UNIVERSITY OF CALIFORNIA, SAN DIEGO

Nonseparable Panel Data Models
Identification, Estimation and Testing

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Economics

by

Dalia A. Ghanem

Committee in charge:

Professor Graham Elliott, Chair
Professor Ian Abramson
Professor Todd Kemp
Professor Ivana Komunjer
Professor Andres Santos
Professor Yixiao Sun
Professor Junjie Zhang

2013

The dissertation of Dalia A. Ghanem is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2013

DEDICATION

To my mother

In memory of my father

EPIGRAPH

Imagination is more important than knowledge.
For knowledge is limited to all we now know and understand,
while imagination embraces the entire world,
and all there ever will be to know and understand.
—Albert Einstein

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

The road to this dissertation was long and far, and I am indebted to everyone who supported me along the way from Egypt to San Diego. First of all, I would like to thank my dissertation committee: Ian Abramson, Graham Elliott, Todd Kemp, Ivana Komunjer, Andres Santos, Yixiao Sun, the late Halbert White, and Junjie Zhang. I am grateful to my advisor, Graham Elliott, for nurturing me intellectually and for allowing me to be stubborn and carve my own way. I will always be indebted to him for his patience and guidance. I also appreciate the support of Andres Santos; I will never forget the crash course he gave me in empirical process theory. I would also like to thank Ivana Komunjer and Yixiao Sun for many helpful discussions. Even though the late Halbert White has not seen most of the work in this dissertation, his early work on misspecification has always been an inspiration to me. I am also grateful for having a deep discussion about what a 'fixed effect' is with him. I would like to thank Junjie Zhang for the research assistant position he offered me. It was a great opportunity for me to devote more of my time to research and to get experience in empirical work. I was also fortunate to meet regularly with members of the faculty at UCSD who were not formally on my committee. Prashant Bharadwaj, Julie Cullen, and Jason Schweinsberg were very generous with their time, and I learnt a lot from my discussions with them. I am grateful to Prashant Bharadwaj for providing the dataset that I use in the empirical illustration in Chapter 3.

It is not an overstatement to say that I would not have pursued my Ph.D. with econometrics in mind, if it was not for William Mikhail, my first professor in econometrics at the American University in Cairo (AUC). His lectures made me realize what an exciting field it is, and since then, I was literally 'hooked'. After

completing my undergraduate degree, I started working and could have strayed away from the Ph.D. route if it was not for the support of Adel Beshai and Karim Abadir. Their conviction that I should do it helped me in making my decision to embark on this journey to the 'wild, wild west'. I am also indebted to Abdel-Aziz Ezz El Arab, Ahmed Kamaly, and Nagla Rizk for their support and enthusiasm. I will never forget how Nagla always told me 'Just Do It'.

I would like to thank my former colleagues and superiors at the Egyptian Center for Economic Studies and the Ministry of Trade and Industry, who had a positive impact on how I see my research in the greater scheme of things. I am especially grateful to H.E. Rachid Mohamed Rachid, Egypt's former Minister of Trade and Industry, for teaching me to be skeptical of econometric models and their relevance for the policymaking process. I hope that this skepticism will always be the driving force for my research.

I definitely could not have made it through the last six years if it was not for the support of my loving husband, Sean, especially after our son Fareed arrived. Sean and I were very lucky to have Sean's parents, Mary Ann and Hilliard, in town to help take care of Fareed. As for Fareed, I would like to thank him for being the most welcome distraction I could have possibly hoped for.

Growing up, my mom's love for knowledge and education was definitely infectious, and her loving support reached across oceans and continents. My father also played an important role in my education. Early on, he decided that I should attend the German School in Alexandria, Egypt. I believe that my life choices, not just my career choices, would have been very different if I had not attended this

school. I would also like to take this opportunity to thank my mother, brother and sister-in-law for supporting my decision to come to the U.S. to pursue my Ph.D., even though it would bring me further away from them. I consider myself very lucky to have such a selfless family. My good fortune was that I also found this support from my uncles and godparents who always took a great interest in my education and career. I am also grateful for my aunt who hosted me during my undergraduate education.

I believe my parents' emphasis on education in my upbringing comes from their own parents. Even though both of my grandmothers were stay-at-home moms, they were extremely knowledgeable, well-read and independent. Thinking back, it was great to have such strong women as a role model.

I would also like to thank my friends who are spread out between Egypt and San Diego for helping me through the trials and tribulations of graduate school. My abisurd group will always be my rock. Shabnam Shalizi was the first friend I had in San Diego, she helped me during my adjustment to San Diego as well as during my preparations for the job market interviews. I will always be grateful for her friendship. I would also like to thank my UCSD friends for their friendship and good company. Special thanks go to Todd Kuffner, my study mate at LSE; Dallas Dotter, Kirti Gupta, and the late Ben Horne, my first-year study group, and Sarojini Hirshleifer whose support during the crunch time of the job market was a godsend.

If I forgot to thank you, thank you and please forgive me.

Chapter 4 is co-authored with Junjie Zhang, School of International Relations and Pacific Studies, UCSD, and has been submitted to the Journal of Environmental Economics and Management for publication.

VITA

2003	B. A. in Economics and Political Science, American University in Cairo
2007	M. Sc. in Econometrics and Mathematical Economics, London School of Economics
2013	Ph. D. in Economics, University of California, San Diego

ABSTRACT OF THE DISSERTATION

Nonseparable Panel Data Models
Identification, Estimation and Testing

by

Dalia A. Ghanem

Doctor of Philosophy in Economics

University of California, San Diego, 2013

Professor Graham Elliott, Chair

Microeconomic panel data, also known as longitudinal data or repeated measures, allow the researcher to observe the same individual across time. One of the advantages of panel data is that they allow the researcher to control for unobservable individual heterogeneity. The linear fixed effects model is the most commonly used method in empirical work to control for unobservable heterogeneity. Chapter 1 reviews the special features of the linear fixed effects model in detail, giving special attention to the definition of fixed effects and correlated random effects. It discusses the issues that arise when we move from a linear model

to fully nonseparable models and reviews the two strands of the literature that are relevant for this dissertation: (1) the literature on nonlinear parametric panel data models with fixed effects, (2) the literature on nonparametric identification in nonseparable panel data models.

Chapter 2 falls under the parametric nonlinear panel data models with fixed effects. Nonlinear panel data models with fixed effects are an important example in econometrics where the incidental parameter problem arises and the maximum likelihood estimator (MLE) is asymptotically biased. Bias correction of the MLE achieves consistency without increasing the asymptotic variance. Chapter 2 proposes a shrinkage estimator that combines that is shown to lead to a higher-order mean-square error improvement over the analytical bias-corrected estimator.

Chapter 3 falls under the literature on nonparametric identification in nonseparable panel data models. Starting from a general DGP that exhibits nonseparability of the structural function, arbitrary individual and time heterogeneity, I give a necessary and sufficient condition for the point-identification of the APE for a subpopulation. This condition is then used to characterize the trade-off between assumptions on unobservable heterogeneity and the structural function that achieve identification. The identifying assumptions here have clear testable implications on the distribution of observables. I hence propose bootstrap-adjusted Kolmogorv-Smirnov and Cramer-von-Mises statistics to test these implications.

Chapter 4 is an empirical paper that studies the issue of manipulation of air pollution data by Chinese cities. It applies tests similar in spirit to the tests proposed in Chapter 3 to test the presence of manipulation.

Chapter 1

From Linear to Nonseparable Panel Data Modeling: A Literature Review

1.1 Introduction

In the past decade, there has been major developments in the literature on identification and estimation of average structural functions and average effects in nonseparable panel data models, both parametric and nonparametric. The goals of this chapter is (1) to revisit the linear model and outline its special features, (2) to clarify the fundamental definition of fixed effects and correlated random effects, (3) to discuss the issues that arise when we move to nonseparable models, and (4) to review the relevant literature. This chapter hence provides a detailed overview of the panel data literature that is used as a foundation for the following chapters.

The general nonseparable model we have in mind is

$$Y_{it} = \phi(X_{it}, Z_{it}, \mathcal{A}_i, \mathcal{L}_t, \mathcal{U}_{it}), \tag{1.1}$$

where Y_{it} is the outcome of interest, X_{it} is a finite-dimensional vector of treatment variables, Z_{it} is a finite-dimensional vector of control variables, $\mathcal{A}_i$ captures unobservable time-invariant individual characteristics (hereinafter, individual effects), $\mathcal{L}_t$ captures unobservable individual-invariant characteristics, and $\mathcal{U}_{it}$ captures time-varying characteristics. The above model is completely nonparametric, $\mathcal{A}_i$ and $\mathcal{L}_t$ here are thought of as unobservable regressors.

We are typically interested in some measure of the effect of X on Y. In a linear model, which is a parametric additively separable version of (1.1)

$$Y_{it} = X_{it}'\beta + Z_{it}'\gamma + \mathcal{A}_i + \mathcal{L}_t + \mathcal{U}_{it}, \tag{1.2}$$

the effect is simply β, which is the average effect at all points of support of X under appropriate mean independence assumptions on $\mathcal{U}_{it}$. Section 1.2 outlines the equivalences that exist in the linear model. Section 1.3 gives special attention to the debate about whether the individual effects are to be thought of as fixed or random. Section 1.4 discusses the issues that arise when we move from the linear to the nonseparable model. Section 1.5 reviews the literature on nonseparable models. If the reader is not concerned about the fixed vs. random effects debate, I would recommend skipping Section 1.3, as it may seem unnecessarily pedantic.

1.2 Equivalences in the Linear Model

For the discussion of the linear model, we will set $\gamma = 0$ and $\mathcal{L}_t = 0$ for all t. The classical example in panel data which appeared in Mundlak (1961) is a log-transformation of a Cobb-Douglas production function.

$$Y_{it} = X'_{it}\beta + \mathcal{A}_i + \mathcal{U}_{it} \tag{1.3}$$

where Y_{it} is log output, X_{it} is a vector of log inputs (such as labor, livestock, and machinery), $\mathcal{A}_i$ is management ability, $\mathcal{U}_{it}$ stochastic input that is outside of farmer's control (rainfall).

