, z)) = \frac{\kappa f_{DXZ}(d, x, z) F_{Y|DXZ}(y|d, x, z) \cdot (1 - F_{Y|DXZ}(y|d, x, z))}{n h_g} + o(\frac{1}{n h_g})$$

Now let us derive the covariances: $\mathbb{C}\big(\mathbb{G}_d(y, x, z), \mathbb{G}_{d'}(y', x', z')\big)$. By i.i.d., the covariance is equal to

$$\frac{1}{n h_g^2}\mathbb{E}\big[L(\frac{y - Y_i}{h_0})L(\frac{y' - Y_i}{h_0})K(\frac{X_i - x}{h_g})K(\frac{X_i - x'}{h_g})\mathbb{1}(D_i = d)\mathbb{1}(D_i = d')\mathbb{1}(Z_i = z)\mathbb{1}(Z_i = z')\big] + O_p(\frac{1}{n})$$

where the $O_p(\frac{1}{n})$ term arises because the bias is of the order h_g^2 by Lemma A.2 in Li and

Racine (2008). It is clear that when $z' \neq z$ or $d' \neq d$, the leading term is o. When $x \neq x'$, for large enough n, $|x' - x| > 2h_g$, and thus $\left|\left|\frac{X_i - x}{h_g}\right| - \left|\frac{X_i - x'}{h_g}\right|\right| > 2$. As $K(\cdot) = 0$ outside $[-1, 1]$, for any X_i, one of the K functions must be o. Therefore, all the covariances are of the order of $O_p(\frac{1}{n})$ which is $o_p(\frac{1}{nh_g})$.

By Lyapunov's Central Limit Theorem and the delta method, we obtain the asymptotic distribution in the theorem. □

A.5.2 Proofs of Lemmas

Lemma A.5.1

We only show equation (A.5.3) as others are standard in the literature of kernel estimation (e.g. Mack and Silverman (1982), Silverman (1986), Härdle, Marron and Wand (1990), Masry (1996)), etc.). Let us first prove the following lemma which generalizes Lemma A.5 in Li and Racine (2008) to a bounded Y.

Lemma A.5.5. *Under the conditions in Lemma A.5.1, we have the following equations for all d and z:*

$$\sup_{\substack{y \in [\underline{y}_{dx}, \bar{y}_{dx}] \\ x \in S_0(X)}} \left| \mathbb{E}\left(L(\frac{y - Y_i}{h_0}) | D_i = d, X_i = x, Z_i = z\right) - F_{Y|DXZ}(y|d, x, z) \right| = O(h_0) \quad \text{(A.5.13)}$$

Proof of Lemma A.5.5. By definition,

$$\mathbb{E}\left(L(\frac{y - Y_i}{h_0}) | D_i = d, X_i = x, Z_i = z\right) = \int_{\underline{y}_{dx}}^{\bar{y}_{dx}} L(\frac{y - y'}{h_0}) dF_{Y|DXZ}(y'|d, x, z)$$

Let $\nu = \frac{y - y'}{h_0}$, by the rule of change-of-variable and integration by part, the right hand

side can be rewritten as

$$
\begin{aligned}
&-\int_{\frac{y-\bar{y}_{dx}}{h_0}}^{\frac{y-\underline{y}_{dx}}{h_0}} L(v)dF_{Y|DXZ}(y-vh_0|d,x,z) \\
=&L(\frac{y-\bar{y}_{dx}}{h_0})+\int_{\frac{y-\bar{y}_{dx}}{h_0}}^{\frac{y-\underline{y}_{dx}}{h_0}} L'(v)F_{Y|DXZ}(y-vh_0|d.x,z)dv \\
=&L(\frac{y-\bar{y}_{dx}}{h_0})+F_{Y|DXZ}(y|d,x,z)\big(L(\frac{y-\underline{y}_{dx}}{h_0})-L(\frac{y-\bar{y}_{dx}}{h_0})\big) \\
&+h_0\int_{\frac{y-\bar{y}_{dx}}{h_0}}^{\frac{y-\underline{y}_{dx}}{h_0}} vf_{Y|DXZ}(y-\tilde{v}h_0|d,x,z)L'(v)dv
\end{aligned}
$$

where $\tilde{v}$ is a mean value between 0 and v. Rearrange the terms, we obtain

$$
\begin{aligned}
&\big|\mathbb{E}\big(L(\frac{y-Y_i}{h_0})|D_i=d,X_i=x,Z_i=z\big)-F_{Y|DXZ}(y|d,x,z)\big| \\
\leq&\big|L(\frac{y-\bar{y}_{dx}}{h_0})\big(1-F_{Y|DXZ}(y|d,x,z)\big)-\big(1-L(\frac{y-\underline{y}_{dx}}{h_0})\big)F_{Y|DXZ}(y|d,x,z)\big| \\
&+h_0\sup_{(y,x)\in S(Y,X|d)} f_{Y|DXZ}(y|d,x,z)\int_0^1 vL'(v)du
\end{aligned}
$$

where the last term is $O(h_0)$. For the first term, if $\underline{y}_{dx}+h_0<y<\bar{y}_{dx}-h_0$, then $L(\frac{y-Y_i}{h_0})=0$ and $\big(1-L(\frac{y-\underline{y}_{dx}}{h_0})\big)=0$. Now we consider the remaining cases. For $\underline{y}_{dx}\leq y\leq \underline{y}_{dx}+h_0$, $L(\frac{y-Y_i}{h_0})=0$. Then the first term is bounded by $F_{Y|DXZ}(\underline{y}_{dx}+h_0|dxz)$, which is $O(h_0)$ by the mean value theorem. Similarly, if $\bar{y}_{dx}-h_0\leq y\leq \bar{y}_{dx}$, then the same term is bounded by $1-F_{Y|DXZ}(\bar{y}_{dx}-h_0|dxz)$, which is again $O(h_0)$. As Y's conditional density is uniformly bounded, these bounds are uniform on $S(Y,X|d)$ □

Remark A.5.1. The rate in Li and Racine (2008) is $O(h_0^2)$, faster than the rate here. Intuitively, this is because at the boundaries, L systematically overestimate (at the lower bound) or underestimate (at the upper bound) the CDF, thus introducing larger bias.

Now we are ready to show equation (A.5.3).

Proof of Equation (A.5.3) *in Lemma A.5.1.* By construction, $\hat{F}_{Y|DXZ}(\cdot|d,x,z)\in[0,1]$ and is

increasing. Therefore,

$$
\begin{aligned}
&\sup_{\substack{y\in S(Y|d)\\ x\in S_0(X)}} \left|\hat{F}_{Y|DXZ}(y|d,x,z) - F_{Y|DXZ}(y|d,x,z)\right| \\
\leq &\sup_{x\in S_0(X)} \sup_{y\leq \underline{y}_{dx}} \left|\hat{F}_{Y|DXZ}(y|d,x,z) - F_{Y|DXZ}(y|d,x,z)\right| \\
&+ \sup_{x\in S_0(X)} \sup_{y\geq \bar{y}_{dx}} \left|\hat{F}_{Y|DXZ}(y|d,x,z) - F_{Y|DXZ}(y|d,x,z)\right| \\
&+ \sup_{x\in S_0(X)} \sup_{\underline{y}_{dx}\leq y\leq \bar{y}_{dx}} \left|\hat{F}_{Y|DXZ}(y|d,x,z) - F_{Y|DXZ}(y|d,x,z)\right| \\
= &\sup_{x\in S_0(X)} \sup_{y\leq \underline{y}_{dx}} \hat{F}_{Y|DXZ}(y|d,x,z) + \sup_{x\in S_0(X)} \sup_{y\geq \bar{y}_{dx}} \left(1 - \hat{F}_{Y|DXZ}(y|d,x,z)\right) \\
&+ \sup_{x\in S_0(X)} \sup_{\underline{y}_{dx}\leq y\leq \bar{y}_{dx}} \left|\hat{F}_{Y|DXZ}(y|d,x,z) - F_{Y|DXZ}(y|d,x,z)\right| \\
\leq &\sup_{x\in S_0(X)} \hat{F}_{Y|DXZ}(\underline{y}_{dx}|d,x,z) + \sup_{x\in S_0(X)} \left(1 - \hat{F}_{Y|DXZ}(\bar{y}_{dx}|d,x,z)\right) \\
&+ \sup_{x\in S_0(X)} \sup_{\underline{y}_{dx}\leq y\leq \bar{y}_{dx}} \left|\hat{F}_{Y|DXZ}(y|d,x,z) - F_{Y|DXZ}(y|d,x,z)\right| \\
\leq &3 \sup_{x\in S_0(X)} \sup_{\underline{y}_{dx}\leq y\leq \bar{y}_{dx}} \left|\hat{F}_{Y|DXZ}(y|d,x,z) - F_{Y|DXZ}(y|d,x,z)\right| \\
= &3 \sup_{x\in S_0(X)} \sup_{\underline{y}_{dx}\leq y\leq \bar{y}_{dx}} \left|\frac{\hat{F}_{YDXZ}(y,d,x,z) - F_{Y|DXZ}(y|d,x,z)\hat{f}_{DXZ}(d,x,z)}{\hat{f}_{DXZ}(d,x,z)}\right|
\end{aligned}
$$

We now show that the right hand side of the last inequality has the desired rate. As the denominator is independent of y, and it's uniformly consistent for $f_{DXZ}(d,x,z)$ on the interior set $S_0(X)$ under the regularity conditions, it's infimum is bounded away from 0.

For the numerator, denote $\hat{q}(y,x) \equiv \hat{F}_{YDXZ}(y,d,x,z) - F_{Y|DXZ}(y|d,x,z)\hat{f}_{DXZ}(d,x,z)$,

then

$$\sup_{x\in S_0(X)}\sup_{\underline{y}_{dx}\le y\le \bar{y}_{dx}}\left|\hat{q}(y,x)\right| \le \sup_{x\in S_0(X)}\sup_{\underline{y}_{dx}\le y\le \bar{y}_{dx}}\left|\mathbb{E}(\hat{q}(y,x))\right| + \sup_{x\in S_0(X)}\sup_{\underline{y}_{dx}\le y\le \bar{y}_{dx}}\left|\hat{q}(y,x)-\mathbb{E}(\hat{q}(y,x))\right|$$

We derive the rate of convergence for each of the two terms.

By the law of iterated expectation,

$$\left|\mathbb{E}(\hat{q}(y,x))\right| = \left|\frac{1}{h_g}\int\left[\mathbb{E}_{Y|DXZ}\left(L(\frac{y-Y_i}{h_0})|d,x',z\right)-F_{Y|DXZ}(y|d,x,z)\right]K(\frac{x'-x}{h_g})f_X(x')dx'\right|$$

By Lemma A.5.5, $\mathbb{E}_{Y|DXZ}\left(L(\frac{y-Y_i}{h_0})|d,x',z\right) = F_{Y|DXZ}(y|d,x',z)+O(h_0)$ uniformly. Therefore, $\left|\mathbb{E}(\hat{q}(y,x))\right|$ is uniformly bounded by

$$\left|\frac{1}{h_g}\int\left[F_{Y|DXZ}(y|d,x',z)-F_{Y|DXZ}(y|d,x,z)\right]K(\frac{x'-x}{h_g})f_X(x')dx'\right|+O(h_0)$$

Recall the supremum is taken on $[\underline{y}_{dx},\bar{y}_{dx}]$, note that it may be the case that y is outside the support for some x'' in between of x and x'. Then $F_{Y|DXZ}(y|d,\cdot,z)$ is non-differentiable, making the second order Taylor expansion invalid at x. However, by Lipschitz continuity of $F_{Y|DXZ}(y|d,\cdot,z)$, we can still bound the CDF difference:

$$\begin{aligned}
&\sup_{x\in S_0(X)}\sup_{\underline{y}_{dx}\le y\le \bar{y}_{dx}}\left|\frac{1}{h_g}\int\left[F_{Y|DXZ}(y|d,x',z)-F_{Y|DXZ}(y|d,x,z)\right]K(\frac{x'-x}{h_g})f_X(x')dx'\right|\\
\le& C\sup_{x\in S_0(X)}\int\left|\frac{x'-x}{h_g}\right|\cdot K(\frac{x'-x}{h_g})f_X(x')dx'\\
\le& h_g C'\int_0^1 \nu K(\nu)d\nu\\
=& O(h_g)
\end{aligned}$$

for some $C, C' > 0$.

Next let us derive the rate of $\sup_{(x,y)\in S_0(X,Y|d)} \left|\hat{q}(y,x) - \mathbb{E}(\hat{q}(y,x))\right|$ where for the ease of notation, $S_0(X,Y|d) \equiv \{(x,y) : x \in S_0(X), \underline{y}_{dx} \leq y \leq \bar{y}_{dx}\}$. As $S_0(X,Y|d)$ is compact, it can be covered by $T_n < \infty$ squares $I_1, I_2, ..., I_{T_n}$ with length τ_n where $\tau_n \propto O(\sqrt{1/T_n})$. Let the center in each square be (y_k, x_k) $(k = 1, 2, ..., T_n)$, then

$$
\begin{aligned}
&\sup_{(x,y)\in S_0(X,Y|d)} \left|\hat{q}(y,x) - \mathbb{E}(\hat{q}(y,x))\right| \\
\leq &\max_{1\leq k\leq T_n} \sup_{S_0(X,Y|d)\cap I_k} \left|\hat{q}(y,x) - \hat{q}(y_k,x_k)\right| \\
&+ \max_{1\leq k\leq T_n} \sup_{S_0(X,Y|d)\cap I_k} \left|\mathbb{E}(\hat{q}(y,x)) - \mathbb{E}(\hat{q}(y_k,x_k))\right| \\
&+ \max_{1\leq k\leq T_n} \left|\hat{q}(y_k,x_k) - \mathbb{E}(\hat{q}(y_k,x_k))\right|
\end{aligned}
$$

Consider the first and the second terms. For the first term,

$$
\begin{aligned}
&\hat{q}(y,x) - \hat{q}(y_k,x_k) \\
=&\frac{1}{nh_g}\sum_{i=1}^{n} \mathbb{1}(D_i = d)\mathbb{1}(Z_i = d)\left(L(\frac{y - Y_i}{h_0})K(\frac{X_i - x}{h_g}) - L(\frac{y_k - Y_i}{h_0})K(\frac{X_i - x_k}{h_g})\right)
\end{aligned}
$$

By Lipschitz continuity of $L(\cdot)$ and $K(\cdot)$, the right hand side is bounded by $\frac{C\tau_n}{h_g h_0}$ given $h_0 < h_g$ uniformly in k. Similarly the second term is also bounded by $\frac{C\tau_n}{h_g h_0}$.