Within-Group Transformation (Invariance). The key feature of the linear model which makes it completely different from the general nonseparable setup is the presence of a transformation that renders β invariant with respect to $\mathcal{A}_i$, as shown in Mundlak (1978). By demeaning,[1] we difference out $\mathcal{A}_i$ and obtain

[1]First-differencing and other forms of differencing ((Angrist and Newey, 1991)) are alternative

the following model

$$Y_{it} - \bar{Y}_i = (X_{it} - \bar{X}_i)'\beta + \mathcal{U}_{it} - \bar{\mathcal{U}}_i. \tag{1.4}$$

Hence, we do not need to observe or estimate $\mathcal{A}_i$ to identify and estimate β. Furthermore, $\mathcal{A}_i$ may be arbitrarily correlated with $X_i = (X_{i1}, X_{i2}, ..., X_{iT})$. Applying OLS to the above equation yields a consistent estimator for β that does not depend on $\mathcal{A}_i$. Arellano (2003) refers to the resulting estimator as the within-group estimator.

Correlated Random Effects . Another interesting, yet less known result in Mundlak (1978) is that the within-group estimator is equivalent to the correlated random effects estimator in Chamberlain (1984), where $\mathcal{A}_i \sim N(\bar{X}_i\pi, \sigma_\alpha^2)$, where $\bar{X}_i = (\sum_{t=1}^{T} X_{it})$. This is an important equivalence, since correlated random effects are typically viewed as restrictive, hence it is important to note that they are equivalent to the within-group estimator and hence also the LSDV in the linear model.

Least Squares Dummy Variable Regression ('Fixed Effects'). It so happens that the within-group estimator is equivalent to the least squares dummy variable (LSDV) regression, where we treat $\mathcal{A}_i$ as an estimand. To be consistent with our notation, whenever we deal with $\mathcal{A}_i$ as an estimand, we will refer to it as α_i, the relevant estimating equation is the following

$$Y_{it} \;=\; X_{it}'\beta + \alpha_i + \mathcal{U}_{it}. \tag{1.5}$$

ways of differencing-out the individual effects.

The above equation is the one more commonly associated with the fixed effects estimator, α_i are also given the unfortunate name, 'fixed effects'. I discuss this common misconception and why it is confusing and misleading when we move to the nonseparable models in Section 1.3. Arellano (2003) refers to this as the joint likelihood approach, where $\mathcal{U}_{it}$ is assumed iid $N(0, \sigma^2)$. The likelihood function is given by the following

$$L(\beta, \sigma^2, \{\alpha_i\}_{i=1}^n; Y, X) = \sum_{i=1}^{n} \log f(Y_i|X_i; \alpha_i, \beta, \sigma^2), \tag{1.6}$$

where L denotes the likelihood function and f denotes the conditional density. The literature on parametric nonlinear models with fixed effects extends this approach to general nonlinear models.[2]

Sufficient Statistic. Arellano (2003) also points out that the within-group estimator can be written as a conditional likelihood estimator where $\bar{Y}_i$ is used as a sufficient statistic for $\mathcal{A}_i$, i.e.

$$\log f(Y_i|X_i, \mathcal{A}_i, \bar{Y}_i) = \log f(Y_i|X_i, \bar{Y}_i). \tag{1.7}$$

It is important to note here that sufficiency is used to estimate 'fixed effects' logit and poisson models in order to circumvent the incidental parameters problem.

Unrestricted Conditional Distribution of $\mathcal{A}_i|X_i$. The key result in Arellano (2003) is that the within-group estimator for β can be obtained from the marginal likelihood approach, where $f(Y_i|X_i) = \int f(Y_i|X_i, \mathcal{A}_i) dF(\mathcal{A}_i|X_i)$ is

[2]This overview shows why this is not a natural way of extending the 'fixed effects' estimator to the nonlinear model in the absence of an invariant transformation.

used as the building block for the likelihood function, where $F(.|.)$ denotes the conditional cdf. Arellano (2003) decomposes the marginal likelihood as follows:

$$\begin{aligned} L(Y_i|X_i;\beta) &= \sum_{i=1}^{n} \log f(Y_i|X_i) \\ &= \sum_{i=1}^{n} \left\{ \log f(Y_i^*|X_i) + \log \int f(\bar{Y}_i|X_i, \mathcal{A}_i) dF(\mathcal{A}_i|X_i) \right\}, \quad (1.8) \end{aligned}$$

where Y_i^* is the demeaned model. Note that $f(Y_i^*|X_i) = f(Y_i^*|X_i, \mathcal{A}_i)$, since the demeaned model is invariant to $\mathcal{A}_i$. The second term on the right-hand side has no effect on the estimation of β. $F(\mathcal{A}_i|X_i)$ may be any distribution that exhibits non-zero correlation between $\mathcal{A}_i$ and X_i. For instance, it cannot be $N(0, \sigma_{\mathcal{A}}^2)$. The marginal MLE is also maximized by the within-group estimator so long as there is some correlation $\mathcal{A}_i$ and X_i.[3]

1.2.1 Summary and Remarks

The above discussion shows that the linear model exhibits many equivalences that are specific to linearity, and more specifically the separability of $\mathcal{A}_i$ from the object of interest β

(i) invariance of the within-group transformation;[4]

(ii) sufficiency of $\bar{Y}_i$ for $\mathcal{A}_i$;

(iii) the equivalence of the within-group estimator and the correlated random effects estimator when $\mathcal{A}_i|X_i \sim N(\bar{X}_i'\delta, \sigma_{\mathcal{A}}^2)$

[3]Note that correlation is an appropriate measure of dependence in linear models, when we move to nonseparable models, it is more appropriate to consider higher-order forms of dependence as well.

[4]There are other transformations such as first-differencing that could also deliver invariance.

(iv) the equivalence of fixed effects estimation (where $\mathcal{A}_i$ is treated as an estimand α_i) and marignal MLE (where $F(\mathcal{A}_i|X_i)$ is nonparametrically defined);

(v) within-group estimator allows for arbitrary correlation between $\mathcal{A}_i$ and X_i by treating $\mathcal{A}_i$ nonparametrically, i.e. unrestricted conditional distribution of $\mathcal{A}_i|X_i$.

It is easy to see how these equivalences led the fixed effects assumption to be the name given to the case where the conditional distribution of $\mathcal{A}_i|X_i$ is unrestricted, as in Evdokimov (2010), Hoderlein and White (2012), and Chernozhukov et al. (2013). It is an unfortunate name, since it may lead us to think of $\mathcal{A}_i$ as fixed constants, which contradicts the idea that they may be correlated or dependent on X_i.

The following section will revisit the origins of the debate about whether the individual effects are fixed (constant) or random. It also attempts to clarify what is meant by fixed (i.e. fixed in repeated sampling) in this context.

1.3 Individual Effects: Fixed or Random?

It may seem pedantic or unnecessary to make the above distinction between a random variable $\mathcal{A}_i$ and a parameter α_i, since it has no effect on the resulting estimator of the object of interest, β, in the linear model. The reason why the distinction is made here is because once we move to the nonlinear model these equivalences no longer hold.

It seems that there was a big debate around the time Mundlak (1978) appeared in *Econometrica* about whether $\mathcal{A}_i$ should be thought of as fixed or random. Mundlak (1978) addresses this issue and points out that the effects are inherently random. In fixed-effects modeling, the idea is that we fix the individual effects across time in the sample, since we observe the same individuals across time. However, it does not mean that we think of $\mathcal{A}_i$ as a fixed constant.[5] The following example from Hsiao (2003) is used to illustrate the distinction.

Example 1.3.1 Two Experiments with Black and Red Balls (Hsiao, 2003)

Let the population be composed of a particular composition of red and black balls, p, which is the ratio of red balls to the entire population of balls. Now the following two experiments generate 2 different $n \times T$ panels, where $i = 1, 2, ..., n$ and $t = 1, 2, ..., T$

Table 1.1: Hsiao (2003): Black and Red Balls Experiment to Compare a Random Effects and Fixed Effects Setting

	Experiment 1	Experiment 2
For each i	each picks k balls randomly from the population	each picks k balls based on preference ($\mathcal{A}_i$)
For each t	random draw from individual's k balls	random draw from individual's k balls

They key difference between the two experiments is that in Experiment 2 the individual's k balls are picked according to the individual's preference. Hence, one would need to allow for arbitrary correlation between $\mathcal{A}_i$ and r_{it}, the indicator for whether a ball is red or black. Experiment 1, on the other hand, fulfills the basic

[5]This misconception may have been a consequence of the equivalence of the LSDV estimator to the within-group estimator.

random effects assumption.

Now Hsiao (2003) points out that, even under Experiment 2, the choice between fixed-effects and (correlated) random-effects modeling depends on the object of interest. If we seek to estimate the ratio of red and black balls in each individual urn, then we should condition on individual preferences ($\mathcal{A}_i = \alpha_i$), i.e. treat them as given.[6] *However, if we seek to estimate p, the population ratio, then we need to integrate out $\mathcal{A}_i$, i.e. control for the effect of preference on the ratio of black and red balls. For Experiment 2, such an approach would involve rectifying the 'bias' introduced in estimating p due to the individual bias toward red or black balls ($\mathcal{A}_i$). This is summarized in the following table.*

Table 1.2: Object of Interest, Model and Estimator: Comparing Fixed Effects and Correlated Random Effects Estimation

Object of interest	Model	Estimator
individual urn's composition	condition on $\mathcal{A}_i = \alpha_i$ fixed-effects	$\hat{p}(r_i, \alpha_i) = \sum_{t=1}^{T} r_{it}/T$
population composition (p)	integrate out $\mathcal{A}_i$ (correlated) random-effects	$p(r_i) = \int p(r_i, \mathcal{A}_i) f(\mathcal{A}_i \vert r_i) d\mathcal{A}_i$

r_{it} is a dummy for whether the ball for i, t is red.

Borrowing from Hsiao (2003) again to give an intuition for the difference between random effects, correlated random effects and fixed effects, let us think of the following states of nature. Under random effects, nature rolls the same dice to draw each individual $\mathcal{A}_i$. Under correlated random effects, nature rolls the same dice for individuals who have the same realizations of $h(X_i)$, which is a function

[6] Of course, in this case, we would need to observe a long time horizon. Since in microeconometric panel models, we do not have a long time horizon, the above shows that treating $\mathcal{A}_i$ as fixed constants is not appropriate, unless it is equivalent to some other estimator with favorable properties, such as the case in the linear model.

with support smaller than the support of X_i.[7] We can think of this as a case where we know which dice are chosen given the realization of X_i. Under fixed effects, nature arbitrarily chooses each individual $\mathcal{A}_i$ using different dice each time but we do not know any feature of the dice given the observable X_i.

In the following section, we will discuss why we need to be very clear about what we mean by fixed effects and correlated random effects when we move to the nonseparable model.

1.4 From Linear to Nonseparable Models

When we move from the linear model to the nonseparable world, there are two key issues that need to be clarified. The first issue is what is meant by the fixed effects assumption, the second is the object of interest. In the linear world, as the previous section shows, the within-group estimator is similar to many different estimators. However, when we move to the nonseparable world, it is no longer true that such equivalences hold in general. Hence, caution has to be exercised in using terms such as the fixed effects assumption.

The problem with interpreting the fixed effects assumption is that it is not an assumption. It is supposed to imply the absence of an assumption on the dependence between $\mathcal{A}_i$ and X_i in the sense of equation (1.8). Graham and Powell (2012) point out that the key difference between fixed effects and correlated random effects in a random coefficient model is that fixed effects leaves $F(\mathcal{A}_i|X_i)$

[7] For instance, it could be the same dice for individuals with the same mean of X_i.

unrestricted, whereas the correlated random effects approach imposes restrictions on $F(\mathcal{A}_i|X_i)$. We will take this to be the difference between fixed effects and correlated random effects in the nonseparable world. It is still important however to keep in mind the distinction between conditioning on $\mathcal{A}_i$ and integrating them out, since it may help to understand different approaches in the nonlinear parametric setup as well as the nonparametric world.

The second major difference between the linear and nonseparable models is the object of interest. In the linear model, under appropriate conditions, β, the slope-coefficient, is itself the overall average partial effect, the average partial effect at the average of the regressor, and it is also the average effect of a unit change in a discrete regressor. However, in the nonseparable world, these are all different objects of interest. Thus, special care is needed in defining what object of interest a researcher intends to identify.

1.5 Literature on Nonseparable Panel Data Models

The literature on nonseparable panel data models is vast. In the following, I will attempt to review recent papers in key segments of the literature. The objective here is to understand the big picture of recent developments in the literature, and to understand which of the properties of the within-group estimator in the linear model are fulfilled. For instance, the "fixed effects" logit and poisson models admit a sufficient statistic for $\mathcal{A}_i$, which is $\sum_{t=1}^{T} Y_{it}$ if our objective is to estimate the common parameter. As in the linear model, the conditional MLE approach fa-

cilitates estimation of the common parameter. However, since in nonlinear setups, the average partial effect is not equal to the slope coefficient or common parameter, focusing on the common parameter may facilitate its estimation at a price. The average partial effect cannot be estimated. In general parametric nonlinear models, to estimate an average partial effect,[8] there are two possible strategies: (correlated) random effects and fixed effects. The difficulty of the former lies in computational integration in order to remove the individual effect, the latter suffers from the incidental parameter problem. For a brief overview of correlated random effects modeling, see Wooldridge (2002). The following subsection reviews the fixed effects literature, since it has witnessed some more recent contributions.

In the nonparametric identification literature, discrete and continuous regressors require different assumptions for identifications, and are hence dealt with separately in the literature. For discrete regressors, Chernozhukov et al. (2013) show under conditions, such as time homogeneity, the average and quantile effects are identified in general nonparametric models. For continuous regressors, Bester and Hansen (2009) propose a nonparametric correlated random effects approach, whereas Hoderlein and White (2012) propose a generalized 'fixed effects' approach. Evdokimov (2010), building on Altonji and Matzkin (2005), studies identification of the structural function and the distribution of the individual effects in a nonparametric model with additively separable idiosyncratic unobservables ($\mathcal{U}_{it}$) under monotonicity and conditional independence assumptions. In the following, we review the two strands of the literature on nonseparable panel data models.

[8]The reason I use 'an' here is because there is not just one effect.

1.5.1 Nonlinear Parametric Panel Data Models

The estimation of nonlinear panel models with fixed effects is a classical example of the incidental parameter problem introduced in Neyman and Scott (1948). The simplest example in economics is that of a binary choice panel data model. So the problem will be illustrated using this example.

Let Y_{it}^* be a latent variable, characterized by

$$Y_{it}^* = X_{it}'\beta + \alpha_i + U_{it}, \tag{1.9}$$

where we observe the binary choice, $Y_{it} = 1_{\{(Y_{it}^* \geq 0\}}$, and X_{it}. α_i is termed the individual effect to be estimated. So in addition to the common parameters, there are n fixed effects to be estimated. Thus, as the sample size n tends to infinity, so does the number of estimands. Thus, the incidental parameter problem arises.

Given a panel data set of n individuals and T time periods and assuming a distribution for u_{it}, one may use MLE to estimate β as follows:

$$\hat{\beta}_{ML} \quad := \quad \arg\max_{\beta} \sum_{i=1}^{n} \sum_{t=1}^{T} l_{it}(\beta, \hat{\alpha}_i(\beta)), \tag{1.10}$$

where $\hat{\alpha}_i(\beta) := \arg\max_\beta \sum_{i=1}^n \sum_{t=1}^T l_{it}(\beta, \alpha_i)$. Now letting $n \to \infty$ with fixed T,

$$\hat{\beta}_{ML} \quad \xrightarrow{p} \quad \beta_T := \arg\max_{\beta} \bar{E}[\sum_{t=1}^{T} l_{it}(\beta, \hat{\alpha}_i(\beta))/T] \tag{1.11}$$

where $\bar{E}[g_i(X_{iT})] = \lim_{N\to\infty} n^{-1} \sum_{i=1}^n E[g_i(X_{iT})]$. Note that $\beta_T \neq \beta_0$, however $\lim_{T\to\infty} \beta_T = \beta_0$.

Expanding β_T in $\hat{\alpha}_i(\beta)$ around α_i leads to a bias of order $1/T$,

$$\beta_T = \beta_0 + \frac{B}{T} + O(\frac{1}{T^2}). \tag{1.12}$$

Hahn and Newey (2004) explains that the bias stems from three sources: (1) the bias of $\hat{\alpha}_i(\beta)$, even as T grows, (2) the data dependence between $\hat{\beta}$ and $\hat{\alpha}_i$, and (3) the variance of $\hat{\alpha}_i$.

Under general conditions, even as $n, T \to \infty$ and $n/T \to \rho$,

$$\begin{aligned} \sqrt{NT}(\hat{\beta}_{ML} - \beta_0) &= \sqrt{NT}(\hat{\beta}_{ML} - \beta_T) + \sqrt{\frac{n}{T}}B + O_p(\sqrt{\frac{n}{T^3}}) \\ &\xrightarrow{d} N(B\sqrt{\rho}, \Omega) \end{aligned} \tag{1.13}$$

Thus, the MLE estimator suffers from an asymptotic bias, even when both n and T grow at the same rate.

There is a vast literature on bias correction in panel models, which includes both Bayesian and frequentist methods. Some works in this literature will be reviewed here. For a more detailed account, see Arellano and Hahn (2005).

Lancaster (2002) proposes a Bayesian method to deal with the incidental parameter problem which entails reparameterizing the fixed effect such that the contribution to the likelihood by each individual can be factored out as follows:

$$\mathcal{L}_i(\alpha_i, \beta) = \mathcal{L}_{1i}(\kappa_i)\mathcal{L}_{2i}(\beta), \tag{1.14}$$

where κ_i is the reparameterized fixed effects. Now the log-likelihood becomes additively separable in κ_i and β.

$$l_i(\alpha_i, \beta) = l_{1i}(\kappa_i) + l_{2i}(\beta) \tag{1.15}$$

This orthogonal reparametrization renders the moment conditions for β, i.e. the first derivative of the log-likelihood with respect to β, independent of the reparameterized fixed effects, κ_i. Hence, the incidental parameter problem subsides and a $\sqrt{n}$-consistent estimator is achieved. Wouterson (2002) generalizes the method in Lancaster (2002) and provides a proof that the order of the bias of the corrected estimator of the common parameter reduces to $O(T^{-2})$. The orthogonalization approach to deal with the problem was mainly geared toward linear dynamic models, where fixed effects need not be estimated. However, in the nonlinear case, estimates of the fixed effects are needed to obtain estimates of average partial effects, which are more meaningful objects than the common parameters.

Arellano and Bonhomme (2009) further generalize Lancaster's approach to a wider class of models, characterize bias-reducing priors, and find that, in the absence of an orthogonal reparametrization, bias-reducing priors have to depend on the data. They also find that bias-reducing priors for common parameters are different from bias-reducing priors for the average partial effect.