For the third term, let

$$
\begin{aligned}
W_i(y,x) =&\frac{1}{nh_g}\Big[\big(L(\frac{y - Y_i}{h_0}) - F_{Y|DXZ}(y|dxz)\big)\mathbb{1}(D_i = d)\mathbb{1}(Z_i = z)K(\frac{x - X_i}{h_g}) \\
&- \mathbb{E}\big(\big(L(\frac{y - Y_i}{h_0}) - F_{Y|DXZ}(y|dxz)\big)\mathbb{1}(D_i = d)\mathbb{1}(Z_i = z)K(\frac{x - X_i}{h_g})\big)\Big]
\end{aligned}
$$

We consider the probability $\mathbb{P}\big(\max_{1\le k\le T_n}\big|\sum_{i=1}^n W_i(y_k,x_k)\big|>C_1\sqrt{\frac{\log(n)}{nh_g}}\big)$.

$$
\begin{aligned}
&\mathbb{P}\big(\max_{1\le k\le T_n}\big|\sum_{i=1}^n W_i(y_k,x_k)\big|>C_1\sqrt{\frac{\log(n)}{nh_g}}\big)\\
\le&\sum_{k=1}^{T_n}\mathbb{P}\big(\big|\sum_{i=1}^n W_i(y_k,x_k)\big|>C_1\sqrt{\frac{\log(n)}{nh_g}}\big)\\
\le&T_n\sup_{(y,x)\in S_0(Y,X|d)}\mathbb{P}\big(\big|\sum_{i=1}^n W_i(y,x)\big|>C_1\sqrt{\frac{\log(n)}{nh_g}}\big)\\
\le&T_n\sup_{(y,x)\in S_0(Y,X|d)}\Big(\mathbb{P}\big(\sum_{i=1}^n W_i(y,x)>C_1\sqrt{\frac{\log(n)}{nh_g}}\big)+\mathbb{P}\big(\sum_{i=1}^n W_i(y,x)<C_1\sqrt{\frac{\log(n)}{nh_g}}\big)\Big)\\
\le&T_n\sup_{(y,x)\in S_0(Y,X|d)}\frac{\mathbb{E}\big(\exp(a_n\sum_i W_i(y,x))\big)+\mathbb{E}\big(\exp(-a_n\sum_i W_i(y,x))\big)}{\exp\big(a_nC_1\sqrt{\frac{\log(n)}{nh_g}}\big)}
\end{aligned}
$$

where the last inequality follows from the Markov inequality for some $a_n>0$. Since K and L are bounded, $|W_i(y,x)|$ is bounded. Let $a_n=\sqrt{\log(n)nh_g}$, then for large enough n, $a_n|W_i(y,x)|<1/2$. Therefore, by the inequality $\exp(c)\le 1+c+c^2$ for any $c\in[-1/2,1/2]$ and $1+c\le\exp(c)$ for $c\ge 0$, we have

$$
\begin{aligned}
\mathbb{E}\big(\exp(\pm a_nW_i(y,x))\big)\le&1\pm\mathbb{E}(a_nW_i(y,x))+\mathbb{E}(a_n^2W_i^2(y,x)))\\
\le&\exp(\mathbb{E}a_n^2W_i^2(y,x)))
\end{aligned}
$$

since $\mathbb{E}(W_i(y,x))$ by construction. Therefore,

$$
\begin{aligned}
&T_n\sup_{(y,x)\in S_0(Y,X|d)}\frac{\mathbb{E}\big(\exp(a_n\sum_i W_i(y,x))\big)+\mathbb{E}\big(\exp(-a_n\sum_i W_i(y,x))\big)}{\exp\big(a_nC_1\sqrt{\frac{\log(n)}{nh_g}}\big)}\\
\le&\frac{2T_n}{n^{C_1}}\cdot\sup_{(y,x)\in S_0(Y,X|d)}\exp(\log(n)nh_g\sum_i\mathbb{E}W_i^2(y,x)))
\end{aligned}
$$

For $\mathbb{E}W_i^2(y,x)$, since $L(\cdot) \in [0,1]$, we have

$$\begin{aligned}
\mathbb{E}W_i^2(y,x) \leq & \frac{1}{n^2 h_g^2}\mathbb{E}\Big[\big(L(\frac{y-Y_i}{h_0}) - F_{Y|DXZ}(y|dxz)\big)\mathbb{1}(D_i=d)\mathbb{1}(Z_i=z)K(\frac{x-X_i}{h_g})\Big]^2 \\
\leq & \frac{1}{n^2 h_g^2}\mathbb{E}\big[K^2(\frac{x-X_i}{h_g})\big] \\
\leq & C_2\frac{1}{n^2 h_g}
\end{aligned}$$

for some $C_2 > 0$. Therefore,

$$\frac{2T_n}{n^{C_1}} \cdot \sup_{(y,x)\in S_0(Y,X|d)} \exp(\log(n) n h_g \sum \mathbb{E}W_i^2(y,x))) \leq \frac{2T_n}{n^{C_1-C_2}}$$

Let $T_n = \frac{n}{\log(n)h_0^2 h_g}$. Then for large enough C_1, there exists $\alpha \geq 2$,

$$\sum_{i=1}^{n} \mathbb{P}\big(\max_{1\leq k\leq T_n} \big|\hat{q}(y_k,x_k) - \mathbb{E}(\hat{q}(y_k,x_k))\big| > C_1\sqrt{\frac{\log(n)}{nh_g}}\big) \leq \sum_n \frac{1}{n^\alpha} < \infty$$

Therefore, by the Borel-Cantelli lemma,

$$\max_{1\leq k\leq T_n} \big|\hat{q}(y_k,x_k) - \mathbb{E}(\hat{q}(y_k,x_k))\big| = O_p(\sqrt{\frac{\log(n)}{nh_g}}).$$

With this choice of T_n, $\tau_n = O(\frac{\sqrt{\log(n)h_g}h_0}{\sqrt{n}})$, so the first two terms are $O_p(\sqrt{\frac{\log(n)}{nh_g}})$ as well. □

Lemma A.5.2

Proof. By the triangle inequality,

$$
\begin{aligned}
&\sup_{y\in[\underline{y}_d,\bar{y}_d]} \left|\hat{F}_{Y|DXZ}(\hat{\varphi}_d(y;\hat{x}_m,z')|d,\hat{x}_m,z) - F_{Y|DXZ}(\varphi_d(y;x_m,z')|d,x_m,z)\right| \\
\leq &\sup_{y\in[\underline{y}_d,\bar{y}_d]} \left|\hat{F}_{Y|DXZ}(\hat{\varphi}_d(y;\hat{x}_m,z')|d,\hat{x}_m,z) - F_{Y|DXZ}(\hat{\varphi}_d(y;\hat{x}_m,z')|d,\hat{x}_m,z)\right| \\
&+ \sup_{y\in[\underline{y}_d,\bar{y}_d]} \left|F_{Y|DXZ}(\hat{\varphi}_d(y;\hat{x}_m,z')|d,\hat{x}_m,z) - F_{Y|DXZ}(\varphi_d(y;x_m,z')|d,x_m,z)\right| \\
\leq &\sup_{\substack{y\in[\underline{y}_d,\bar{y}_d]\\ x\in S_0(X)}} \left|\hat{F}_{Y|DXZ}(y|d,x,z) - F_{Y|DXZ}(y|d,x,z)\right| \\
&+ \sup_{y\in[\underline{y}_d,\bar{y}_d]} \left|F_{Y|DXZ}(\hat{\varphi}_d(y;\hat{x}_m,z')|d,\hat{x}_m,z) - F_{Y|DXZ}(\varphi_d(y;x_m,z')|d,x_m,z)\right| \\
\leq &\sup_{\substack{y\in[\underline{y}_d,\bar{y}_d]\\ x\in S_0(X)}} \left|\hat{F}_{Y|DXZ}(y|d,x,z) - F_{Y|DXZ}(y|d,x,z)\right| \\
&+ \sup_{y\in[\underline{y}_d,\bar{y}_d]} \left|F_{Y|DXZ}(\hat{\varphi}_d(y;\hat{x}_m,z')|d,x_m,z) - F_{Y|DXZ}(\varphi_d(y;x_m,z')|d,x_m,z)\right| + O_p(a_n)
\end{aligned}
$$

where the last inequality holds because of Lipschitz continuity of $F_{Y|DXZ}(\cdot|d,\cdot,z)$. Note the first term on the right hand side is $O_p(r_n)$ by Lemma A.5.1. Now we only need to show that

$$
\sup_{y\in[\underline{y}_d,\bar{y}_d]} \left|F_{Y|DXZ}(\hat{\varphi}_d(y;\hat{x}_m,z')|d,x_m,z) - F_{Y|DXZ}(\varphi_d(y;x_m,z')|d,x_m,z)\right| = O_p(r_n+a_n)
$$

This is straightforward if $\hat{\varphi}_d(y)$ is uniformly consistent at that rate. However, as $[\underline{y}_d,\bar{y}_d]$ is larger than $[\underline{y}_{dx_0},\bar{y}_{dx_0}]$, $\hat{\varphi}_d(y)$ is not consistent when y is outside $[\underline{y}_{dx_0},\bar{y}_{dx_0}]$ because then $\varphi_d(y)$ is not unique. Let us prove the following equation first:

$$
\sup_{y\in[\underline{y}_d,\bar{y}_d]} \left|F_{Y|DXZ}(\hat{\varphi}_d(y;\hat{x}_m,z')|d,x_m,z') - F_{Y|DXZ}(\varphi_d(y;x_m,z')|d,x_m,z')\right| = O_p(r_n+a_n)
$$

By adding and subtracting $\hat{F}_{Y|DXZ}(\hat{\varphi}_d(y;\hat{x}_m,z')|d,\hat{x}_m,z')$,

$$
\begin{aligned}
&\sup_{y\in[\underline{y}_d,\bar{y}_d]} \left|F_{Y|DXZ}(\hat{\varphi}_d(y;x_m,z')|d,\hat{x}_m,z') - F_{Y|DXZ}(\varphi_d(y;x_m,z')|d,x_m,z')\right| \\
\leq &\sup_{y\in[\underline{y}_d,\bar{y}_d]} \left|\hat{F}_{Y|DXZ}(\hat{\varphi}_d(y;\hat{x}_m,z')|d,\hat{x}_m,z') - F_{Y|DXZ}(\hat{\varphi}_d(y;\hat{x}_m,z')|d,x_m,z')\right| \\
&+ \sup_{y\in[\underline{y}_d,\bar{y}_d]} \left|\hat{F}_{Y|DXZ}(\hat{\varphi}_d(y;\hat{x}_m,z')|d,\hat{x}_m,z') - F_{Y|DXZ}(\varphi_d(y;x_m,z')|d,x_m,z')\right| \\
\leq &\sup_{y\in[\underline{y}_d,\bar{y}_d]} \left|\hat{F}_{Y|DXZ}(\hat{\varphi}_d(y;\hat{x}_m,z')|d,x_m,z') - F_{Y|DXZ}(y|d,x_0,z)\right| + O_p(r_n+a_n)
\end{aligned}
$$

where we use $F_{Y|DXZ}(\varphi_d(y;x_m,z')|d,x_m,z') = F_{Y|DXZ}(y|d,x_0,z)$ for the last inequality by Theorem MEQ. By adding and subtracting $\hat{F}_{Y|DXZ}(y|d,x_0,z)$,

$$
\begin{aligned}
&\sup_{y\in[\underline{y}_d,\bar{y}_d]} \left|\hat{F}_{Y|DXZ}(\hat{\varphi}_d(y;\hat{x}_m,z')|d,\hat{x}_m,z') - F_{Y|DXZ}(y|d,x_0,z)\right| \\
\leq &\sup_{y\in[\underline{y}_d,\bar{y}_d]} \left|\hat{F}_{Y|DXZ}(\hat{\varphi}_d(y;\hat{x}_m,z')|d,\hat{x}_m,z') - \hat{F}_{Y|DXZ}(y|d,x_0,z)\right| + O_p(r_n) \\
\leq &\sup_{y\in[\underline{y}_d,\bar{y}_d]} \left|\hat{F}_{Y|DXZ}(\varphi_d(y;x_m,z')|d,\hat{x}_m,z') - \hat{F}_{Y|DXZ}(y|d,x_0,z)\right| + O_p(r_n) \\
= &\sup_{y\in[\underline{y}_d,\bar{y}_d]} \left|\hat{F}_{Y|DXZ}(\varphi_d(y;x_m,z')|d,\hat{x}_m,z') - F_{Y|DXZ}(\varphi_d(y;x_m,z')|d,x_m,z')\right| + O_p(r_n) \\
\leq &\sup_{y\in[\underline{y}_d,\bar{y}_d]} \left|\hat{F}_{Y|DXZ}(\varphi_d(y;x_m,z')|d,\hat{x}_m,z') - F_{Y|DXZ}(\varphi_d(y;x_m,z')|d,\hat{x}_m,z')\right| + O_p(r_n+a_n) \\
= &O_p(r_n+a_n)
\end{aligned}
$$

where the third line follows from the definition of $\hat{\varphi}_d(y;\hat{x}_m,z')$. Finally, let us establish the relation between $F_{Y|DXZ}(y'|d,x,z) - F_{Y|DXZ}(y|d,x,z)$ and $F_{Y|DXZ}(y'|d,x,z') - F_{Y|DXZ}(y|d,x,z')$ for any $y,y' \in [\underline{y}_d,\bar{y}_d]$. By the exogeneity condition of Z in Assumption E-NSP, $S(Y|d,x,z) = S(Y|d,x,z') = S(Y|d)$. Therefore, we have the following two

equations:

$$F_{Y|DXZ}(y'|d,x,z) - F_{Y|DXZ}(y|d,x,z) = f_{Y|DXZ}(\tilde{y}_1|d,x,z)\big[\big(\underline{y}_{dx} \vee (y' \wedge \bar{y}_{dx})\big) - \big(\underline{y}_{dx} \vee (y \wedge \bar{y}_{dx})\big)\big],$$

and

$$F_{Y|DXZ}(y'|d,x,z') - F_{Y|DXZ}(y|d,x,z') = f_{Y|DXZ}(\tilde{y}_2|d,x,z')\big[\big(\underline{y}_{dx} \vee (y' \wedge \bar{y}_{dx})\big) - \big(\underline{y}_{dx} \vee (y \wedge \bar{y}_{dx})\big)\big].$$

where $\tilde{y}_1$ and $\tilde{y}_2$ are the mean values between $\big(\underline{y}_{dx} \vee (y' \wedge \bar{y}_{dx})\big)$ and $\big(\underline{y}_{dx} \vee (y \wedge \bar{y}_{dx})\big)$.