Hahn and Newey (2004) propose a Jackknife as well as an analytical bias-corrected estimator (BCE) for the nonlinear panel model with exogenous regressors. The Jackknife method is a subsampling technique, which results in the fol-

lowing estimator,

$$\hat{\beta}_{JK} \;=\; T\hat{\beta}_{ML} - (T-1)\sum_{t=1}^{T}\frac{\hat{\beta}_{ML,(t)}}{T}, \tag{1.16}$$

whereby $\hat{\beta}_{ML}$ and $\hat{\beta}_{ML,t}$ are MLE estimators using all T observations and using the leave-t-out subsample, respectively. The analytical BCE, $\hat{\beta}_{BC}$, is of the form,

$$\hat{\beta}_{BC} \;=\; \hat{\beta}_{ML} - \frac{\bar{B}(\hat{\beta}_{ML})}{T}, \tag{1.17}$$

where $\bar{B}(\hat{\beta}_{ML})$ is an estimator of the product of the expected first-order condition with respect to β and the inverse of the Jacobian thereof, using Bartlett identities.[9] Using large-n, large-T asymptotics, Hahn and Newey (2004) shows that the MLE estimator is asymptotically biased and that the two proposed BCEs of the common parameter are asymptotically normal with zero mean. Hahn and Kuersteiner (2011) propose an analytical BCE for common parameters in dynamic nonlinear panel data models, similar to the analytical BCE in Hahn and Newey (2004). The BCE in Hahn and Kuersteiner (2011) is consistent and asymptotically normal. For dynamic binary choice models, Fernandez-Val (2009) proposes bias-corrections for the common parameter and the average partial effect that uses estimates of expected quantities rather than observed quantities, which leads to some improvement in the finite-sample properties compared to other BCEs.

Chapter 2 falls under this literature and shows how shrinkage can achieve a higher-order mean-square error improvement.

[9]There is a number of other analytical BCEs proposed in Hahn and Newey (2004), but we will only use this one for the following analysis.

1.5.2 Nonparametric Identification of Nonseparable Panel Data Models

In the literature on nonparametric identification, we assume that T is fixed. This literature is divided into two main approaches. The first focuses on a particular object of interest, such as in Chernozhukov et al. (2013), Hoderlein and White (2012), and Bester and Hansen (2009). Chapter 3 gives a unifying framework for this literature and relates it to the other approach in the literature which is the classical identification approach as in Evdokimov (2010), and Evdokimov (2011). These papers attempt to identify all structural objects including the conditional distribution of unobservables.

Chernozhukov et al. (2013), Hoderlein and White (2012) and Bester and Hansen (2009) focus on the identificaiton and estimation of average partial effects. Chernozhukov et al. (2013)'s model admits discrete regressors and discusses identification of average and quantile effects. Their basic model is

$$Y_{it} = \phi(X_{it}, \mathcal{A}_i, \mathcal{U}_{it}), \tag{1.18}$$

where X_{it} is a vector of discrete regressors and are all part of the treatment, i.e. there are no control variables in this model. For fixed T, we can only point-identify the effect of the treatment x to $\tilde{x}$ for individuals for whom X_{it} takes on both x and $\tilde{x}$ at some point during the time series. The authors call this quantity the *identified effect*, δ. The following example shows how the identified effect is estimated in that model.

Example 1.5.1 *Nonparametric Panel Data Model with Scalar Binary Regressor*
Let $Y_{it} = \phi(X_{it}, \mathcal{A}_i, \mathcal{U}_{it})$, where X_{it} is a scalar binary regressor. The estimator for

the identified effect of changing X from 0 to 1 is given by

$$\hat{\delta} = \sum_{i=1}^{n_I} \left(\frac{\sum_{t=1}^{T} X_{it} Y_{it}}{\sum_{t=1}^{T} X_{it}} - \frac{\sum_{t=1}^{T} (1 - X_{it}) Y_{it}}{\sum_{t=1}^{T} (1 - X_{it})} \right) / n_I \tag{1.19}$$

where n_I is the number of individuals for whom the effect of interest is identified. In order to understand the intuition behind the above quantity we can rewrite it as follows

$$\hat{\delta} = \sum_{i=1}^{n_I} \left(\bar{Y}_{i1} - \bar{Y}_{i0} \right) / n_I, \tag{1.20}$$

$\bar{Y}_{ix}$ is the average of Y_{it} for individual i for periods where $X_{it} = x$. This is a very intuitive way of estimating the identified effect.

Since any multidimensional treatment with discrete regressors can be written as the change from a scalar binary variable to another, the above can be generalized for a k-dimensional treatment vector, x to $\tilde{x}$, where we use two binary variables, $1\{X_{it} = x\}$ and $1\{X_{it} = \tilde{x}\}$. The example shows the importance of panel data for the treatment effect literature, since we are able to observe an individual in both 'states' of the world, treatment $\tilde{x}$ and non-treatment x. A key ingredient for identification that Chernozhukov et al. (2013) point out clearly is *time homogeneity*, $\mathcal{U}_{it}|X_i, \mathcal{A}_i \stackrel{d}{=} \mathcal{U}_{i1}|X_i, \mathcal{A}_i$. For practical purposes, this is an extremely strong assumption, especially since we typically have a small number of years that we are averaging over. Chernozhukov et al. (2013)'s main contribution points out that for fully nonseparable panel data models, time homogeneity is key to identifying average and quantile treatment effects for finite T.

Hoderlein and White (2012)'s object of interest is the local average structural derivative, which is the local average effect of a continuous variable. The model is similar to (1.1)

$$Y_{it} \quad = \quad \phi(X_{it}, Z_{it}, \mathcal{A}_i, \mathcal{U}_{it}), \tag{1.21}$$

where X_{it} is a vector of continuous variables. Hence, one may look at this paper as the continuous-variable counterpart to Chernozhukov et al. (2013). The assumption that corresponds to the time homogeneity assumption is a conditional (strict) stationarity assumption requiring $U_{i1}|Z_{i1}, \Delta Z_i, \mathcal{A}_i \stackrel{d}{=} U_{i2}|Z_{i2}, \Delta Z_i, \mathcal{A}_i$. The other key assumption in Hoderlein and White (2012) is conditional independence assumption that there exists $\epsilon > 0$ such that

$$\mathcal{U}_{it} \quad \perp \quad (\mathbb{I}\{\|\Delta X_i\| < \epsilon\}\Delta X, X_{i1})|\mathcal{A}_i, \Delta Z_i = 0, Z_{i1}, t = 1, 2. \tag{1.22}$$

Under the conditional stationarity and independence assumptions together with some further regularity conditions, Hoderlein and White (2012) show that the local average structural derivative at $X = x$ and $Z = z$ is identified for the stayers, i.e. for the subpopulation whose Z and X does not change over time.

Bester and Hansen (2009) establish identification of local average partial effects of a scalar continuous regressor in a nonparametric correlated random effects framework. It seems that a key advantage of the correlated random effects specification is that Bester and Hansen (2009) are able to allow for time heterogeneity in their model. Their model is given by the following,

$$Y_{it}|X_i, \mathcal{A}_i \quad \sim \quad G_t(Y_{it}|X_i, \mathcal{A}_i) = G_t(Y_{it}|X_{it}, \mathcal{A}_i). \tag{1.23}$$

The identifiable object of interest here is the following

$$b_t(X_i) \quad = \quad \int_{\mathcal{A}} \left(\frac{\partial}{\partial X_{it}} E[m(Y_{it})|X_i, \mathcal{A}_i] \right) dQ(\mathcal{A}_i|X_i), \tag{1.24}$$

where we seek to measure the average effect of X on $m(Y)$. Note that the integrand is the effect while fixing $\mathcal{A}_i$, whereas $b_t(X_i)$ is the effect averaged over $\mathcal{A}_i|X_i$. One interpretation of the latter effect is that it is the effect of the average individual with observables X_i, Z_i over the unobservable effect $\mathcal{A}_i$. This is what Hsiao (2003) and Arellano (2003) termed marginal inference, where $\mathcal{A}_i$ is integrated out. Recall that in the linear model such a distinction was not made, because both effects were the same quantity equal to the common parameter β. In a fully nonseparable model, in order to integrate out $\mathcal{A}_i$, one needs to impose some restrictions on the distribution of $\mathcal{A}_i|X_i$. Bester and Hansen (2009) impose index restrictions that allow the identification of $b_t(X_i)$. However, they only assume the existence of such restrictions without assuming the distribution or the index function to be known, as is typically assumed in correlated random effects models, such as in Mundlak (1978) and Chamberlain (1984), where $\mathcal{A}_i|X_i \sim N(X_i'\Pi, \sigma^2_{\mathcal{A}})$.

Evdokimov (2010) builds on Altonji and Matzkin (2005) and proposes a method to identify all structural objects, i.e. the structural function and the distribution of unobservables. This of course requires stronger assumptions than the more focused identification approach. To start with, the model assumes the additive separability of $\mathcal{U}_{it}$ as follows

$$Y_{it} \quad = \quad \mu(X_{it}, Z_{it}, \mathcal{A}_i) + \mathcal{U}_{it}. \tag{1.25}$$

The identification of the structural function $\mu(X_{it}, Z_{it}, \mathcal{A}_i)$ and the conditional distribution of $\mathcal{U}_{it}|X_{it}, Z_{it}$ under both the random effects and the fixed effects assumptions is possible by imposing monotonicity of μ in $\mathcal{A}_i$ as well as independence assumptions on $\mathcal{U}_{it}|X_{it}, Z_{it}$ of $X_{i(-t)}$, $Z_{i(-t)}$, $\mathcal{U}_{i(-t)}$ and $\mathcal{A}_i$.[10] The approach here can be applied to both continuous and discrete regressors, however there are some restrictions on the regressors. For instance, regressors that are time-varying with probability one, such as time effects, time trends or age, have to enter the structural function separably from $\mathcal{A}_i$. Finally, it is important to note that the identification of μ and the distribution of $\mathcal{A}_i$ and $\mathcal{U}_{it}$ allows us to identify interesting objects of interest such as the effect of a change in x on the median of $\mathcal{A}_i$. This paper illustrates how stronger assumptions on the structure of the model may help us identify more objects that may be interesting for application. Evdokimov (2011) has a similar goal in mind but attempts to allow for some time heterogeneity.

Chapter 3 will revisit this literature and follows the focused identification approach. It starts from a general data-generating process that exhibits both time and individual heterogeneity and tries to characterize the different strategies that can be used to identify an average partial effect. One of the key motivations behind the identification section of Chapter 3 is to provide a unifying framework for the existing literature.

[10]Note that $X_{i(-t)} \equiv (X_{i1}, X_{i2}, ..., X_{i,t-1}, X_{i,t+1}, ..., X_i)$

Chapter 2

Shrinkage and Higher-Order Improvements in Nonlinear Panel Data Models with Fixed Effects

2.1 Introduction

Nonlinear panel data models with fixed effects constitute an important example of the incidental parameter problem in econometrics. Due to the fixed effects assumption, the dimension of the nuisance parameters increases with the sample size leading to asymptotically biased maximum likelihood estimators (MLEs) of common parameters and other objects of interest, such as the average partial effects (APEs). Nonlinear models encompass many useful models in empirical applications, such binary choice models, count data models and duration models. As economic panel data tend to be observational, controlling for individual heterogeneity is key to identifying the effects of interest. Unfortunately, it becomes a double-edged sword in nonlinear models due to the incidental parameter problem.

For panel probit models, the Chamberlain-Mundlak device (Mundlak, 1978; Chamberlain, 1984; Wouterson, 2002) is a quick-fix, whereby each fixed effect, α_i can be transformed into a random effect, c_i assuming that $\alpha_i = \delta\bar{x}_i + c_i$, where $\bar{x}_i$ is the average of the regressor for individual i across the time horizon. Assuming that c_i is not correlated with x_{it}, the regressor, it is treated as a non-confounding disturbance term. This manipulation only necessitates the estimation of the common parameters, and accordingly the total number of parameters does not increase with sample size. Chamberlain (1984) discusses the linearization of the relationship between α_i and x_{it} and points out that despite its convenience there is a strong assumption underneath this manipulation. It not only assumes a linear relationship between α_i and x_{it}, but also assumes that there is no unobservable ξ_{it} that can lead to a systematic relationship between α_i and x_{it} — a shortcoming of the random effects assumption.

Empirical work using the Chamberlain-Mundlak device includes Papke and Wooldridge (2008) and Christiansen, Joensen, and Rangvid (2008). An interesting application of count data models to the management of fisheries is Zhang (2008). Due to the incidental parameter problem in the case of the Negative Binomial distribution, Zhang (2008) employs a bootstrap variant of the jackknife estimator proposed in Hahn and Newey (2004) (hereinafter, HN04). [1]

The work in the theoretical literature has been geared toward bias correction of the MLE, the moment equations, and the objective function. The paper at hand falls under the first category. A bias-corrected estimator (BCE), $\hat{\theta}_{BC}$, is constructed by subtracting an estimator of the bias term from the MLE as follows,

$$\hat{\theta}_{BC} \;=\; \hat{\theta}_{ML} - \frac{\hat{B}}{T}, \tag{2.1}$$

where $\hat{B}$ may be an automatic or analytical estimator of the bias. HN04 propose analytical as well as jackknife estimators of the bias. The paper at hand proposes a shrinkage estimator, which consists of a weighted average of the MLE and analytical BCE, where the weights are chosen to minimize the mean square error (MSE) criterion. The shrinkage estimator leads to a higher-order MSE improvement over the analytical BCE, but is first-order equivalent to the analytical BCE. Hence, it maintains favorable properties such as consistency and asymptotic normality. Monte Carlo evidence shows that the shrinkage estimator for the common parameter leads to further bias reduction and improved coverage probabilities in finite samples, especially for small T. As for the APE, the shrinkage estimator does not perform very well for small T. We conjecture that this is due to the

[1] Another interesting application of panel probit models is Debaere and Mostashari (2005).

reliance of the shrinkage estimator on the asymptotic distribution to estimate the shrinkage weights, and numerical evidence shows that this distribution is a poor approximation of the finite-sample distribution.

The paper is organized as follows. It describes the incidental parameter problem in nonlinear panel data models and reviews some relevant work in the literature on bias correction (Section 2.2). Section 2.3 motivates the shrinkage approach proposed here and describes the shrinkage estimator for the common parameter and APE. Monte Carlo evidence is provided to show its finite sample performance. Section 2.4 concludes.

2.2 The Incidental Parameter Problem and Bias Correction

The estimation of nonlinear panel data models with fixed effects is a classic example of the incidental parameter problem introduced in Neyman and Scott (1948). The simplest example in economics is that of a binary choice panel data model. So the problem will be illustrated using this example. However, the aim of the project is to deal with nonlinear panel data models at large.

Let y_{it}^* be a latent variable, characterized by

$$y_{it}^* = x_{it}'\theta + v_{it}, \tag{2.2}$$

where we observe the binary choice, $y_{it} = 1_{\{y_{it}^* \geq 0\}}$, and x_{it}. As panel data track different individuals i across time t, it is often the case that the (unobservable)

disturbance term v_{it} includes a time-invariant, individual-specific term, which represents unobservable individual characteristics. Hence, one can rewrite v_{it} and y_{it} as follows:

$$v_{it} = \alpha_i + u_{it}, \tag{2.3}$$

$$y_{it} = 1_{\{x'_{it}\theta+\alpha_i+u_{it}\geq 0\}}, \tag{2.4}$$

where α_i is termed the individual effect. If α_i is uncorrelated with the regressors, then one could ignore its presence and estimate the model using maximum likelihood. This is the motivation behind random effects estimation.

However, if there is any correlation between individual heterogeneity and the regressors, then ignoring the individual effect confounds estimates of common parameters and APEs. Hence, fixed effects estimation would be appropriate, and α_i is dealt with as a parameter to be estimated. As a result, there are n fixed effects to be estimated aside from the common parameters. As the sample size tends to infinity, so does the number of parameters to be estimated. This is how controlling for unobservable individual heterogeneity in nonlinear panel data models leads to the incidental parameter problem.

Let α_i be a fixed effect. Given a panel data set of n individuals and T time periods and assuming a distribution for u_{it}, one may use MLE to estimate θ as follows:

$$\hat{\theta}_{ML} \equiv \arg\max_{\theta} \sum_{i=1}^{n}\sum_{t=1}^{T} \log f(z_{it};\theta,\hat{\alpha}_i(\theta)), \tag{2.5}$$

where $\hat{\alpha}_i(\theta) \equiv \arg\max_\alpha \sum_{t=1}^{T} \log f(z_{it};\theta,\alpha_i)$ and $z_{it} \equiv (y_{it}, x_{it}^T)^T$. Now letting $n \to \infty$ with fixed T,

$$\hat{\theta} \xrightarrow{p} \theta_{ML,T} \equiv \arg\max_\theta \bar{E}[\sum_{t=1}^{T} \log f(z_{it};\theta,\hat{\alpha}_i(\theta))/T] \tag{2.6}$$

where $\bar{E}[m(Z_{iT};\theta,\alpha_i)] = \lim_{N\to\infty} n^{-1} \sum_{i=1}^{n} E[m(Z_{iT};\theta,\alpha_i)]$, where $Z_{iT} \equiv (z_{i1}, ..., z_{iT})$.

Let θ_0 denote the true parameter. Note that $\theta_{ML,T} \neq \theta_0$. However, $\lim_{T\to\infty} \theta_{ML,T} = \theta_0$. Expanding $\theta_{ML,T}$ in $\hat{\alpha}_i(\theta)$ around α_i leads to a bias of order $1/T$,

$$\theta_{ML,T} = \theta_0 + \frac{B}{T} + O\left(\frac{1}{T^2}\right). \tag{2.7}$$

HN04 point out three sources of bias in this setting: (1) the bias of $\hat{\alpha}_i(\theta)$, even as T grows, (2) the data dependence between $\hat{\theta}$ and $\hat{\alpha}_i$, and (3) the variance of $\hat{\alpha}_i$.

To continue the analysis, we will need to introduce more notation. Let f$(z_{it};\theta,\alpha)$ be the contribution of the i^{th} individual in the t^{th} period.

$$u_{it}(\theta,\alpha) \equiv \frac{\partial}{\partial\theta} \log \text{f}(z_{it};\theta,\alpha), \tag{2.8}$$

$$v_{it}(\theta,\alpha) \equiv \frac{\partial}{\partial\alpha_i} \log \text{f}(z_{it};\theta,\alpha), \tag{2.9}$$

$$v_{it\alpha}(\theta,\alpha) \equiv \frac{\partial v_{it}(\theta,\alpha)}{\partial\alpha}. \tag{2.10}$$

Let g_{it} denote $g_{it}(\theta_0, \alpha_0)$.

$$\begin{aligned} U_{it} &\equiv u_{it} - v_{it}\frac{\sum_{s=1}^{T} u_{it}v_{it}}{\sum_{s=1}^{T} v_{is}^2}, && (2.11)\\ V_{2it} &\equiv v_{it}^2 + v_{it\alpha}, && (2.12)\\ [\bar{E}[\mathcal{I}_i]]^{-1} &= \bar{E}[U_{it}U_{it}^T]. && (2.13) \end{aligned}$$

Now HN04 impose the following conditions:

Assumption 2.2.1 *$n, T \to \infty$ such that $n/T \to \rho$, where $0 < \rho < \infty$.*

Assumption 2.2.2 *(i)The function $\log f(.;\theta,\alpha)$ is continuous in $(\theta,\alpha) \in \mathcal{Y}$; (ii) the parameter space $\mathcal{Y}$ is a compact subset of $\mathcal{R}^{k+1}$;*

Assumption 2.2.3 *For each $\eta > 0$,*

$$\inf_i [G_{(i)}(\theta_0, \alpha_{i0}) - \sup_{\{(\theta,\alpha):|(\theta,\alpha)-(\theta_0,\alpha_{i0})|>\eta\}} G_{(i)}(\theta,\alpha)] > 0,$$

where $\hat{G}_{(i)}(\theta,\alpha_i) \equiv T^{-1}\sum_{t=1}^{T}\log f(z_{it};\theta,\alpha_i) \equiv T^{-1}\sum_{t=1}^{T} g(z_{it};\theta,\alpha_i)$, $G_{(i)}(\theta,\alpha_i) \equiv E[\log f(z_{it};\theta,\alpha_i)]$.