By Assumption Reg-NSP, $\frac{f_{Y|DXZ}(\tilde{y}_1|d,x,z)}{f_{Y|DXZ}(\tilde{y}_2|d,x,z')}$ is uniformly bounded from above on $S(Y|d)$ $\times S(Y|d)$. Therefore, there exists a constant $C > 0$ such that

$$\begin{aligned}
&\sup_{y\in[\underline{y}_d,\bar{y}_d]} \big|F_{Y|DXZ}(\hat{\varphi}_d(y;\hat{x}_m,z')|d,x_m,z') - F_{Y|DXZ}(\varphi_d(y;x_m,z')|d,x_m,z')\big| \\
\leq & C \sup_{y\in[\underline{y}_d,\bar{y}_d]} \big|F_{Y|DXZ}(\hat{\varphi}_d(y;\hat{x}_m,z')|d,x_m,z) - F_{Y|DXZ}(\varphi_d(y;x_m,z')|d,x_m,z)\big|
\end{aligned}$$

This completes the proof. □

Lemma A.5.3

Proof of Lemma A.5.3. By the triangle inequality,

$$
\begin{aligned}
&\sup_{g\in\hat{\mathcal{G}}}\Big|\frac{1}{J}\sum_{j=1}^{J}\hat{Q}_{NSP}(g(u_j),u_j)-\int_0^1 Q_{NSP}(g(u),u)du\Big| \\
\leq&\sup_{g\in\hat{\mathcal{G}}}\Big|\frac{1}{J}\sum_{j=1}^{J}\hat{Q}_{NSP}(g(u_j),u_j)-\frac{1}{J}\sum_{j=1}^{J}Q_{NSP}(g(u_j),u_j)\Big| \\
&+\sup_{g\in\hat{\mathcal{G}}}\Big|\frac{1}{J}\sum_{j=1}^{J}Q_{NSP}(g(u_j),u_j)-\int_0^1 Q_{NSP}(g(u),u)du\Big| \\
\leq&\sup_{y\in\prod_d S(Y|d)}\big|\hat{Q}_{NSP}(y,u_j)-Q_{NSP}(y,u_j)\big| \\
&+\sup_{g\in\hat{\mathcal{G}}}\Big|\frac{1}{J}\sum_{j=1}^{J}Q_{NSP}(g(u_j),u_j)-\int_0^1 Q_{NSP}(g(u),u)du\Big| \\
=&\sup_{g\in\hat{\mathcal{G}}}\Big|\frac{1}{J}\sum_{j=1}^{J}Q_{NSP}(g(u_j),u_j)-\int_0^1 Q_{NSP}(g(u),u)du\Big|+o_p(1)
\end{aligned}
$$

where the $o_p(1)$ in the last inequality follows from Lemma A.5.2. We now show the remaining term is also $o_p(1)$.

$$
\begin{aligned}
&\sup_{g\in\hat{\mathcal{G}}}\Big|\frac{1}{J}\sum_{j=1}^{J}Q_{NSP}(g(u_j),u_j)-\int_0^1 Q_{NSP}(g(u),u)du\Big| \\
=&\sup_{g\in\hat{\mathcal{G}}}\Big|\frac{1}{J}\sum_{j=1}^{J}Q_{NSP}(g(u_j),u_j)-\sum_{j=1}^{J}\int_{\frac{j-1}{J}}^{\frac{j}{J}}Q_{NSP}(g(u),u)du\Big| \\
=&\sup_{g\in\mathcal{G}^*}\Big|\sum_{j=1}^{J}\Big(\frac{1}{J}Q_{NSP}(g(u_j),u_j)-\int_{\frac{j-1}{J}}^{\frac{j}{J}}Q_{NSP}(g(u),u)du\Big)\Big| \\
=&\sup_{g\in\mathcal{G}^*}\Big|\sum_{j=1}^{J}\Big(\int_{\frac{j-1}{J}}^{\frac{1}{J}}Q_{NSP}(g(u_j),u_j)du-\int_{\frac{j-1}{J}}^{\frac{j}{J}}Q_{NSP}(g(u),u)du\Big)\Big| \\
\leq&\sup_{g\in\mathcal{G}^*}\Big|\sum_{j=1}^{J}\Big(\int_{\frac{j-1}{J}}^{\frac{1}{J}}C\cdot(g(u_j)-g(u_{j-1}))du\Big)\Big| \\
\leq&\frac{1}{J}C\cdot\sum_{j=1}^{J}\big|g(u_j)-g(u_{j-1})\big|
\end{aligned}
$$

where in the second equality, $\hat{\mathcal{G}}$ is changed to $\mathcal{G}^*$ because for any $\boldsymbol{g}(u)$ that is outside $\prod_d S(Y|d, x_0)$ gives the same value of Q_{NSP} as that at the boundaries of $\prod_d S(Y|d, x_0)$. The third equality is due to the fact that $Q_{NSP}(\boldsymbol{g}(u_j), u_j)$ is a constant so $\int_{\frac{j-1}{J}}^{\frac{j}{J}} Q_{NSP}(\boldsymbol{g}(u_j), u_j)du = \frac{1}{J} Q_{NSP}(\boldsymbol{g}(u_j), u_j)$. The first inequality holds because for all values in $\prod_d S(Y|d, x_0)$, Q_{NSP} is differentiable and the derivative is uniformly bounded. Finally, by monotonicity, $\boldsymbol{g}(u_j) - \boldsymbol{g}(u_{j-1}) > \mathbf{0}$. Hence

$$\frac{1}{J} C \cdot \sum_{j=1}^{J} \left| \boldsymbol{g}(u_j) - \boldsymbol{g}(u_{j-1}) \right| \leq \frac{1}{J} C \sum_d (\bar{y}_{dx_0} - \underline{y}_{dx_0}) = O(\frac{1}{J}) = o(1)$$

□

Lemma A.5.4

Proof of Lemma A.5.4. For each u_j, let $\boldsymbol{g}_0(u_j) = \boldsymbol{g}^*(x_0, u_j)$. Then for any u between nodes u_{j-1} and u_j, by monotonicity,

$$|\boldsymbol{g}_0(u) - \boldsymbol{g}^*(x_0, u)| \leq \boldsymbol{g}^*(x_0, u_j) - \boldsymbol{g}^*(x_0, u_{j-1}) = O(\frac{1}{J}) = o(1).$$

□

Appendix B

Appendix to Chapter 2

B.1 Proof of Lemma 2.3.2

The proof closely follows the proof of Lemmas 3 and 4 in Candès and Recht (2009) and the proof of Lemmas 2.8 and 2.9 in Candès et al. (2011). We record it here only to make it easy to see where modifications are made to accommodate a converging δ by construction. Throughout, we condition on the event that the conditions in Assumption 2.3.1 hold.

We begin by stating three lemmas in Candès and Recht (2009) and Candès et al. (2011).

Lemma B.1.1 (Candès and Recht (2009), Theoerm 4.1; Candès et al. (2011), Theorem 2.6)**.** *Suppose* $\Omega_0 \sim Ber(\delta_0)$. *Then with high probability,*

$$||\mathcal{P}_\Phi - \delta_0^{-1}\mathcal{P}_\Phi\mathcal{P}_{\Omega_0}\mathcal{P}_\Phi|| \leq \varepsilon$$

provided that $\delta_0 \geq C\varepsilon^{-2}(\mu r \log(N \vee T)/(N \wedge T)$.

Lemma B.1.2 (Candès et al. (2011), Lemma 3.1)**.** *Suppose* $X \in \Phi$ *is a fixed matrix, and*

$\Omega_0 \sim Ber(\delta_0)$. *Then, with high probability,*

$$||X - \delta_0^{-1}\mathcal{P}_\Phi \mathcal{P}_{\Omega_0} X||_\infty \leq \varepsilon ||X||_\infty$$

provided that $\delta_0 \geq C\varepsilon^{-2}(\mu r \log(N \vee T)/(N \wedge T)$.

Lemma B.1.3 (Candès and Recht (2009), Theorem 6.3; Candès et al. (2011), Lemma 3.2)**.** *Suppose* X *is fixed, and* $\Omega_0 \sim Ber(\delta_0)$. *Then, with high probability,*

$$||(I - \delta_0^{-1}\mathcal{P}_{\Omega_0})X|| \leq C\sqrt{\frac{(N \vee T)\log(N \vee T)}{\delta_0}}||X||_\infty$$

for some small numerical constant $C > 0$ *provided that* $\delta_0 \geq C(\mu r \log(N \vee T)/(N \wedge T)$.

Under the conditions in Lemma 2.3.2, $N \asymp T$, so in the subsequent proof, $N \vee T$ and $N \wedge T$ will simply be written as N. Following Candès et al. (2011), let $X_j = UV^* - \mathcal{P}_\Phi Q_j$ where Q_j is defined in (2.3.6). Then although U and V can be random, the three lemmas can be first proved conditional on U and V and since the bounds for U and V in Assumption 2.3.1 are uniform, the conclusions hold unconditionally.

Proof of a)

From (2.3.6) and the expression of X_j, $Q_j = Q_{j-1} + q^{-1}\mathcal{P}_{\Omega_j} X_{j-1}, \forall j \geq 1$. Therefore, $Q_{j_0} = \sum_{j=1} q^{-1}\mathcal{P}_{\Omega_j} X_{j-1}$ with $Q_0 = 0$. Then

$$\begin{aligned}
||W_L|| = ||\mathcal{P}_{\Phi^\perp} Q_{j_0}|| &\leq \sum_j ||q^{-1}\mathcal{P}_{\Phi^\perp}\mathcal{P}_{\Omega_j} X_{j-1}|| \\
&= \sum_{j=1} ||\mathcal{P}_{\Phi^\perp}(q^{-1}\mathcal{P}_{\Omega_j} X_{j-1} - X_{j-1})|| \\
&\leq \sum_{j=1} ||q^{-1}\mathcal{P}_{\Omega_j} X_{j-1} - X_{j-1}|| \\
&\leq C\sqrt{\frac{N\log N}{q}} \sum_{j=1} ||X_{j-1}||_\infty \\
&\leq C'\sqrt{\frac{N\log N}{q}} ||UV^*||_\infty
\end{aligned}$$

The second line follows from $\mathcal{P}_{\Phi^\perp} X_{j-1} = 0$. The second inequality follows from $||\mathcal{P}_{\Phi^\perp}|| \leq 1$. The third inequality follows from Lemma B.1.3 by setting $\delta_0 = q$. Similarly, the last inequality holds by noticing $||X_j||_\infty = ||(\mathcal{P}_\Phi - q^{-1}\mathcal{P}_\Phi\mathcal{P}_{\Omega_j}\mathcal{P}_\Phi)X_{j-1}||_\infty \leq ||X_{j-1}||_\infty$ by Lemma B.1.2, and thus $||X_j||_\infty \leq \varepsilon^j ||UV^*||_\infty$. To see q indeed satisfies the conditions in Lemma B.1.2 and Lemma B.1.3, note that since $j_0 = 4$, $(1-q)^4 = 1 - \delta$, so

$$q = \frac{\delta}{(1 + (1-\delta)^{\frac{1}{2}})(1 + (1-\delta)^{\frac{1}{4}})} \in [\frac{\delta}{4}, \delta]$$

Therefore, q and δ are of the same order, and thus under the conditions in Lemma 2.3.2, the conditions in Lemma B.1.2 and Lemma B.1.3 are satisfied. Then the right-hand side of the last inequality is of the order $\frac{\log N\sqrt{\mu r}}{N^{1/3}} \to 0$.

Proof of b)

The right-hand side $\frac{\lambda\delta}{16}$ is of the order of $\frac{\mu^{4/3} r}{N}$.

For the left-hand side, since $\Omega^c = \cup_{j=1}^{j_0} \Omega_j$, $||\mathcal{P}_\Omega Q_{j_0}|| = 0$. Then

$$\mathcal{P}_\Omega(UV^* + W_L) = \mathcal{P}_\Omega(UV^* + \mathcal{P}_{\Phi^\perp} Q_{j_0}) = \mathcal{P}_\Omega(UV^* - \mathcal{P}_\Phi Q_{j_0}) = \mathcal{P}_\Omega(X_{j_0}) \tag{B.1.1}$$

Recall $X_j = (\mathcal{P}_\Phi - q^{-1}\mathcal{P}_\Phi \mathcal{P}_{\Omega_j} \mathcal{P}_\Phi) X_{j-1}$. So by Lemma B.1.1, $||X_j||_F \leq \varepsilon ||X_{j-1}||_F$, where ε is the same as the one in the condition in Lemma 2.3.2 since the orders of q and δ are equal. Therefore, $||X_{j_0}||_F \leq \varepsilon^{j_0} ||UV^*||_F = \varepsilon^4 \sqrt{r} = O(\frac{(\log N)^4 r^{1/2}}{N^{4/3}})$ which is smaller than $\frac{\mu^{4/3} r}{N}$ with large enough N.

Proof of c)

Since $UV^* + W_L = X_{j_0} + Q_{j_0}$ and Q_{j_0} is supported on Ω^c,

$$\begin{aligned} ||\mathcal{P}_{\Omega^\perp}(UV^* + W_L)||_\infty =& ||\mathcal{P}_{\Omega^\perp}(X_{j_0} + Q_{j_0})||_\infty \\ =& ||\mathcal{P}_{\Omega^\perp} X_{j_0} + Q_{j_0}||_\infty \\ \leq& ||X_{j_0}||_F + ||Q_{j_0}||_\infty \end{aligned}$$

From b) we already have $||X_{j_0}||_F \leq \lambda/8$ for large enough N because $\lambda \asymp \frac{\mu^{1/3}}{N^{2/3}}$, and thus,

$$\begin{aligned} ||Q_{j_0}||_\infty &= ||\sum_{j=1} q^{-1} \mathcal{P}_{\Omega_j} X_{j-1}||_\infty \\ &\leq q^{-1} \sum_{j=1} ||X_{j-1}||_\infty \\ &\leq C q^{-1} ||UV||_\infty \\ &= O(\frac{1}{N^{2/3}}) \end{aligned}$$

where the second inequality follows from $||X_j||_\infty \leq \varepsilon^j ||UV^*||_\infty$ which we derived in the proof of a). It is clear that the right hand side of the inequality is smaller than $\lambda/8$ for large enough N.

Proof of d)

By construction of the Bernoulli device, for any element E_{it}, $P(E_{it} = 0) = \delta$ and $P(E_{it} = 1) = P(E_{it} = -1) = \frac{1-\delta}{2}$. Also, conditional on Ω, the signs of E are i.i.d. symmetric. By the definition of W_S,

$$W_S = \lambda \mathcal{P}_{\Phi^\perp} E + \lambda \mathcal{P}_{\Phi^\perp} \sum_{k \geq 1} (\mathcal{P}_\Omega \mathcal{P}_\Phi \mathcal{P}_\Omega)^k E$$

For the first term, $||\lambda \mathcal{P}_{\Phi^\perp} E|| \leq \lambda ||E|| \leq \lambda C \sqrt{N}$ with large probability because E has i.i.d. and mean 0 entries (see Candès et al. (2011) and the reference therein).

For the second term, denote $\mathcal{R} = \mathcal{P}_{\Phi^\perp} \sum_{k \geq 1} (\mathcal{P}_\Omega \mathcal{P}_\Phi \mathcal{P}_\Omega)^k$. Then under the event $\{||\mathcal{P}_\Omega \mathcal{P}_\Phi||^2 \leq 1 - \delta + \varepsilon\delta\} \cap \{||\mathcal{P}_{\Omega^\perp} \mathcal{P}_\Phi||^2 \leq \delta + \varepsilon(1-\delta)\}$ which occurs with high probability under Lemma 2.3.1, we have

$$\begin{aligned}
||R|| &= ||\mathcal{P}_{\Phi^\perp} \sum_{k \geq 1} (\mathcal{P}_\Omega \mathcal{P}_\Phi \mathcal{P}_\Omega)^k|| \\
&\leq ||\mathcal{P}_{\Phi^\perp} \mathcal{P}_\Omega \mathcal{P}_\Phi \mathcal{P}_\Omega|| \cdot \sum_{k \geq 0} ||(\mathcal{P}_\Omega \mathcal{P}_\Phi \mathcal{P}_\Omega)^k|| \\
&\leq ||\mathcal{P}_{\Phi^\perp} \mathcal{P}_\Omega \mathcal{P}_\Phi|| \cdot ||\mathcal{P}_\Omega|| \cdot \sum_{k \geq 0} ||(\mathcal{P}_\Omega \mathcal{P}_\Phi)||^{2k} \\
&= ||\mathcal{P}_{\Phi^\perp} \mathcal{P}_{\Omega^\perp} \mathcal{P}_\Phi|| \cdot ||\mathcal{P}_\Omega|| \cdot \sum_{k \geq 0} ||(\mathcal{P}_\Omega \mathcal{P}_\Phi)||^{2k} \\
&\leq \frac{||\mathcal{P}_{\Omega^\perp} \mathcal{P}_\Phi||}{1 - ||(\mathcal{P}_\Omega \mathcal{P}_\Phi)||^2} \\
&\leq \frac{\sqrt{\delta + \varepsilon(1-\delta)}}{\delta(1-\varepsilon)} \\
&\leq \frac{2}{\sqrt{\delta}}
\end{aligned}$$

where the last inequality is from the fact that for large enough N and T, $\varepsilon < \delta$ and $1 - \varepsilon$ is arbitrarily close to 1.

For any $\tau \in (0,1)$, let N_τ and T_τ denote an τ-net for $\mathbb{S}^{N-1}$ and $\mathbb{S}^{T-1}$, which are $N-1$

and $T-1$ dimensional unit sphere respectively. The sizes of the spheres are at most $(3/\tau)^N$ and $(3/\tau)^T$ (see Ganesh et al. (2010)). Then

$$||\mathcal{R}E|| = \sup_{x\in\mathbb{S}^{N-1},y\in\mathbb{S}^{T-1}} \langle y,(\mathcal{R}E)'x\rangle \leq (1-\tau)^{-2} \sup_{x\in N_\tau,y\in T_\tau} \langle y,(\mathcal{R}E)'x\rangle$$

For a fixed pair $(x,y) \in N_\tau \times T_\tau$, define $X(x,y) \equiv \langle y,(\mathcal{R}E)'x\rangle = \langle \mathcal{R}xy', E\rangle$. Then conditional on Ω, U and V, by Hoeffding's inequality,

$$\begin{aligned} P\big(|X(x,y)| > t\big|\Omega,U,V\big) \leq& 2\exp\big(-\frac{2t^2}{||\mathcal{R}(xy')||_F^2}\big) \\ \leq& 2\exp\big(-\frac{2t^2}{||\mathcal{R}||^2}\big) \\ \leq& 2\exp\big(-\frac{\delta t^2}{2}\big) \end{aligned}$$

Then using the union bound,

$$P\big(\sup_{x\in N_\tau,y\in T_\tau} |X(x,y)| > t\big|\Omega,U,V\big) \leq 2(\frac{3}{\tau})^{N+T}\exp(-\frac{\delta t^2}{2})$$

Therefore,

$$\begin{aligned} P\big(||\mathcal{R}E|| > C\sqrt{\frac{N}{\delta}}\big|\Omega,U,V\big) \leq& 2(\frac{3}{\tau})^{N+T}\exp(-\frac{\delta C^2(1-\tau)^4N}{2\delta}) \\ =& 2(\frac{3}{\tau})^{N+T}\exp(-\frac{C^2(1-\tau)^4N}{2}) \end{aligned}$$

Since $N \asymp T$, for any τ, there exists a finite C such that $\big(\frac{3}{\tau}\big)^2 < \exp(\frac{C^2(1-\tau)^4}{2})$. Also as the probability bound is not a function of Ω, U or V, the inequality holds unconditionally. Hence with high probability,

$$||W_S|| \leq C\lambda\sqrt{N}(1+\frac{1}{\sqrt{\delta}}) \leq C'\mu^{-1/6}r^{-1/2} \to 0$$

Proof of e)

By construction $\mathcal{P}_{\Omega^\perp} E = 0$ and $\mathcal{P}_{\Omega^\perp}(\mathcal{P}_\Omega \mathcal{P}_\Phi \mathcal{P}_\Omega)^k E = 0$ for all k, so we have

$$\begin{aligned}\mathcal{P}_{\Omega^\perp} W_S =& \lambda \mathcal{P}_{\Omega^\perp} \mathcal{P}_{\Phi^\perp} \sum_{k\geq 0} (\mathcal{P}_\Omega \mathcal{P}_\Phi \mathcal{P}_\Omega)^k E \\ =& -\lambda \mathcal{P}_{\Omega^\perp} \mathcal{P}_\Phi \sum_{k\geq 0} (\mathcal{P}_\Omega \mathcal{P}_\Phi \mathcal{P}_\Omega)^k E \\ =& -\lambda \mathcal{P}_{\Omega^\perp} \mathcal{P}_\Phi (\mathcal{P}_\Omega - \mathcal{P}_\Omega \mathcal{P}_\Phi \mathcal{P}_\Omega)^{-1} E\end{aligned}$$

Therefore, $||\mathcal{P}_{\Omega^\perp} W_S||_\infty \leq \lambda ||\mathcal{P}_\Phi (\mathcal{P}_\Omega - \mathcal{P}_\Omega \mathcal{P}_\Phi \mathcal{P}_\Omega)^{-1} E||_\infty$.

Let $\tilde{W}_S \equiv \mathcal{P}_\Phi (\mathcal{P}_\Omega - \mathcal{P}_\Omega \mathcal{P}_\Phi \mathcal{P}_\Omega)^{-1} E$. Then $\tilde{W}_{S,it} = \langle e_i e_t', \tilde{W}_S \rangle$. Let $X(i,t) = (\mathcal{P}_\Omega - \mathcal{P}_\Omega \mathcal{P}_\Phi \mathcal{P}_\Omega)^{-1} \mathcal{P}_\Omega \mathcal{P}_\Phi (e_i e_t')$, then $\langle e_i e_t', \tilde{W}_S \rangle = \langle X(i,t), E \rangle$ since $\mathcal{P}_\Omega (\mathcal{P}_\Omega - \mathcal{P}_\Omega \mathcal{P}_\Phi \mathcal{P}_\Omega)^{-1} = (\mathcal{P}_\Omega - \mathcal{P}_\Omega \mathcal{P}_\Phi \mathcal{P}_\Omega)^{-1}$. Therefore, by Hoeffding's inequality, conditional on Ω, U and V,

$$P\big(|\tilde{W}_{S,it}| > t \big| \Omega, U, V\big) \leq 2 \exp\Big(-\frac{2t^2}{||X(i,t)||_F^2}\Big)$$

and by using the union bound,

$$P\big(||\tilde{W}_S||_\infty > t \big| \Omega, U, V\big) \leq 2NT \exp\Big(-\frac{2t^2}{||X(i,t)||_F^2}\Big)$$

Under the conditions in Lemma 2.3.1 and the incoherence conditions,

$$||\mathcal{P}_\Omega \mathcal{P}_\Phi (e_i e_t')||_F \leq ||\mathcal{P}_\Omega \mathcal{P}_\Phi|| \cdot ||\mathcal{P}_\Phi (e_i e_t')||_F \leq \sqrt{1 - \delta + \varepsilon\delta} \sqrt{C\mu r / N}$$

Meanwhile, from the proof of d), we know $||(\mathcal{P}_\Omega - \mathcal{P}_\Omega\mathcal{P}_\Phi\mathcal{P}_\Omega)^{-1}|| \leq \frac{1}{\delta(1-\varepsilon)}$. Therefore,

$$\begin{aligned} P\big(||W_S||_\infty > \lambda t|\Omega, U, V\big) \leq & P\big(||\tilde{W}_{S,it}||_\infty > t\big|\Omega, U, V\big) \\ \leq & 2NT\exp\big(-\frac{C't^2N\delta^2(1-\varepsilon)^2}{\mu r(1-\delta+\varepsilon\delta)}\big) \\ \to & 0, \forall t > 0 \end{aligned}$$

Since the right-hand side does not depend on Ω, U and V, the inequality holds unconditionally, which completes the proof.

Appendix C

Appendix to Chapter 3

C.1 Proof of Lemma 3.3.1

Let $\nabla\boldsymbol{\rho}_u(\cdot)$ be the subgradiant of $\boldsymbol{\rho}_u$ evaluated at $\cdot$, then by definition, $\nabla\boldsymbol{\rho}_u(V(u))$ is an $N \times T$ matrix of which the (i,t)-th element is

$$\left(\nabla\boldsymbol{\rho}_u\left(V(u)\right)\right)_{it} = u\mathbb{1}_{V_{it}(u)>0} + (u-1)\mathbb{1}_{V_{it}(u)<0}.$$

By Assumption 3.3.1, $\nabla\boldsymbol{\rho}_u(V(u))$ is thus a random matrix with i.i.d. mean 0 entries bounded by $u \vee (1-u)$ conditional on $(X_1, ..., X_p)$. We have the following lemma for $\nabla\boldsymbol{\rho}_u(V)$:

Lemma C.1.1. *Under Assumption 3.3.1, there exist two constants $C_1, C_2 > 0$ such that the following inequalities hold with high probabilities:*

$$\sup_{u\in\mathcal{U}} \max_{1\le j\le p} \left|\left\langle \nabla\boldsymbol{\rho}_u(V(u)), X_j\right\rangle\right| \le C_1\sqrt{NT\log(NT)}, \tag{C.1.1}$$

$$\sup_{u\in\mathcal{U}} \max_{1\le j\le p} ||\nabla\boldsymbol{\rho}_u(V(u))|| \le C_2\sqrt{N\vee T} \tag{C.1.2}$$

Proof. See Appendix C.3. □

In what follows, the derivation is under the event that inequalities (C.1.1) and (C.1.2) hold and conditional on $(X_1, ..., X_p)$. By the definition of $(\hat{\beta}(u), \hat{L}(u))$, the following holds uniformly in $u \in \mathcal{U}$:

$$\frac{1}{NT}\left[\boldsymbol{\rho}_u(Y - \sum_{j=1}^{p} X_j\hat{\beta}(u) - \hat{L}(u)) - \boldsymbol{\rho}_u(Y - \sum_{j=1}^{p} X_j\beta_0(u) - L_0(u))\right] + \lambda\left[||\hat{L}(u)||_* - ||L_0(u)||_*\right] \leq 0$$

$$\implies \frac{1}{NT}\left[\boldsymbol{\rho}_u(V(u) - \sum_{j=1}^{p} X_j\hat{\Delta}_{\beta,j}(u) - \hat{\Delta}_L(u)) - \boldsymbol{\rho}_u(V(u))\right] + \lambda\left[||\hat{L}(u)||_* - ||L_0(u)||_*\right] \leq 0$$

Let us first consider $\frac{1}{NT}\left[\boldsymbol{\rho}_u\left(V(u) - \sum_{j=1}^{p} X_j\hat{\Delta}_{\beta,j}(u) - \hat{\Delta}_L(u)\right) - \boldsymbol{\rho}_u(V(u))\right]$. By the definition of the subgradient,

$$\begin{aligned}
&\frac{1}{NT}\left[\boldsymbol{\rho}_u\left(V(u) - \sum_{j=1}^{p} X_j\hat{\Delta}_{\beta,j}(u) - \hat{\Delta}_L(u)\right) - \boldsymbol{\rho}_u(V(u))\right] \\
\geq& -\frac{1}{NT}\left|\left\langle \nabla\boldsymbol{\rho}_u(V(u)), \sum_{j=1}^{p} X_j\hat{\Delta}_{\beta,j}(u) + \hat{\Delta}_L(u)\right\rangle\right| \\
\geq& -\frac{1}{NT}||\hat{\Delta}_\beta(u)||_1 \max_{1\leq j\leq p}\left|\left\langle \nabla\boldsymbol{\rho}_u(V(u)), X_j\right\rangle\right| - \frac{1}{NT}||\nabla\boldsymbol{\rho}_u(V(u))|| \cdot ||\hat{\Delta}_L(u)||_* \\
\geq& -C_1\sqrt{\frac{\log(NT)}{NT}}||\hat{\Delta}_\beta(u)||_1 - \frac{C_2\sqrt{N \vee T}}{NT}||\hat{\Delta}_L(u)||_*
\end{aligned}$$

The first term in the third line is elementary. The second term is from Lemma 3.2 in Candès and Recht (2009) which says for any two matrices A and B of the same size, $|\langle A, B\rangle| \leq ||A|| \cdot ||B||_*$. The last inequality is from equations (C.1.1) and (C.1.2) in Lemma C.1.1.