Assumption 2.2.4 *(i) There exists some $M(z_{it})$ such that $|\frac{\partial^{m_1+m_2}\log f(z_{it};\theta,\alpha_i)}{\partial\theta^{m_1}\partial\alpha_i^{m_2}}| \leq M(z_{it})$, $0 \leq m_1 + m_2 \leq 1, ..., 6$*
and $\sup_i E[M(z_{it})^Q] < \infty$ for some $Q > 64$; (ii)$\bar{E}[\mathcal{I}_i] > 0$; (iii) $\min_i E[v_{it}^2] > 0$.

See Section 2.5 for definitions of U_{it}, v_{it}, $\mathcal{I}_i$, and $\bar{B}$. Under conditions 1-4, HN04 show that

$$\begin{aligned} \sqrt{NT}(\hat{\theta} - \theta_0) &= \sqrt{NT}(\hat{\theta} - \theta_{ML,T}) + \sqrt{\frac{N}{T}}B + O_p(\sqrt{\frac{N}{T^3}})\\ &\xrightarrow{d} N(B\sqrt{\rho}, [\bar{E}[\mathcal{I}_i]]^{-1}) && (2.14)\\ \text{where}\quad B &\equiv -\frac{1}{2}[\bar{E}[\mathcal{I}_i]]^{-1}\bar{E}[E[U_{it}V_{2it}]/E[v_{it}]^2]. && (2.15) \end{aligned}$$

Thus, the MLE suffers from an asymptotic bias, even when both N and T grow at the same rate.

There is a vast literature on bias correction in panel data models, which includes both Bayesian and frequentist methods. Some works in this literature will be reviewed here. For a more detailed account, see Arellano and Hahn (2005).

Lancaster (2002) proposes a Bayesian method to deal with the incidental parameter problem which entails reparameterizing the fixed effect, α_i, such that the contribution to the likelihood by the reparametrized individual effect, κ_i, can be factored out as follows:

$$\mathcal{L}_i(\alpha_i, \theta) \quad = \quad \mathcal{L}_{1i}(\kappa_i)\mathcal{L}_{2i}(\theta). \tag{2.16}$$

Now the log-likelihood becomes additively separable in κ_i and θ.

$$l_i(\alpha_i, \theta) \quad = \quad l_{1i}(\kappa_i) + l_{2i}(\theta) \tag{2.17}$$

This orthogonal reparameterization renders the moment conditions for θ, i.e. the first derivative of the log-likelihood with respect to θ, independent of κ_i. Hence, the incidental parameter problem subsides and a $\sqrt{n}$-consistent estimator is achieved. Wouterson (2002) generalizes the method in Lancaster (2002) and provides a proof that the order of the bias of the corrected estimator of the common parameter reduces to $O(T^{-2})$. The orthogonalization approach to deal with the problem was mainly geared toward linear dynamic models, where fixed effects need not be estimated. However, in the nonlinear case, estimates of the fixed effects are needed to obtain estimates of APEs, which are more meaningful objects than the com-

mon parameters. Arellano and Bonhomme (2009) further generalize Lancaster's approach to a wider class of models, characterize bias-reducing priors, and find that, in the absence of an orthogonal reparametrization, bias-reducing priors have to depend on the data.

HN04 propose a jackknife as well as an analytical BCE for the nonlinear panel data model with exogenous regressors. The jackknife method is a subsampling technique, which results in the following estimator,

$$\begin{aligned} \hat{\theta}_{JK} &= T\hat{\theta}_{ML} - (T-1)\sum_{t=1}^{T}\frac{\hat{\theta}_{ML,(t)}}{T} \\ &= \hat{\theta}_{ML} - (T-1)\left(\sum_{t=1}^{T}\frac{\hat{\theta}_{ML,(t)}}{T} - \hat{\theta}_{ML}\right) \end{aligned} \tag{2.18}$$

whereby $\hat{\theta}_{ML}$ and $\hat{\theta}_{ML,(t)}$ denote the MLEs using all T observations and using the leave-t-out subsample, respectively. The analytical BCE, $\hat{\theta}_{BC}$, is of the form,

$$\hat{\theta}_{BC} = \hat{\theta}_{ML} - \frac{\bar{B}(\hat{\theta}_{ML})}{T}, \tag{2.19}$$

where $\bar{B}(\hat{\theta}_{ML})$ is the sample analogue of equation (2.15), see Section 2.5 for a detailed definition.[2] Using large-N, large-T asymptotics, HN04 shows that the MLE is asymptotically biased and that the two proposed BCEs of the common parameter are asymptotically normal with zero mean. HN04 also proposes iterating the analytical estimator of the bias, as follows

$$\hat{\theta}^{k}_{BC} = \hat{\theta}_{ML} - \frac{\bar{B}(\hat{\theta}^{k-1}_{BC})}{T}, \tag{2.20}$$

[2]There is a number of other analytical BCEs proposed in HN04, but we will only use this one for the following analysis.

until convergence. Hahn and Kuersteiner (2011) propose an analytical BCE for common parameters in dynamic nonlinear panel data models, similar to the analytical BCE in HN04. Fernandez-Val (2009) applies the analytical bias correction of Hahn and Kuersteiner (2011) to dynamic binary choice models and uses estimators of expected quantities instead of observed quantities. This manipulation leads to an improvement in the finite-sample properties for both common parameters and APEs.

Noting that fixed effects still restrict the form of unobservable heterogeneity, some recent work in the literature generalizes the notion of fixed effects to better suit nonlinear models. For dynamic binary choice models, Browning and Carro (2007) propose a BCE and a minimum integrated MSE (MIMSE) estimator, both allowing for state-dependent heterogeneity in addition to the traditional fixed effects. Furthermore, Hoderlein and White (2012) uses a generalization of first-differencing to identify local average structural derivatives (LASD) in nonseparable panel data models, which allow for an arbitrary relationship between regressors and individual heterogeneity. Evdokimov (2010) examines nonparametric identification of structural functions under random effects and correlated random effects assumptions and proposes a nonparametric estimation strategy.

The paper at hand will focus on the analytical BCEs for the common parameter and the APE proposed in HN04 in order to illustrate the possible benefits of shrinkage in the traditional fixed effects setup. The approach can be extended to other analytical BCEs, such as those proposed in Hahn and Kuersteiner (2011) and Fernandez-Val (2009).

2.3 A Shrinkage Approach

Since Stein (1955) and James and Stein (1961) showed that the MLE for the mean of a multivariate normal can be improved upon in terms of squared loss, when the dimension of the normal vector is greater than 2, shrinkage methods have been widely used to improve finite sample performance of consistent estimators. The shrinkage approach proposed here will be in the same vein, and the resulting estimator will be equivalent to the analytical BCE asymptotically. The motivation here is to ensure that the shrinkage estimator is consistent and asymptotically normal. Kim and White (2000) show that one can propose shrinkage estimators that are asymptotically biased but achieve minimum asymptotic risk. The approach taken in this paper will lead to higher-order improvements, but no first-order gain.

2.3.1 A Shrinkage Estimator for the Common Parameter

The shrinkage estimator proposed here consists of a weighted average of the MLE and the analytical BCE from HN04,

$$\begin{aligned} \hat{\theta}_{SH} &= \Lambda\hat{\theta}_{BC} + (I - \Lambda)\hat{\theta}_{ML} \\ &= \hat{\theta}_{ML} - \frac{\Lambda\bar{B}(\hat{\theta}_{ML})}{T}. \end{aligned} \tag{2.21}$$

Λ is chosen to minimize the MSE of $\hat{\theta}_{SH}$. For a vector u, let $\|u\|_T \equiv uu^T$. For illustration purposes, I will assume that the MSE of $\hat{\theta}_{ML}$ exists. This is a strong assumption but it is only for illustration purposes here. In Proposition 3.1, we

work with the MSE up to order $O(T^{-3})$.

$$\begin{aligned}\Lambda_{nT} &= \arg\min E\|\hat{\theta}_{SH}(\Lambda) - \theta_0\|_T \\ &= E\left[(\hat{\theta}_{ML} - \theta_0)\frac{\bar{B}(\hat{\theta}_{ML})^T}{T}\right]\left\{E\left\|\frac{\bar{B}(\hat{\theta}_{ML})}{T}\right\|_T\right\}^{-1}. \end{aligned} \tag{2.22}$$

Let $\hat{\theta}_{SH} \equiv \hat{\theta}_{SH}(\Lambda_{nT})$. One can show the following relationship between the MSE of $\hat{\theta}_{SH}$ and $\hat{\theta}_{ML}$.

$$\begin{aligned} & E\|\hat{\theta}_{SH} - \theta_0\|_T \\ = \; & E\|\hat{\theta}_{ML} - \theta_0\|_T \\ - \; & E\left[(\hat{\theta}_{ML} - \theta_0)\frac{\bar{B}(\hat{\theta}_{ML})^T}{T}\right]\left\{E\left\|\frac{\bar{B}(\hat{\theta}_{ML})}{T}\right\|_T\right\}^{-1} \\ & \times E\left[\frac{\bar{B}(\hat{\theta}_{ML})}{T}(\hat{\theta}_{ML} - \theta_0)^T\right], \end{aligned} \tag{2.23}$$

where the last term on the right-hand side is positive definite. For the full derivation, see Section 2.5.

Contrasting (2.23) with the following relationship between the MSE of $\hat{\theta}_{BC}$ and $\hat{\theta}_{ML}$ helps illustrate what shrinkage helps achieve.

$$\begin{aligned} E\|\hat{\theta}_{BC} - \theta_0\|_T &= E\|\hat{\theta}_{ML} - \theta_0\|_T - E\left[(\hat{\theta}_{ML} - \theta_0)\frac{\bar{B}(\hat{\theta}_{ML})^T}{T}\right] \\ &- E\left[(\hat{\theta}_{ML} - \theta_0)\frac{\bar{B}(\hat{\theta}_{ML})^T}{T}\right]^T + E\left\|\frac{\bar{B}(\hat{\theta}_{ML})}{T}\right\|_T. \end{aligned} \tag{2.24}$$

Note that the analytical BCE reduces the first-order bias without increasing the first-order variance. However, the uncertainty due to estimating the bias leads to higher-order terms. The net effect of these terms depends on the relative magni-

tude of the last three terms on the right-hand side of (2.24).