Next, consider $\lambda\left[||\hat{L}(u)||_* - ||L_0(u)||_*\right]$. By construction, $P_{\Phi^\perp(u)}L_0(u) = 0$. Hence,

$$\begin{aligned}
||\hat{L}(u)||_* - ||L_0(u)||_* =& ||P_{\Phi(u)}L_0(u) + P_{\Phi(u)}\hat{\Delta}_L(u)||_* + ||P_{\Phi^\perp(u)}\hat{\Delta}_L(u)||_* - ||P_{\Phi(u)}L_0(u)||_* \\
\geq& ||P_{\Phi^\perp(u)}\hat{\Delta}_L(u)||_* - ||P_{\Phi(u)}\hat{\Delta}_L(u)||_*
\end{aligned}$$

Combining the two pieces with rearrangement, we have shown that the following

inequality holds uniformly in $u \in \mathcal{U}$ conditional on $(X_1, ..., X_p)$:

$$(\lambda - \frac{C_2\sqrt{N \vee T}}{NT})||P_{\Phi^{\perp}(u)}\hat{\Delta}_L(u)||_* \leq C_1\sqrt{\frac{p\log(NT)}{NT}}||\hat{\Delta}_\beta(u)||_F + (\lambda + \frac{C_2\sqrt{N \vee T}}{NT})||P_{\Phi(u)}\hat{\Delta}_L(u)||_*$$

Then by Lemma C.1.1 and integrating out $(X_1, ..., X_p)$, we obtain the desired result.

C.2 Proof of Theorem 3.4.1

Throughout, the analysis is under the event that $(\hat{\Delta}_\beta(u), \hat{\Delta}_L(u)) \in \mathcal{R}_u$ uniformly in $u \in \mathcal{U}$ and $\max_{j=1,...,p} ||X_j||_\infty < \frac{c\sqrt{N \wedge T}}{\alpha\sqrt{\log(NT)}}$ for any fixed $c > 0$. Both events hold with high probability because of Lemma 3.3.1 and Assumption 3.4.2. To prove the theorem, we want to show the following is impossible with high probability:

$$\exists u \in \mathcal{U} : ||\hat{\Delta}_\beta(u)||_F^2 + \frac{1}{NT}||\hat{\Delta}_L(u)||_F^2 \geq t^2$$

for some t^2 is of the order of the right hand side of inequality (3.4.1).

Under Assumption 3.4.2 and by the definition of the estimator, $||\hat{\Delta}_L(u)||_\infty \leq 2\alpha$ a.s. by the triangle inequality. Let $\mathcal{D} \equiv \{(\Delta_\beta, \Delta_L) : \Delta_\beta \in \mathbb{R}^p, \Delta_L \in \mathbb{R}^{N\times T}, ||\Delta_L||_\infty \leq 2\alpha\}$. Then by convexity of the objective function and the definition of the estimator, the inequality above implies that

$$\begin{aligned} 0 \geq \min_{\substack{(\Delta_\beta,\Delta_L)\in\mathcal{R}_u\cap\mathcal{D} \\ ||\Delta_\beta||_F^2+\frac{1}{NT}||\Delta_L||_F^2\geq t^2}} & \frac{1}{NT}\left[\boldsymbol{\rho}_u\left(V(u) - \sum_{j=1}^p X_j\Delta_{\beta,j} - \Delta_L\right) - \boldsymbol{\rho}_u(V(u))\right] \\ & + \lambda\left[||L_0(u) + \Delta_L||_* - ||L_0(u)||_*\right] \end{aligned}$$

The proof consists of two main steps. In the first step (Lemmas C.2.1 and C.2.3), I bound the norm of $\sum_{j=1}^p X_j\Delta_{\beta,j} + \Delta_L$. In the second step (Lemma C.2.4), I separate $\Delta_{\beta,j}$ and Δ_L by invoking Assumption 3.4.4. The first step is adapted from the proofs of Theorem 2 in Belloni and Chernozhukov (2011) and Theorem 3.2 in Chao, Härdle and

Yuan (2019). A new theoretical challenge arises in the minoration step (Lemma C.2.1) because of the high dimensional object Δ_L. I develop a new argument to handle it. Lemma C.2.3 follows the two papers cited closely where I only highlight the differences that Δ_L brings into.

Since $\mathcal{R}_u$ is a cone and $\mathcal{D}$ is convex, by convexity of the objective function, the inequality sign in the constraint can be replaced with equality:

$$0 \geq \min_{\substack{(\Delta_\beta,\Delta_L)\in\mathcal{R}_u\cap\mathcal{D} \\ ||\Delta_\beta||_F^2+\frac{1}{NT}||\Delta_L||_F^2=t^2}} \frac{1}{NT}\left[\boldsymbol{\rho}_u\left(V(u)-\sum_{j=1}^{p}X_j\Delta_{\beta,j}-\Delta_L\right)-\boldsymbol{\rho}_u(V(u))\right] + \lambda\left[||L_0(u)+\Delta_L||_*-||L_0||_*\right]$$

Let us rewrite the minimand as follows:

$$\begin{aligned}
&\frac{1}{NT}\left[\boldsymbol{\rho}_u\left(V(u)-\sum_{j=1}^{p}X_j\Delta_{\beta,j}-\Delta_L\right)-\boldsymbol{\rho}_u(V(u))\right]+\lambda\left[||L_0(u)+\Delta_L||_*-||L_0(u)||_*\right] \\
=&\frac{1}{NT}\mathbb{E}\left[\boldsymbol{\rho}_u\left(V(u)-\sum_{j=1}^{p}X_j\Delta_{\beta,j}-\Delta_L\right)-\boldsymbol{\rho}_u(V(u))\right] \\
&+\frac{1}{\sqrt{NT}}\mathbb{G}_{NT}\left(\rho_u\left(V_{it}(u)-X_{it}'\Delta_\beta-\Delta_{L,it}\right)-\rho_u(V_{it}(u))\right) \\
&+\lambda\left[||L_0(u)+\Delta_L||_*-||L_0(u)||_*\right]
\end{aligned}$$

Lemmas C.2.1 and C.2.3 provide bounds for the first two terms on the right hand side respectively.

Lemma C.2.1 (Minoration). *Under Assumptions 3.3.1-3.4.2, there exists a constant $c>0$ such that the following holds uniformly in u.*

$$\frac{1}{NT}\mathbb{E}\left[\boldsymbol{\rho}_u\left(V(u)-\sum_{j=1}^{p}X_j\Delta_{\beta,j}-\Delta_L\right)-\boldsymbol{\rho}_u(V(u))\right] \geq \frac{c\underline{f}^4}{(\bar{\alpha}\bar{f}')^2NT}\mathbb{E}||X_{it}'\Delta_\beta+\Delta_{L,it}||_F^2 \quad \text{(C.2.1)}$$

To prove Lemma C.2.1, I need the following result which will be used to handle the

high dimensional object Δ_L.

Lemma C.2.2. *A2 For all $w_1, w_2 \in \mathbb{R}$ and all $\kappa \in (0,1)$,*

$$\int_0^{w_2} \big(\mathbb{1}(w_1 \leq z) - \mathbb{1}(w_1 \leq 0)\big)dz \geq \int_0^{\kappa w_2} \big(\mathbb{1}(w_1 \leq z) - \mathbb{1}(w_1 \leq 0)\big)dz \geq 0$$

Proof. See Appendix C.3. □

Proof of Lemma C.2.1. By the Knight's identity (Knight, 1998), for any two scalars w_1 and w_2,

$$\rho_u(w_1 - w_2) - \rho_u(w_1) = -w_2(u - \mathbb{1}_{w_1 \leq 0}) + \int_0^{w_2} (\mathbb{1}_{w_1 \leq s} - \mathbb{1}_{w_1 \leq 0})ds$$

Let $w_1 = V_{it}(u)$ and $w_2 = X_{it}'\Delta_\beta + \Delta_{L,it}$, then by construction $\mathbb{E}(-w_2(u - \mathbb{1}_{w_1 \leq 0})) = 0$. Also, by Lemma C.2.2, with probability 1,

$$\begin{aligned}
&\int_0^{X_{it}'\Delta_\beta + \Delta_{L,it}} \big(\mathbb{1}(V_{it}(u) \leq s) - \mathbb{1}(V_{it}(u) \leq 0)\big)ds \\
\geq &\int_0^{\kappa(X_{it}'\Delta_\beta + \Delta_{L,it})} \big(\mathbb{1}(V_{it}(u) \leq s) - \mathbb{1}(V_{it}(u) \leq 0)\big)ds
\end{aligned}$$

where $\kappa \equiv \frac{3\underline{f}^2}{8\alpha\bar{f}'} \in (0,1)$ for large N and T. Then by the law of iterated expectation and mean value theorem,

$$\begin{aligned}
&\mathbb{E}\int_0^{X_{it}'\Delta_\beta + \Delta_{L,it}} \big(\mathbb{1}(V_{it}(u) \leq s) - \mathbb{1}(V_{it}(u) \leq 0)\big)ds \\
\geq &\mathbb{E}\int_0^{\kappa(X_{it}'\Delta_\beta + \Delta_{L,it})} \big(\mathbb{1}(V_{it}(u) \leq s) - \mathbb{1}(V_{it}(u) \leq 0)\big)ds \\
\geq &\mathbb{E}\int_0^{\kappa(X_{it}'\Delta_\beta + \Delta_{L,it})} \big(F_{V_{it}(u)|X_{it}}(s) - F_{V_{it}(u)|X_{it}}(0)\big)\big)ds \\
= &\mathbb{E}\int_0^{\kappa(X_{it}'\Delta_\beta + \Delta_{L,it})} \big(sf_{V_{it}(u)|X_{it}}(0) + \frac{s^2}{2}f'_{V_{it}(u)|X_{it}}(\tilde{s})\big)\big)ds \\
\geq &\frac{\kappa^2\underline{f}^2}{4}\mathbb{E}\big(X_{it}'\Delta_\beta + \Delta_{L,it}\big)^2 + \mathbb{E}\big[\frac{\kappa^2\underline{f}^2}{4}\big(X_{it}'\Delta_\beta + \Delta_{L,it}\big)^2\big(1 - |\frac{2\kappa\bar{f}'}{3\underline{f}^2}(X_{it}'\Delta_\beta + \Delta_{L,it}|)\big)\big] \\
\geq &\frac{\kappa^2\underline{f}^2}{4}\mathbb{E}\big(X_{it}'\Delta_\beta + \Delta_{L,it}\big)^2
\end{aligned}$$

where the third line is from the law of iterated expectation. The last inequality holds because under the choice of κ and t, $1 - \left|\frac{2\kappa\bar{f}'}{3\underline{f}^2}\left(X_{it}'\Delta_\beta + \Delta_{L,it}\right)\right| > 0$ given the upper bound of the magnitude of $||X_j||_\infty$. □

Remark C.2.1. As mentioned, Δ_L introduces new difficulties for minoration. Specifically, $||\Delta_L||_F^2$ can be greater than $\sum_{i,t}|\Delta_{L,it}|^3$ even in the restricted set $\mathcal{R}_u$ for $||\Delta_L||_F^2 = NTt^2$. As a consequence, standard argument fails because after Taylor expansion, the higher order term may dominate the leading term, resulting in a negative lower bound for the expectation under investigation. I overcome this difficulty by exploiting monotonicity in the integral in the Knight's identity and imposing conditions that directly restrict the tail behavior of $X_{j,it}$ and the magnitude of $L_{0,it}$.

Lemma C.2.3 (Bound on the Empirical Process)**.** *Under Assumptions 3.3.1, 3.4.1 and 3.4.3, there exists a constant $C_0 > 0$ such that with high probability*

$$\begin{aligned}&\sup_{\substack{u\in\mathcal{U}\\(\Delta_\beta,\Delta_L)\in\mathcal{R}_u\cap\mathcal{D}\\||\Delta_\beta||_F^2+\frac{1}{NT}||\Delta_L||_F^2=t^2}}\left|\mathbb{G}_{NT}\left(\rho_u\left(V_{it}(u)-X_{it}'\Delta_\beta-\Delta_{L,it}\right)-\rho_u(V_{it}(u))\right)\right|\\ \leq& C_0\left(\sqrt{p\log(NT)}+\sqrt{N\vee T}\sqrt{\bar{r}}\right)\sqrt{\log(NT)}t\end{aligned}$$

Proof of Lemma C.2.3. Note that the check function is a contraction. Hence, similar to Belloni and Chernozhukov (2011) and Chao, Härdle and Yuan (2019), there exists $C > 0$

such that

$$
\begin{aligned}
&\mathrm{Var}\big(\mathbb{G}_{NT}\big(\rho_u\big(V_{it}(u) - X_{it}'\Delta_\beta - \Delta_{L,it}\big) - \rho_u(V_{it}(u))\big)\big) \\
\leq &\frac{1}{NT}\sum_{i,t}\mathbb{E}(X_{it}'\Delta_\beta + \Delta_{L,it})^2 \\
\leq &\frac{2}{NT}\sum_{i,t}\mathbb{E}(X_{it}'\Delta_\beta)^2 + \frac{2}{NT}||\Delta_L||_F^2 \\
\leq &C\big(||\Delta_\beta||_F^2 + \frac{1}{NT}||\Delta_L||_F^2\big) \\
\leq &Ct^2
\end{aligned}
$$

where the third inequality follows from $\mathbb{E}(X_{it}'\Delta_\beta)^2 = \Delta_\beta'\mathbb{E}(X_{it}X_{it}')\Delta_\beta \leq C'\sigma_{\max}^2||\Delta_\beta||_F^2$ for some $C' > 0$, where $\sigma_{\max}^2$ is the largest eigenvalue of $\mathbb{E}(X_{it}X_{it}')$. Let $\mathcal{A}(t)$ denote the empirical process under investigation. Then by Lemma 2.3.7 in van der Vaart and Wellner (1996), let $s \geq 4t$, we have

$$
\mathbb{P}(\mathcal{A}(t) > s) \leq C''\mathbb{P}(\mathcal{A}^0(t) > \frac{s}{4})
$$

for some $C'' > 0$ where $\mathcal{A}^0(t)$ is the symmetrized version of $\mathcal{A}(t)$ by replacing $\mathbb{G}_{NT}$ with the symmetrized version $\mathbb{G}_{NT}^0$.

Consider the random variable $\rho_u\big(V_{it}(u) - X_{it}'\Delta_\beta - \Delta_{L,it}\big) - \rho_u(V_{it}(u))$:

$$
\rho_u\big(V_{it}(u) - X_{it}'\Delta_\beta - \Delta_{L,it}\big) - \rho_u(V_{it}(u)) = u(X_{it}'\Delta_\beta + \Delta_{L,it}) + \delta_{it}\big(X_{it}'\Delta_\beta + \Delta_{L,it}, u\big)
$$

where $\delta_{it}\big(X_{it}'\Delta_\beta + \Delta_{L,it}, u\big) = \big(V_{it}(u) - X_{it}'\Delta_\beta - \Delta_{L,it}\big)_- - \big(V_{it}(u)\big)_-$. Let

$$
\mathcal{B}_1^0(t) \equiv \sup_{\substack{u\in\mathcal{U} \\ (\Delta_\beta,\Delta_L)\in\mathcal{R}_u\cap\mathcal{D} \\ ||\Delta_\beta||_F^2\leq t^2}} \big|\mathbb{G}^0\big(X_{it}'\Delta_\beta\big)\big|,
$$

$$\mathcal{B}_2^0(t) \equiv \sup_{\substack{u\in\mathcal{U} \\ (\Delta_\beta,\Delta_L)\in\mathcal{R}_u\cap\mathcal{D} \\ ||\Delta_\beta||_F^2+\frac{1}{NT}||\Delta_L||_F^2\leq t^2}} \left|\mathbb{G}^0\left(\Delta_{L,it}\right)\right|,$$

and

$$\mathcal{C}^0(t) \equiv \sup_{\substack{u\in\mathcal{U} \\ (\Delta_\beta,\Delta_L)\in\mathcal{R}_u\cap\mathcal{D} \\ ||\Delta_\beta||_F^2+\frac{1}{NT}||\Delta_L||_F^2\leq t^2}} \left|\mathbb{G}^0\left(\delta_{it}\left(X_{it}'\Delta_\beta+\Delta_{L,it},u\right)\right)\right|,$$

then $\mathcal{A}^0(t) \leq \mathcal{B}_1^0(t)+\mathcal{B}_2^0(t)+\mathcal{C}^0(t)$. Next I bound $\mathcal{B}_1^0(t)$, $\mathcal{B}_2^0(t)$ and $\mathcal{C}^0(t)$ respectively.

Bound on $\mathcal{B}_1^0(t)$. Since $\mathcal{B}_1^0(t)$ does not contain Δ_L, the bound is identical to that in Belloni and Chernozhukov (2011), i.e., $\mathcal{B}_1^0(t) \leq C_1\sqrt{p\log(NT)}t$ with high probability.

Bound on $\mathcal{B}_2^0(t)$. Let $(\varepsilon_{it} : i \in \{1,..,N\}, t \in \{1,...,T\})$ be the Rademacher sequence in the symmetrized process. Let $\boldsymbol{\varepsilon}$ be the $N \times T$ matrix containing all the elements in the sequence. Then

$$\begin{aligned}
\mathcal{B}_2^0(t) \leq & \frac{1}{\sqrt{NT}} \sup_{\substack{u\in\mathcal{U} \\ (\Delta_\beta,\Delta_L)\in\mathcal{R}_u\cap\mathcal{D} \\ ||\Delta_\beta||_F^2+\frac{1}{NT}||\Delta_L||_F^2\leq t^2}} |\sum_{i,t}\varepsilon_{it}\Delta_{L,it}| \\
= & \frac{1}{\sqrt{NT}} \sup_{\substack{(\Delta_\beta,\Delta_L)\in\mathcal{R}_u\cap\mathcal{D} \\ ||\Delta_\beta||_F^2+\frac{1}{NT}||\Delta_L||_F^2\leq t^2}} |\langle\boldsymbol{\varepsilon},\Delta_L\rangle| \\
\leq & \frac{1}{\sqrt{NT}}||\boldsymbol{\varepsilon}||\cdot \sup_{\substack{(\Delta_\beta,\Delta_L)\in\mathcal{R}_u\cap\mathcal{D} \\ ||\Delta_\beta||_F^2+\frac{1}{NT}||\Delta_L||_F^2\leq t^2}} ||\Delta_L||_* \\
\leq & \frac{1}{\sqrt{NT}}||\boldsymbol{\varepsilon}||\cdot \sup_{\substack{(\Delta_\beta,\Delta_L)\in\mathcal{R}_u\cap\mathcal{D} \\ ||\Delta_\beta||_F^2+\frac{1}{NT}||\Delta_L||_F^2\leq t^2}} (||P_{\Phi(u)}\Delta_L||_* + ||P_{\Phi^\perp(u)}\Delta_L||_*) \\
\leq & \frac{1}{\sqrt{NT}}||\boldsymbol{\varepsilon}||\cdot \sup_{\substack{(\Delta_\beta,\Delta_L)\in\mathcal{R}_u\cap\mathcal{D} \\ ||\Delta_\beta||_F^2+\frac{1}{NT}||\Delta_L||_F^2\leq t^2}} \left(C_3\sqrt{p(N\wedge T)\log(NT)}||\Delta_\beta||_F + 4||P_{\Phi(u)}\Delta_L||_*\right) \\
\leq & \frac{1}{\sqrt{NT}}||\boldsymbol{\varepsilon}||\cdot \left(C_3\sqrt{p(N\wedge T)\log(NT)} + C_4\sqrt{NT\bar{r}}\right)t
\end{aligned}$$

where the second to the last inequality is from the definition of the restricted set $\mathcal{R}_u$.

Finally, since elements in ε are i.i.d. mean 0 and uniformly bounded in magnitude by 1, the operator norm is bounded by $C_5\sqrt{N \vee T}$ with high probability (Corollary 2.3.5 in Tao (2012)). Therefore,

$$\mathcal{B}_2^0(t) \leq \left(C_6\sqrt{p\log(NT)} + C_7\sqrt{N \vee T}\sqrt{\bar{r}}\right)t$$

with high probability.

Bound on $\mathcal{C}^0(t)$. Assumption 3.4.3 allows us to follow a similar ϵ-net argument in Belloni and Chernozhukov (2011) with necessary modifications made to accommodate the new term Δ_L. Let $\mathcal{U}_l = \{u_1, ..., u_l\}$ be an ϵ-net in $\mathcal{U}$ where $\epsilon \leq t$. By the triangle inequality, we have

$$\begin{aligned}
\mathcal{C}^0(t) \leq & \sup_{\substack{u\in\mathcal{U},|u-u_l|<\epsilon,u_l\in\mathcal{U}_l \\ (\Delta_\beta,\Delta_L)\in\mathcal{R}_u\cap\mathcal{D} \\ ||\Delta_\beta||_F^2+\frac{1}{NT}||\Delta_L||_F^2\leq t^2}} \left|\mathbb{G}_{NT}^0\left(\delta_{it}[X_{it}'(\Delta_\beta+\beta_0(u)-\beta_0(u_l))+\Delta_L+L_0(u)-L_0(u_l)]),u_l)\right)\right| \\
& + \sup_{\substack{u\in\mathcal{U},|u-u_l|<\epsilon,u_l\in\mathcal{U}_l \\ (\Delta_\beta,\Delta_L)\in\mathcal{R}_u\cap\mathcal{D} \\ ||\Delta_\beta||_F^2+\frac{1}{NT}||\Delta_L||_F^2\leq t^2}} \left|\mathbb{G}_{NT}^0\left(\delta_{it}[X_{it}'(\beta_0(u)-\beta_0(u_l))+L_0(u)-L_0(u_l),u_l]\right|\right| \\
\leq & 2\cdot \sup_{\substack{u_l\in\mathcal{U}_l \\ (\Delta_\beta,\Delta_L)\in\bar{\mathcal{R}}\cap\bar{\mathcal{D}} \\ ||\Delta_\beta||_F^2+\frac{1}{NT}||\Delta_L||_F^2\leq(\zeta_1^2+\zeta_2^2+1)t^2}} \left|\mathbb{G}_{NT}^0\left(\delta_{it}(X_{it}'\Delta_\beta+\Delta_L,u_l)\right)\right|
\end{aligned}$$

where the last inequality follows from Assumption 3.4.3 that $\sup_{|u-u_l|<\epsilon}||\beta_0(u)-\beta_0(u_l)||_F \leq \zeta_1\epsilon$ and $\frac{1}{\sqrt{NT}}\sup_{|u-u_l|<\epsilon}||L_0(u)-L_0(u_l)||_F \leq \zeta_2\epsilon$, and thus under the choice of ϵ, we can treat $\Delta_\beta+\beta_0(u)-\beta_0(u_l)$ and $\beta_0(u)-\beta_0(u_l)$ as new Δ_β. Similarly, I treat $\Delta_L + L_0(u) - L_0(u_l)$ and $L_0(u) - L_0(u_l)$ as new Δ_L by enlarging $\mathcal{R}_u$ and $\mathcal{D}$. $\bar{\mathcal{D}}$ contains all $||\Delta_L||_\infty \leq 4\alpha$. For $\bar{\mathcal{R}}$, note that $L_0(u_l)$ may not be in the space of $L_0(u)$, so it may be the case that $L_0(u) - L_0(u_l) \notin \mathcal{R}_u$. However, since

$$\text{rank}(L_0(u) - L_0(u_l)) \leq r(u) + r(u_l) \leq 2\bar{r},$$

$||L_0(u) - L_0(u_l)||_* \leq \sqrt{2\bar{r}}||L_0(u) - L_0(u_l)||_F \leq \sqrt{2\bar{r}}\zeta_2\sqrt{NT}t$ by Assumption 3.4.3. From the derivation for the bound on $\mathcal{B}_2^0(t)$, set $\bar{\mathcal{R}} = \{\Delta_L : ||L||_* \leq (C_3\sqrt{pNT\log(NT)} + (C_4 + \zeta_2)\sqrt{NT}\sqrt{2\bar{r}})t\}$.

Now by Markov inequality,

$$\begin{aligned}
&\mathbb{P}\big(\mathcal{C}^0(t) \geq \big(C_8\sqrt{p\log(NT)} + C_9\sqrt{N\vee T}\sqrt{\bar{r}}\big)\sqrt{\log(NT)t}\big) \\
\leq &\min_{\tau\geq 0} e^{-\tau(C_8\sqrt{p\log(NT)}+C_9\sqrt{N\vee T}\sqrt{\bar{r}})\log(NT)t}\mathbb{E}[e^{\tau\mathcal{C}^0(t)}]
\end{aligned}$$

By Theorem 4.12 of Ledoux and Talagrand (1991), contractivity of $\delta_{it}(\cdot)$ implies

$$\begin{aligned}
\mathbb{E}[e^{\tau\mathcal{C}^0(t)}] \leq &(1/\epsilon)\max_{u_l\in\mathcal{U}_l}\mathbb{E}\big[\exp\big(2\tau \sup_{\substack{(\Delta_\beta,\Delta_L)\in\bar{\mathcal{R}}\cap\bar{\mathcal{D}} \\ ||\Delta_\beta||_F^2+\frac{1}{NT}||\Delta_L||_F^2\leq(1+\zeta_1^2+\zeta_2^2)t^2}} \big|\mathbb{G}_{NT}^0\big(\delta_{it}(X_{it}'\Delta_\beta + \Delta_{L,it}, u_l)\big)\big|\big)\big] \\
\leq &(1/\epsilon)\mathbb{E}\big[\exp\big(4\tau \sup_{\substack{(\Delta_\beta,\Delta_L)\in\bar{\mathcal{R}}\cap\bar{\mathcal{D}} \\ ||\Delta_\beta||_F^2+\frac{1}{NT}||\Delta_L||_F^2\leq(1+\zeta_1^2+\zeta_2^2)t^2}} \big|\mathbb{G}_{NT}^0\big(X_{it}'\Delta_\beta + \Delta_{L,it}\big)\big|\big)\big] \\
\leq &(1/\epsilon)\mathbb{E}\big[\exp\big(4\tau \sup_{\substack{(\Delta_\beta,\Delta_L)\in\bar{\mathcal{R}}\cap\bar{\mathcal{D}} \\ ||\Delta_\beta||_F^2\leq(1+\zeta_1^2+\zeta_2^2)t^2}} \big|\mathbb{G}_{NT}^0\big(X_{it}'\Delta_\beta\big)\big| + \sup_{\substack{(\Delta_\beta,\Delta_L)\in\bar{\mathcal{R}}\cap\bar{\mathcal{D}} \\ \frac{1}{NT}||\Delta_L||_F^2\leq(1+\zeta_1^2+\zeta_2^2)t^2}} \big|\mathbb{G}_{NT}^0\Delta_{L,it}\big)\big|\big)\big]
\end{aligned}$$

Following exactly the same argument for $\mathcal{B}_1^0$ and $\mathcal{B}_2^0$, we obtain

$$\mathcal{C}^0(t) \leq \big(C_8\sqrt{p\log(NT)} + C_9\sqrt{N\vee T}\sqrt{\bar{r}}\big)\sqrt{\log(NT)t}$$

with high probability. □

Finally consider the difference in the penalty. From the derivation of the bound on $\mathcal{B}_2^0$, we have

$$\lambda\big|||L_0(u) + \Delta_L||_* - ||L_0(u)||_*\big| \leq \lambda||\Delta_L||_* \leq \lambda\big(C_3\sqrt{p(N\wedge T)\log(NT)} + C_4\sqrt{NT\bar{r}}\big)t$$

By the choice of λ, the right hand side is $(C_{10}\frac{\sqrt{p\log(NT)}}{\sqrt{NT}} + C_{11}\frac{\sqrt{\bar{r}}}{\sqrt{N\wedge T}})t$ for some $C_{10}, C_{11} > 0$.