Rilstone, Srivastava, and Ullah (1996) point out that if the bias term were known, i.e. the bias estimator was a constant, B, then

$$E\|\hat{\theta}_{BC} - \theta_0\|_T = E\|\hat{\theta}_{ML} - \theta_0\|_T - \frac{BB^T}{T^2} \tag{2.25}$$

This is equivalent to the case, where the last two expected quantities on the right-hand side of (2.24) were equal and

$$E\|\hat{\theta}_{BC} - \theta_0\|_T = E\|\hat{\theta}_{ML} - \theta_0\|_T - E\left\|\frac{\bar{B}(\hat{\theta}_{ML})}{T}\right\|_T. \tag{2.26}$$

If (2.26) were true, then Λ_{nT} would be equal to the identity matrix to minimize the MSE of the shrinkage estimator.[3] It would also follow that the right-hand side of (2.23) and (2.26) would be equal, again implying that Λ_{nT} is equal to the identity matrix. However, this is unlikely to be the case in finite samples. Hence, if the determinant of the numerator in (2.22) is larger than the denominator, then the determinant of the weighting matrix on the bias estimator will be larger than 1, and vice versa. To illustrate the intuition behind putting a larger or smaller weight on the bias estimator, suppose that the bias estimator perfectly estimated the actual bias of the MLE,[4] i.e.

$$\frac{\bar{B}(\hat{\theta}_{ML})}{T} = \hat{\theta}_{ML} - \theta_0 \tag{2.27}$$

for all $\hat{\theta}_{ML}$. In that case, Λ_{nT} would always be equal to identity and the shrinkage

[3] In the limit experiment, the bias term is indeed constant and hence $\Lambda_0 \equiv \lim_{n,T\to\infty} \Lambda_{nT}$ equals the identity matrix.

[4] Note that a constant bias term is not perfectly accurate.

estimator would be equal to the anaytical BCE for any realization of the MLE. Now to illustrate what happens in practice, let

$$\frac{\bar{B}(\hat{\theta}_{ML})}{T} + \epsilon \;\; = \;\; \hat{\theta}_{ML} - \theta_0, \tag{2.28}$$

where ϵ reflects the estimation error of the bias estimator. If the bias estimator over-estimates the magnitude of the actual bias of the MLE, i.e. the covariance of the bias estimator and ϵ is negative,[5], then the determinant of Λ_{nT} will be less than 1 in order to reduce the magnitude of bias correction for the shrinkage estimator. The contrary is also true; if the bias estimator tends to under-estimate the magnitude of the actual bias of the MLE, then the determinant of Λ_{nT} will be greater than 1 in order to increase the magnitude of bias correction for the shrinkage estimator. Hence, the shrinkage estimator can be viewed as a modified or adjusted BCE.

The above explains the intuition behind the shrinkage weighting matrix. However, our goal is not to improve on the MLE, since the analytical BCE already improves on it. The following proposition explains how the shrinkage estimator improves on the analytical BCE in terms of higher-order MSE.

Proposition 2.3.1 *Let* $\theta_{ML,T} \equiv plim_{n\to\infty}\hat{\theta}_{ML}$, $\theta_{BC,T} \equiv plim_{n\to\infty}\hat{\theta}_{BC}$, $\theta_{SH,T} \equiv plim_{n\to\infty}\hat{\theta}_{SH}$, $\overline{\bar{B}(\hat{\theta}_{ML})} = plim_{n\to\infty}\bar{B}(\hat{\theta}_{ML})$ *and* $\Lambda_T \equiv lim_{n\to\infty}\Lambda_{nT}$. *Let conditions*

[5] $cov\left(\hat{\theta}_{ML} - \theta_0, \frac{\bar{B}(\hat{\theta}_{ML})}{T}\right) = var\left(\frac{\bar{B}(\hat{\theta}_{ML})}{T}\right) + cov\left(\frac{\bar{B}(\hat{\theta}_{ML})}{T}, \epsilon\right)$.

2-4 hold and let MSE denote the mean squared error matrix to order $O(T^{-3})$

$$\begin{aligned} MSE(\theta_{BC,T}) &= MSE(\theta_{ML,T}) - \frac{BB^{'}}{T^2} + \frac{E[(\overline{\bar{B}(\hat{\theta}_{ML})} - B)(\overline{\bar{B}(\hat{\theta}_{ML})} - B)^{'}]}{T^2} \\ &\equiv MSE(\theta_{ML,T}) - \frac{BB^{'}}{T^2} + \frac{\Delta_2}{T^3}, where|\Delta_2| > 0, \\ MSE(\theta_{SH,T}) &= MSE(\theta_{BC,T}) - \frac{\Delta_1}{T^3}, where|\Delta_1| > 0, \\ |\Delta_2 - \Delta_1| &> 0. \end{aligned}$$

The proof is given in Section 2.5. From the previous proposition, we learn that the analytical BCE reduces the first-order bias, however, the uncertainty that results from estimating the bias leads to a positive definite $O(T^{-3})$ term in the MSE. Shrinkage is able to remove part of this term. To further illustrate the intuition behind how the shrinkage estimator achieves this improvement, it is useful to take a closer look at the numerator of the shrinkage weighting matrix, $E[(\hat{\theta}_{ML} - \theta_0)\bar{B}(\hat{\theta}_{ML})^{'}/T]$. This term exploits the joint distribution of the MLE and the bias estimator, which can be viewed as a joint distribution of a first-order term with a higher-order term,

$$\sqrt{nT}\begin{pmatrix} \hat{\theta}_{ML} - \theta_0 \\ \bar{B}(\hat{\theta}_{ML}) - B \end{pmatrix} = \begin{pmatrix} \sqrt{nT}\hat{\theta}_{ML} - \theta_0 \\ \sqrt{nT^3}\frac{(\bar{B}(\hat{\theta}_{ML}) - B)}{T} \end{pmatrix}.$$

This is not surprising since the estimation uncertainty due to $\bar{B}(\hat{\theta}_{ML})$ only has higher-order effects on the MSE. However, by exploiting this joint distribution, the shrinkage estimator is able to reduce the effects of uncertainty due to bias estimation.

A plug-in estimator of Λ_{nT}, $\hat{\Lambda}_{nT}$, can be obtained by applying mean-variance

decompositions of the MSE terms and estimating both mean and variance using the δ-method. It is important to note here that $\bar{B}(\theta)$ is a random function. Hence, the asymptotic validity of the application of the δ-method here is formally addressed in Lemma 2.5.2.

$$\begin{aligned}\hat{\Lambda}_{nT} &= \left\{\left\|\frac{\bar{B}(\hat{\theta}_{ML})}{T}\right\|_T + \hat{cov}\left(\hat{\theta}_{ML}-\theta_0, \frac{\bar{B}(\hat{\theta}_{ML})}{T}\right)\right\} \\ &\quad \times \left\{\left\|\frac{\bar{B}(\hat{\theta}_{ML})}{T}\right\|_T + \hat{var}\left(\frac{\bar{B}(\hat{\theta}_{ML})}{T}\right)\right\}^{-1},\end{aligned}$$

$$\begin{aligned}\text{where} \quad \hat{cov}\left(\hat{\theta}_{ML}-\theta_0, \frac{\bar{B}(\hat{\theta}_{ML})}{T}\right) &= \frac{1}{nT}\hat{E}_n[\mathcal{I}_i]^{-1}\bar{B}'(\hat{\theta}_{ML})^T, \\ \hat{var}(\bar{B}(\hat{\theta}_{ML})) &= \frac{1}{nT}\bar{B}'(\hat{\theta}_{ML})\hat{E}_n[\mathcal{I}_i]^{-1}\bar{B}'(\hat{\theta}_{ML})^T,\end{aligned}$$

where $\hat{E}_n[m(Z_{iT};\theta,\alpha_i)] \equiv (nT)^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T} m(Z_{it};\theta,\alpha_i)$.

A natural question that arises now is the effect of the estimation uncertainty due to estimating the shrinkage weighting on the MSE. It is important to note that this estimation uncertainty will lead to a $O(T^{-4})$ term. Numerical evidence can show whether in finite samples the two effects outweigh each other.

It has been noted above that the shrinkage estimator is first-order equivalent to the analytical BCE. The following proposition proves this result by showing that $\hat{\Lambda}_{nT} \xrightarrow{p} I_k$. The following conditions need to be imposed.

Assumption 2.3.1 *(i) There exists some $M(z_{it})$ such that*

$|\frac{\partial^{m_1+m_2}\log f(z_{it};\theta,\alpha_i)}{\partial\theta^{m_1}\partial\alpha_i^{m_2}}| \leq M(z_{it})$, $0 \leq m_1+m_2 \leq 1,...,6$

and $\sup_i E[M(z_{it})^Q] < \infty$ *for some* $Q > 64$*;*

(ii) there exists some $m(z_{it})$ *such that*

$|\frac{\partial^{m_3}\log f(z_{it};\theta,\alpha_i)}{\partial\alpha_i^{m_3}}| \geq m(z_{it})$, $m_3 = 1$

and $\inf_i E[m(z_{it})^2] > 0$*;*

(iii)$\bar{E}[\mathcal{I}_i] > 0$.

Assumption 2.3.2 $\Lambda = (\lambda_{ij}) \in \mathcal{L}$*, a compact subset of* $\mathbb{R}^{k\times k}$*, and there exists* M *such that* $|\lambda_{ij}| \leq M$ *for all* i, j, $\Lambda \in \mathcal{L}$.

Condition 5 is a slightly stronger variant of condition 4. Condition 6 is needed to show that the MSE of the feasible shrinkage estimator converges to the minimum MSE of $\hat{\theta}_{SH}(\Lambda)$. Let $\|.\|_F$ denote the Frobenius norm.