Combining all the pieces together, we have

$$\min_{\substack{(\Delta_\beta,\Delta_L)\in\mathcal{R}_u\cap\mathcal{D}\\ ||\Delta_\beta||_F^2+\frac{1}{NT}||\Delta_L||_F^2=t^2}} \frac{1}{NT}\left[\boldsymbol{\rho}_u\left(V(u)-\sum_{j=1}^{p}X_j\Delta_{\beta,j}-\Delta_L\right)-\boldsymbol{\rho}_u(V(u))\right]+\lambda\left[||L_0(u)+\Delta_L||_*-||L_0(u)||_*\right]$$

$$\geq \frac{c\underline{f}^4}{(\alpha\bar{f}')^2NT}\mathbb{E}||\sum_{j=1}^{p}X\Delta_{\beta,j}+\Delta_L||_F^2-\bar{C}\left(\frac{\sqrt{p\log(NT)}}{\sqrt{NT}}+\frac{\sqrt{r}}{\sqrt{N\wedge T}}\right)\sqrt{\log(NT)}t$$

uniformly in u over $\mathcal{U}$ for some $\bar{C}>0$.

To obtain the result in Theorem 3.4.1, I need to separate the two terms in the the expectation $\mathbb{E}||\sum_{j=1}^{p}X\Delta_{\beta,j}+\Delta_L||_F^2$. This is guaranteed by Assumption 3.4.4.

Lemma C.2.4 (Separation). *Under Assumption 3.3.1 and 3.4.4, for any $(\Delta_\beta,\Delta_L)\in\mathbb{R}_u$ such that $||\Delta_\beta||_F^2+\frac{1}{NT}||\Delta_L||_F^2=t^2$, there exists $C>0$ such that*

$$\frac{1}{NT}\mathbb{E}||\sum_{j=1}^{p}X_j\Delta_{\beta,j}+\Delta_L||_F^2\geq Ct^2$$

Proof of Lemma C.2.4. Given $||\Delta_\beta||_F^2+\frac{1}{NT}||\Delta_L||_F^2=t^2$, let $||\Delta_\beta||_F^2=\gamma t^2$. Then $\frac{1}{NT}||\Delta_L||_F^2=(1-\gamma)t^2$. I am to express the expectation $\frac{1}{NT}\mathbb{E}||\sum_{j=1}^{p}X_j\Delta_{\beta,j}+\Delta_L||_F^2$ as a function of γ and show it is bounded away from a fixed fraction of t^2 uniformly in $\gamma\in[0,1]$ under Assumptions 3.3.1 and 3.4.4.

First consider $\mathbb{E}||\sum_{j=1}^{p}X_j\Delta_{\beta,j}||_F^2$. Let $\sigma^2_{\min}$ denote the smallest eigenvalues of $\mathbb{E}(X_{it}X_{it}')$. Let $\sigma^2_{\max}$ and $\sigma^2_{\min}$ denote the largest and the smallest eigenvalues of $\mathbb{E}(X_{it}X_{it}')$. By i.i.d.,

$$\mathbb{E}||\sum_{j=1}^{p}X_j\Delta_{\beta,j}||_F^2=NT\Delta_\beta'\mathbb{E}(X_{it}X_{it}')\Delta_\beta\in[NT\sigma^2_{\min}\gamma t^2,NT\sigma^2_{\max}\gamma t^2]$$

By Assumption 3.4.4, $\sigma^2_{\min}>0$. Hence, there exists a positive $c_0>0$ such that $\mathbb{E}||\sum_{j=1}^{p}X_j\Delta_{\beta,j}||_F^2=NTc_0\gamma t^2$.

Meanwhile, by the Pythagoras theorem,

$$\begin{aligned}\mathbb{E}||\sum_{j=1}^{p} X_j\Delta_{\beta,j}||_F^2 =&\mathbb{E}||P_{\Phi(u)}(\sum_{j=1}^{p} X_j\Delta_{\beta,j})||_F^2 + \mathbb{E}||P_{\Phi^\perp(u)}(\sum_{j=1}^{p} X_j\Delta_{\beta,j})||_F^2\\ =&\Delta_\beta'\sum_{i,t}\mathbb{E}\big((P_{\Phi(u)}\boldsymbol{X})_{it}(P_{\Phi(u)}\boldsymbol{X})_{it}'\big)\Delta_\beta + \Delta_\beta'\sum_{i,t}\mathbb{E}\big((P_{\Phi^\perp(u)}\boldsymbol{X})_{it}(P_{\Phi^\perp(u)}\boldsymbol{X})_{it}'\big)\Delta_\beta\end{aligned}$$

where $(P_{\Phi(u)}\boldsymbol{X})_{it}$ and $(P_{\Phi^\perp(u)}\boldsymbol{X})_{it}$ are defined in Assumption 3.4.4.

Let $c_1\Delta_\beta'\sum_{i,t}\mathbb{E}\big((P_{\Phi(u)}\boldsymbol{X})_{it}(P_{\Phi(u)}\boldsymbol{X})_{it}'\big)\Delta_\beta = \Delta_\beta'\sum_{i,t}\mathbb{E}\big((P_{\Phi^\perp(u)}\boldsymbol{X})_{it}(P_{\Phi^\perp(u)}\boldsymbol{X})_{it}'\big)\Delta_\beta$.

Next consider Δ_L. Since it is in the restricted set, by the derivation in Section 3.3,

$$||P_{\Phi^\perp(u)}\Delta_L||_F \leq \frac{C_1\sqrt{p\log(NT)(N\wedge T)}}{C_2}\sqrt{\gamma}t + 3\sqrt{3\bar{r}}||P_{\Phi(u)}\Delta_L||_F$$

Let $c_2||P_{\Phi(u)}\Delta_L||_F^2 = ||P_{\Phi^\perp(u)}\Delta_L||_F^2$, then by $||P_{\Phi(u)}\Delta_L||_F^2 + ||P_{\Phi^\perp(u)}\Delta_L||_F^2 = NT(1-\gamma)t^2$, we have

$$\sqrt{\frac{c_2}{c_2+1}}\sqrt{1-\gamma} \leq \frac{C_1\sqrt{\log(NT)}}{C_2\sqrt{N\vee T}}\sqrt{p\gamma} + 3\sqrt{3\bar{r}}\sqrt{\frac{1-\gamma}{c_2+1}} \tag{C.2.2}$$

Now let us consider the expectation under investigation.

$$
\begin{aligned}
&\mathbb{E}||\sum_{j=1}^{p} X_j\Delta_{\beta,j}+\Delta_L||_F^2 \\
=&\mathbb{E}||P_{\Phi(u)}(\sum_{j=1}^{p} X_j\Delta_{\beta,j})||_F^2+||P_{\Phi(u)}\Delta_L||_F^2+2\mathbb{E}\langle\sum_{j=1}^{p} P_{\Phi(u)}X_j\Delta_{\beta,j},P_{\Phi(u)}\Delta_L\rangle \\
&+\mathbb{E}||P_{\Phi^\perp(u)}(\sum_{j=1}^{p} X_j\Delta_{\beta,j})||_F^2+||P_{\Phi^\perp(u)}\Delta_L||_F^2+2\mathbb{E}\langle\sum_{j=1}^{p} P_{\Phi^\perp(u)}X_j\Delta_{\beta,j},P_{\Phi^\perp(u)}\Delta_L\rangle \\
\geq&\mathbb{E}||P_{\Phi(u)}(\sum_{j=1}^{p} X_j\Delta_{\beta,j})||_F^2+||P_{\Phi(u)}\Delta_L||_F^2-2\mathbb{E}||\sum_{j=1}^{p} P_{\Phi(u)}X_j\Delta_{\beta,j}||_F\cdot||P_{\Phi(u)}\Delta_L||_F \\
&+\mathbb{E}||P_{\Phi^\perp(u)}(\sum_{j=1}^{p} X_j\Delta_{\beta,j})||_F^2+||P_{\Phi^\perp(u)}\Delta_L||_F^2-2\mathbb{E}||\sum_{j=1}^{p} P_{\Phi^\perp(u)}X_j\Delta_{\beta,j}||_F\cdot||P_{\Phi^\perp(u)}\Delta_L||_F \\
\geq&\mathbb{E}||P_{\Phi(u)}(\sum_{j=1}^{p} X_j\Delta_{\beta,j})||_F^2+||P_{\Phi(u)}\Delta_L||_F^2-2\sqrt{\mathbb{E}||\sum_{j=1}^{p} P_{\Phi(u)}X_j\Delta_{\beta,j}||_F^2}\cdot||P_{\Phi(u)}\Delta_L||_F \\
&+\mathbb{E}||P_{\Phi^\perp(u)}(\sum_{j=1}^{p} X_j\Delta_{\beta,j})||_F^2+||P_{\Phi^\perp(u)}\Delta_L||_F^2-2\sqrt{\mathbb{E}||\sum_{j=1}^{p} P_{\Phi^\perp(u)}X_j\Delta_{\beta,j}||_F^2}\cdot||P_{\Phi^\perp(u)}\Delta_L||_F \\
=&\Big(\sqrt{\mathbb{E}||P_{\Phi(u)}(\sum_{j=1}^{p} X_j\Delta_{\beta,j})||_F^2}-||P_{\Phi(u)}\Delta_L||_F\Big)^2+\Big(\sqrt{\mathbb{E}||P_{\Phi^\perp(u)}(\sum_{j=1}^{p} X_j\Delta_{\beta,j})||_F^2}-||P_{\Phi^\perp(u)}\Delta_L||_F\Big)^2 \\
=&\Big[\Big(\sqrt{\frac{c_0\gamma}{1+c_1}}-\sqrt{\frac{1-\gamma}{1+c_2}}\Big)^2+\Big(\sqrt{\frac{c_0c_1\gamma}{1+c_1}}-\sqrt{\frac{c_2(1-\gamma)}{1+c_2}}\Big)^2\Big]NTt^2
\end{aligned}
$$

where the last inequality follows from convexity of $||\cdot||_F^2$ and Jensen's inequality.

Now it is sufficient to show that $\big[\big(\sqrt{\frac{c_0\gamma}{1+c_1}}-\sqrt{\frac{1-\gamma}{1+c_2}}\big)^2+\big(\sqrt{\frac{c_0c_1\gamma}{1+c_1}}-\sqrt{\frac{c_2(1-\gamma)}{1+c_2}}\big)^2\big]$ has a positive minimum.

First, from equation (C.2.2), if $\frac{C_1\sqrt{\log(NT)}}{C_2\sqrt{N\vee T}}\sqrt{p\gamma}$ dominates $3\sqrt{3\bar{r}}\sqrt{\frac{1-\gamma}{c_2+1}}$ in order, then $(1-\gamma)=o(1)$, and both $\sqrt{\frac{1-\gamma}{c_2+1}}$ and $\sqrt{\frac{c_2}{c_2+1}}\sqrt{1-\gamma}$ are dominated by $\sqrt{p\gamma}$. Therefore, as long as there exists some positive constant ϖ such that $c_1>\varpi$, the second term in the bracket is greater than $\frac{c_0\varpi}{2(1+\varpi)}$ for large enough N and T.

Now consider the case that $3\sqrt{3\bar{r}}\sqrt{\frac{1-\gamma}{c_2+1}}$ weakly dominates $\frac{C_1\sqrt{\log(NT)}}{C_2\sqrt{N\vee T}}\sqrt{p\gamma}$. Then for

large enough N and T, $c_2 \leq C_0\bar{r}$ for some $C_0 > 0$. We have

$$\begin{aligned}
&(\sqrt{\frac{c_0\gamma}{1+c_1}} - \sqrt{\frac{1-\gamma}{1+c_2}})^2 + (\sqrt{\frac{c_0c_1\gamma}{1+c_1}} - \sqrt{\frac{c_2(1-\gamma)}{1+c_2}})^2 \\
=&c_0\gamma + (1-\gamma) - 2(\sqrt{\frac{1}{(1+c_1)(1+c_2)}} + \sqrt{\frac{c_1c_2}{(1+c_1)(1+c_2)}})\sqrt{c_0\gamma(1-\gamma)}
\end{aligned}$$

It can be verified that the right hand side is bounded away from 0 if there exists a constant $0 < \eta < 1$ such that

$$\sqrt{\frac{1}{(1+c_1)(1+c_2)}} + \sqrt{\frac{c_1c_2}{(1+c_1)(1+c_2)}} < 1 - \eta$$

The inequality holds if $c_1 - c_2 > \eta'$ for some $\eta' > 0$. Since $c_2 \leq C_0\bar{r}$, this is the case if $c_1 > C_0'\bar{r}$ for some $C_0' > C_0$. By the definition of c_1, it is equivalent to

$$\Delta_\beta' \sum_{i,t} \mathbb{E}\big((P_{\Phi^\perp(u)}\boldsymbol{X})_{it}(P_{\Phi^\perp(u)}\boldsymbol{X})_{it}' - C_0'\bar{r}(P_{\Phi(u)}\boldsymbol{X})_{it}(P_{\Phi(u)}\boldsymbol{X})_{it}'\big)\Delta_\beta > 0,$$

which is guaranteed by Assumption 3.4.4. □

This completes the proof.

C.3 Proof of Lemmas C.1.1 and C.2.2

Proof of Lemma C.1.1. Under Assumption 3.3.1, there exists a constant $C > 0$ such that $\max_{1\leq j\leq p} ||X_j||_F^2 \leq C\sqrt{NT}$ with high probability. In what follows, all probabilities and expectations are implicitly taken conditional on this event and on $\{X_j\}_{j=1}^p$.

Proof of Equation (C.1.1).

Let $\mathcal{U}_K = (u_1, u_2, ..., u_K)$ be an ε-net in $\mathcal{U}$. Let $\varepsilon = \frac{1}{\sqrt{NT}}$. Then

$$\begin{aligned}
&\sup_{u\in\mathcal{U}} \left|\left\langle \nabla\boldsymbol{\rho}_u(V(u)), X_j\right\rangle\right| \\
\leq &\max_{u_k\in\mathcal{U}_K} \left|\left\langle \nabla\boldsymbol{\rho}_{u_k}(V(u_k)), X_j\right\rangle\right| + \sup_{|u-u_k|\leq\varepsilon, u_k\in\mathcal{U}_K} \left|\left\langle \nabla\boldsymbol{\rho}_u(V(u)) - \nabla\boldsymbol{\rho}_{u_k}(V(u_k)), X_j\right\rangle\right|
\end{aligned}$$

For the first term, since the length of $\mathcal{U}_K$ is no greater than 1,

$$\begin{aligned}
&\mathbb{P}\big(\max_{u_k\in\mathcal{U}_K} \left|\left\langle \nabla\boldsymbol{\rho}_{u_k}(V(u_k)), X_j\right\rangle\right| \geq C_1\sqrt{NT\log(NT)}\big) \\
\leq &\frac{1}{\varepsilon}\max_{u_k\in\mathcal{U}_K} \mathbb{P}\big(\left|\left\langle \nabla\boldsymbol{\rho}_{u_k}(V(u_k)), X_j\right\rangle\right| \geq C_1\sqrt{NT\log(NT)}\big) \\
\leq &\frac{2}{\varepsilon}\exp\big(-\frac{2C_1^2\log(NT)NT}{||X_j||_F^2}\big) \\
\leq &\frac{2}{\varepsilon}\exp\big(-\frac{2C_1^2\log(NT)NT}{CNT}\big) \\
\leq &\frac{C_1'}{\sqrt{NT}}
\end{aligned}$$

where the third line is by Hoeffding's inequality for $C_1 > \frac{\sqrt{2}}{2}$.

For the second term, by definition, the (i,t)-th element in $\nabla\boldsymbol{\rho}_u(V(u)) - \nabla\boldsymbol{\rho}_{u_k}(V(u_k))$ is

$$\begin{aligned}
&u\mathbb{1}_{V_{it}(u)>0} + (u-1)\mathbb{1}_{V_{it}(u)<0} - [u_k\mathbb{1}_{V_{it}(u_k)>0} + (u_k-1)\mathbb{1}_{V_{it}(u_k)<0}] \\
=&(u-u_k) + u_k(\mathbb{1}_{V_{it}(u)>0} - \mathbb{1}_{V_{it}(u_k)>0}) + (u_k-1)(\mathbb{1}_{V_{it}(u)<0} - \mathbb{1}_{V_{it}(u_k)<0})
\end{aligned}$$

The first term forms a constant matrix M_1 such that $||M_1||_F = \sqrt{NT}(u-u_k) \leq \varepsilon\sqrt{NT}$. Therefore, by the Cauchy-Schwartz inequality,

$$\left|\left\langle M_1, X_j\right\rangle\right| \leq ||M_1||_F||X_j||_F \leq CNT\varepsilon = C\sqrt{NT}$$

For the remaining terms, let $\xi_{it}(u,u_k) = u_k(\mathbb{1}_{V_{it}(u)>0} - \mathbb{1}_{V_{it}(u_k)>0}) + (u_k-1)(\mathbb{1}_{V_{it}(u)<0} -$

$\mathbb{1}_{V_{it}(u_k)<0}) = \mathbb{1}_{V_{it}(u_k)<0} - \mathbb{1}_{V_{it}(u)<0} = \mathbb{1}_{U_{it}<u_k} - \mathbb{1}_{U_{it}<u}$ where the last equality is from the definition of $V_{it}(\cdot)$. Therefore, if $u < u_k$, $0 \leq \xi_{it}(u, u_k) \leq \xi_{it}^{(1)}(u_k) = \mathbb{1}_{U_{it}<u_k} - \mathbb{1}_{U_{it}<u_{k-1}} \leq 1$. If $u \geq u_k$, $0 \geq \xi_{it}(u, u_k) \geq \xi_{it}^{(2)}(u_k) = \mathbb{1}_{U_{it}<u_k} - \mathbb{1}_{U_{it}<u_{k+1}} \geq -1$. Let $M_2^o = (\xi_{it}(u, u_k))_{1\leq i\leq N, 1\leq j\leq T}$, $M_2^{(1)} \equiv (\xi_{it}^{(1)}(u_k))_{1\leq i\leq N, 1\leq j\leq T}$, and $M_2^{(2)} \equiv (\xi_{it}^{(2)}(u_k))_{1\leq i\leq N, 1\leq j\leq T}$, we have

$$\begin{aligned}
\sup_{|u-u_k|\leq\varepsilon, u_k\in\mathcal{U}_k} \left|\langle M_2^o, X_j\rangle\right| &\leq \sup_{u_k-\varepsilon\leq u\leq u_k, u_k\in\mathcal{U}_k} \left|\langle M_2^o, X_j\rangle\right| + \sup_{u_k\leq u\leq u_k+\varepsilon, u_k\in\mathcal{U}_k} \left|\langle M_2^o, X_j\rangle\right| \\
&\leq \sup_{u_k-\varepsilon\leq u\leq u_k, u_k\in\mathcal{U}_k} \left|\langle M_2^o, |X_j|\rangle\right| + \sup_{u_k\leq u\leq u_k+\varepsilon, u_k\in\mathcal{U}_k} \left|\langle M_2^o, |X_j|\rangle\right| \\
&\leq \max_{u_k\in\mathcal{U}_k} \left|\langle M_2^{(1)}, |X_j|\rangle\right| + \max_{u_k\in\mathcal{U}_k} \left|\langle M_2^{(2)}, |X_j|\rangle\right|
\end{aligned}$$

The second inequality holds because in each of the two cases, all the elements in M_2^o have the same sign. So the inner product is maximized if the elements in X_j also have the same sign. The third inequality then follows because now the magnitude of the inner product is increasing in the magnitude of any elements in M_2^o.

I now only discuss $\max_{u_k\in\mathcal{U}_k} \left|\langle M_2^{(1)}, |X_j|\rangle\right|$ as the argument for $\max_{u_k\in\mathcal{U}_k} \left|\langle M_2^{(2)}, |X_j|\rangle\right|$ is identical. The expectation of a generic element in $M_2^{(1)}$ satisfies $\mu_k \equiv \mathbb{E}(\xi_{it}^{(1)}(u_k)) = \mathbb{P}(u_k - \varepsilon < U_{it} \leq u_k) = \varepsilon$ because U_{it} follows Unif$[0, 1]$. Let $\bar{M}_2^{(1)}$ be an $N \times T$ matrix of εs. Then we have

$$\begin{aligned}
\max_{u_k\in\mathcal{U}_k} \left|\langle M_2^{(1)}, |X_j|\rangle\right| &\leq \max_{u_k\in\mathcal{U}_k} \left|\langle M_2^{(1)} - \bar{M}_2^{(1)}, |X_j|\rangle\right| + \max_{u_k\in\mathcal{U}_k} \left|\langle \bar{M}_2^{(1)}, |X_j|\rangle\right| \\
&\leq \max_{u_k\in\mathcal{U}_k} \left|\langle M_2^{(1)} - \bar{M}_2^{(1)}, |X_j|\rangle\right| + \varepsilon\sqrt{NT}||X_j||_F \\
&\leq \max_{u_k\in\mathcal{U}_k} \left|\langle M_2^{(1)} - \bar{M}_2^{(1)}, |X_j|\rangle\right| + C\sqrt{NT}
\end{aligned}$$

Finally, the first term is also bounded by $C_1\sqrt{NT\log(NT)}$ following exactly the same argument as for $\max_{u_k\in\mathcal{U}_K} \left|\left\langle \nabla\boldsymbol{\rho}_{u_k}(V(u_k)), X_j\right\rangle\right|$ because elements in $(M_2^{(1)} - \bar{M}_2^{(1)})$ are i.i.d, mean zero, and bounded in magnitude by 1. Therefore, as the bound does not

depend on j and p is a constant, equation (C.1.1) follows.

Proof of Equation (C.1.2). The proof follows similar argument as for equation (C.1.1). Again, let $\mathcal{U}_K = (u_1, u_2, ..., u_K)$ be an ε-net in $\mathcal{U}$. This time let $\varepsilon = \frac{1}{\sqrt{N \vee T}}$. Then

$$\sup_{u \in \mathcal{U}} ||\nabla \boldsymbol{\rho}_u(V(u))|| \leq \max_{u_k \in \mathcal{U}_K} ||\nabla \boldsymbol{\rho}_{u_k}(V(u_k))|| + \sup_{|u-u_k| \leq \varepsilon, u_k \in \mathcal{U}_K} ||\nabla \boldsymbol{\rho}_u(V(u)) - \nabla \boldsymbol{\rho}_{u_k}(V(u_k))||$$

For the first term,

$$\begin{aligned}
\mathbb{P}\big(\max_{u_k \in \mathcal{U}_K} ||\nabla \boldsymbol{\rho}_{u_k}(V(u_k))|| > C_2 \sqrt{N \vee T} \big) &\leq \frac{1}{\varepsilon} \max_{u_k \in \mathcal{U}_K} \mathbb{P}\big(||\nabla \boldsymbol{\rho}_{u_k}(V(u_k))|| > C_2 \sqrt{N \vee T} \big) \\
&\leq \frac{C_2'}{\varepsilon} \exp(-C_2'' \sqrt{NT}) \\
&= \frac{C_2'' \sqrt{N \vee T}}{\exp(C_2'' \sqrt{NT})}
\end{aligned}$$

where the second line is from Corollary 2.3.5 in Tao (2012) (p.129) that bounds the operator norm of a matrix with i.i.d., mean zero entries that are bounded in magnitude by 1.

For the second term, from the proof of equation (C.1.1),

$$\sup_{|u-u_k| \leq \varepsilon, u_k \in \mathcal{U}_K} ||\nabla \boldsymbol{\rho}_u(V(u)) - \nabla \boldsymbol{\rho}_{u_k}(V(u_k))|| \leq \sup_{|u-u_k| \leq \varepsilon, u_k \in \mathcal{U}_K} \big(||M_1|| + ||M_2^o|| \big)$$

where M_1 and M_2^o are defined in the proof of equation (C.1.1). By definition, the operator norm of a generic matrix A is $\sup_{||x||_F = 1} ||Ax||_F$ where x is a vector of unit Euclidean norm. When all the elements in A have the same sign, the supremum is achieved if all elements in x also have the same sign. Therefore, $\sup_{||x||_F=1} ||Ax||_F \leq \sup_{||x||_F=1} ||A \cdot |x|||_F$. Then for a matrix B such all elements in B also have the same sign and have weakly larger magnitude than those in A, $||A|| = \sup_{||x||_F=1} ||A \cdot |x|||_F \leq \sup_{||x||_F=1} ||B \cdot |x|||_F =$

$\sup_{||x||_F=1} ||Bx||_F = ||B||$. Hence,

$$\begin{aligned}\sup_{\varepsilon, u_k \in \mathcal{U}_K} \left(||M_1|| + ||M_2^o||\right) \leq & \sup_{u_k - \varepsilon \leq u \leq u_k, u_k \in \mathcal{U}_k} \left(||M_1|| + ||M_2^o||\right) + \sup_{u_k \leq u \leq u_k + \varepsilon, u_k \in \mathcal{U}_k} \left(||M_1|| + ||M_2^o||\right) \\ \leq & 2\varepsilon ||\mathbf{1}_{N\times T}|| + \sup_{u_k - \varepsilon \leq u \leq u_k, u_k \in \mathcal{U}_k} ||M_2^{(1)}|| + \sup_{u_k \leq u \leq u_k + \varepsilon, u_k \in \mathcal{U}_k} ||M_2^{(2)}|| \\ = & 2\varepsilon ||\mathbf{1}_{N\times T}|| + \max_{u_k \in \mathcal{U}_k} ||M_2^{(1)}|| + \max_{u_k \in \mathcal{U}_k} ||M_2^{(2)}||\end{aligned}$$

where $\mathbf{1}_{N\times T}$ is a constant matrix of ones whose operator norm is $O(N \vee T)$. It arises from M_1. $M_2^{(1)}$ and $M_2^{(2)}$ follow the same definitions in the proof of equation (C.1.1).

Again, let us only consider $\max_{u_k \in \mathcal{U}_k} ||M_2^{(1)}||$.

$$\max_{u_k \in \mathcal{U}_k} ||M_2^{(1)}|| \leq \max_{u_k \in \mathcal{U}_k} ||M_2^{(1)} - \bar{M}_2^{(1)}|| + \max_{u_k \in \mathcal{U}_k} ||\bar{M}_2^{(1)}||$$

Note that elements in $M_2^{(1)} - \bar{M}_2^{(1)}$ are again i.i.d., mean zero, and bounded in magnitude by 1, so it has the same upper bound as $\max_{u_k \in \mathcal{U}_K} ||\nabla \boldsymbol{\rho}_{u_k}(V(u_k))||$. For the second term, $\bar{M}_2^1 = \varepsilon \mathbf{1}_{N\times T}$. Therefore $\max_{u_k \in \mathcal{U}_k} ||\bar{M}_2^{(1)}|| \leq C_2''' \sqrt{N \vee T}$ by the choice of ε. □

Proof of Lemma C.2.2. For any fixed $w_1 \in \mathbb{R}$, $\mathbb{1}(w_1 \leq z)$ is weakly increasing in z. Therefore, if $w_2 > 0$, $z \geq 0$, so $\mathbb{1}(w_1 \leq z) - \mathbb{1}(w_1 \leq 0) \geq 0$, the second inequality thus holds. Similarly,

$$\begin{aligned}&\int_0^{w_2} \big(\mathbb{1}(w_1 \leq z) - \mathbb{1}(w_1 \leq 0)\big) dz - \int_0^{\kappa w_2} \big(\mathbb{1}(w_1 \leq z) - \mathbb{1}(w_1 \leq 0)\big) dz \\ = &\int_{\kappa w_2}^{w_2} \big(\mathbb{1}(w_1 \leq z) - \mathbb{1}(w_1 \leq 0)\big) dz\end{aligned}$$

Since $\kappa w_2 < w_1$, the right hand side is nonnegative. Hence the first inequality holds.

When $w_2 \leq 0$, note that $\int_0^{w_2} \big(\mathbb{1}(w_1 \leq z) - \mathbb{1}(w_1 \leq 0)\big) dz = \int_{w_2}^0 \big(\mathbb{1}(w_1 \leq 0) - \mathbb{1}(w_1 \leq z)\big) dz$ and $\int_0^{\kappa w_2} \big(\mathbb{1}(w_1 \leq z) - \mathbb{1}(w_1 \leq 0)\big) dz = \int_{\kappa w_2}^0 \big(\mathbb{1}(w_1 \leq 0) - \mathbb{1}(w_1 \leq z)\big) dz$. Now that $z \leq 0$, $\mathbb{1}(w_1 \leq 0) - \mathbb{1}(w_1 \leq z) \geq 0$. Therefore, following the same argument in the previous case, we obtain the desired result. □