Proposition 2.3.2 *Under conditions 1-3 and 5-6,*

$$\hat{\Lambda}_{nT} \xrightarrow{p} I$$

and

$$\left\| E\|\hat{\theta}_{SH}(\hat{\Lambda}_{nT}) - \theta_0\|_T - \min_{\Lambda\in\mathcal{L}} E\|(\hat{\theta}_{SH}(\Lambda) - \theta_0\|_T \right\|_F = o(1)$$

Section 2.5 contains the proof. Inference on $\hat{\theta}_{SH}$ can be performed using the asymptotic variance estimator $\{\sum_{i=1}^n \sum_{t=1}^T U_{it}(\hat{\theta}_{SH})U_{it}(\hat{\theta}_{SH})^T\}$. The finite-sample performance of the shrinkage estimator for θ will be explored in the Monte Carlo study, Subsection 3.3.

2.3.2 Shrinkage Estimators for the APE

Unlike the common parameter, proposing a shrinkage estimator for the APE is not as straightforward and there are different avenues that one may take

to derive such an estimator. In order to discuss these avenues, let us list the MLE and analytical BCE, given in HN04, of the APE:

$$\begin{aligned}\hat{\mu}(\hat{\theta}_{ML}) &= \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} m(z_{it};\hat{\theta}_{ML},\hat{\alpha}_i(\hat{\theta}_{ML})),\\ \tilde{\mu}(\hat{\theta}_{BC}) &= \hat{\mu}(\hat{\theta}_{BC}) - \frac{\hat{\Delta}(\hat{\theta}_{BC})}{T},\end{aligned}$$

where $\hat{\Delta}$ is equivalent to $\bar{B}$ in the case of the common parameter. Note that z_{it} refers to the vector of regressors and regressand of individual i at time t.

Using the definitions in Section 2.5, let

$$\begin{aligned}\bar{\sigma}_i^2(\theta) &= T/\sum_{s=1}^{T} v_{is}^2(\theta),\\ \bar{\beta}_i(\theta) &= -\bar{\sigma}_i^2\sum_{s=1}^{T} v_{is}(\theta)V_{2it}(\theta)/2T,\\ \bar{\psi}_{it}(\theta) &= \bar{\sigma}_i^2(\theta)v_{it}(\theta).\end{aligned}$$

Let m_α and $m_{\alpha\alpha}$ denote the first and second order derivatives of m with respect to α. Now

$$\hat{\Delta}(\theta) = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\{m_\alpha(z_{it};\theta,\hat{\alpha}_i(\theta))[\bar{\beta}_i(\theta)+\bar{\psi}_{it}(\theta)] + m_{\alpha\alpha}(z_{it};\theta,\hat{\alpha}_i(\theta))\bar{\sigma}_i^2(\theta)/2\}$$

Example 2.3.1 *APE at $\bar{x}$ in a Binary Probit Model*

Let $y_{it} = 1_{\{x_{it}\theta+\alpha_i+u_{it}\geq 0\}}$, ϕ be the standard normal density, and $\bar{x}$ be the average

of x_{it}

$$\begin{aligned}\hat{\mu}_{ML} &= \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{ML}\phi(\bar{x}\hat{\theta}_{ML}+\hat{\alpha}_i(\hat{\theta}_{ML})) \\ \tilde{\mu}_{BC} &= \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{BC}\phi(\bar{x}\hat{\theta}_{BC}+\hat{\alpha}_i(\hat{\theta}_{BC}))-\frac{\hat{\Delta}(\hat{\theta}_{BC})}{T}.\end{aligned}$$

Let ϕ_{it} *denote* $\phi(x_{it}\theta+\hat{\alpha}_i(\theta))$ *and* Φ_{it} *denote* $\Phi(x_{it}\theta+\hat{\alpha}_i(\theta))$*, where* Φ *is the standard normal cumulative distribution function.*

$$\begin{aligned}m_\alpha(z_{it};\theta,\hat{\alpha}_i(\theta)) &= -\theta\phi_{it}[\bar{x}\theta+\hat{\alpha}_i(\theta)] \\ m_{\alpha\alpha}(z_{it};\theta,\hat{\alpha}_i(\theta)) &= -\theta\phi_{it}\{1-[\bar{x}\theta+\hat{\alpha}_i(\theta)]^2\}\end{aligned}$$

Now plugging the above into the definitions for $\bar{\sigma}_i^2(\theta)$ *and* $\bar{\beta}_i(\theta)$*, we can get an expression for* $\hat{\Delta}(\theta)$*. Note that as* m_α *only depends on* $\bar{x}$ *we do not need to include* ψ_{it} *as pointed out in Hahn and Newey (2004).*

Now there are two main ways that one may choose to take the shrinkage approach. First, one can mimic the approach taken for the common parameter, θ, where the resulting shrinkage estimator for the APE would be given as follows:

$$\tilde{\mu}_{SH,1}(\lambda_{\mu_1}) = \hat{\mu}(\hat{\theta}_{BC})-\lambda_{\mu_1}\frac{\hat{\Delta}(\hat{\theta}_{BC})}{T} \tag{2.29}$$

Here, λ_{μ_1} will be chosen by minimizing the MSE of the above estimator. The resulting $\lambda_{\mu_1,nT}$ is given by:

$$\lambda_{\mu_1,nT} = \frac{E[(\hat{\mu}(\hat{\theta}_{BC})-\mu_0)\hat{\Delta}(\hat{\theta}_{BC})/T]}{E[\hat{\Delta}(\hat{\theta}_{BC})/T]^2} \tag{2.30}$$

Similar to the common parameter case, the resulting MSE is

$$\begin{aligned} & MSE(\tilde{\mu}_{SH,1}(\lambda_{\mu_1,nT})) \\ = \; & E[\tilde{\mu}(\hat{\theta}_{BC}) - \theta_0]^2 \\ - \; & \frac{\{E[\hat{\Delta}(\hat{\theta}_{BC})/T]^2 - E[(\hat{\mu}(\hat{\theta}_{BC}) - \mu_0)\hat{\Delta}(\hat{\theta}_{BC})/T]\}^2}{E[\hat{\Delta}(\hat{\theta}_{BC})/T]^2}, \end{aligned} \tag{2.31}$$

where the second term on the right-hand side of (2.31) is $O(T^{-3})$, similar to the shrinkage estimator for θ. To derive a plug-in estimator of $\lambda_{\mu_1,nT}$, we would need to obtain the joint distribution of $(\mu(\hat{\theta}_{BC}), \Delta(\hat{\theta}_{BC}))^{'}$. [6]

Alternatively, one may want to combine the MLE and BCE of the APE as follows:

$$\tilde{\mu}_{SH,2}(\lambda_{\mu_2}) \;=\; \lambda_{\mu_2}\tilde{\mu}(\hat{\theta}_{BC}) + (1-\lambda_{\mu_2})\hat{\mu}(\hat{\theta}_{ML}). \tag{2.32}$$

Now to minimize the MSE of $\tilde{\mu}_{SH,2}$

$$\begin{aligned} & MSE(\tilde{\mu}_{SH,2}(\lambda_{\mu_2})) \\ = \; & E[\lambda_{\mu_2}\tilde{\mu}(\hat{\theta}_{BC}) + (1-\lambda_{\mu_2})\hat{\mu}(\hat{\theta}_{ML}) - \mu_0]^2 \\ = \; & \lambda_{\mu_2}^2 E[\tilde{\mu}(\hat{\theta}_{BC}) - \mu_0]^2 \\ + \; & 2\lambda_{\mu_2}(1-\lambda_{\mu_2})E[(\tilde{\mu}(\hat{\theta}_{BC}) - \mu_0)\hat{\mu}(\hat{\theta}_{ML})] + (1-\lambda_{\mu_2})^2 E[\hat{\mu}(\hat{\theta}_{ML})]^2 \\ \lambda_{\mu_2,nT} \; = \; & \arg\min MSE(\tilde{\mu}_{SH,2}(\lambda_{\mu_2})) \\ = \; & \frac{E[\hat{\mu}(\hat{\theta}_{ML})]^2 - E[(\tilde{\mu}(\hat{\theta}_{BC}) - \mu_0)\hat{\mu}(\hat{\theta}_{ML})]}{E[\tilde{\mu}(\hat{\theta}_{BC}) - \mu_0]^2 - 2E[(\tilde{\mu}(\hat{\theta}_{BC}) - \mu_0)\hat{\mu}(\hat{\theta}_{ML})] + E[\hat{\mu}(\hat{\theta}_{ML})]^2} \end{aligned}$$

[6] Note that μ and Δ are the expected quantities of $\hat{\mu}$ and $\hat{\Delta}$, respectively. We thus have to show that
$\sqrt{nT}(\hat{\mu}(\hat{\theta}_{BC}) - \mu(\hat{\theta}_{BC}), \hat{\Delta}(\hat{\theta}_{BC}) - \Delta(\hat{\theta}_{BC}))^{'} = o_p(1)$.

After some manipulation, $MSE(\tilde{\mu}_{SH,2}(\lambda_{\mu_2,nT}))$ simplifies to the following. For details of the derivation, see Section 2.5.

$$\begin{aligned} & MSE(\tilde{\mu}_{SH,2}(\lambda_{\mu_2,nT})) \\ = \; & E[\tilde{\mu}(\hat{\theta}_{BC}) - \mu_0]^2 \\ - \; & \frac{\{E[\tilde{\mu}(\hat{\theta}_{BC}) - \mu_0]^2 - E[(\tilde{\mu}(\hat{\theta}_{BC}) - \mu_0)\hat{\mu}(\hat{\theta}_{ML})]\}^2}{E[\tilde{\mu}(\hat{\theta}_{BC}) - \mu_0]^2 - 2E[(\tilde{\mu}(\hat{\theta}_{BC}) - \mu_0)\hat{\mu}(\hat{\theta}_{ML})] + E[\hat{\mu}(\hat{\theta}_{ML})]^2} \end{aligned} \tag{2.33}$$

The second term on the right-hand side of (2.33) is positive, hence there is an improvement over $\tilde{\mu}(\hat{\theta}_{BC})$ in terms of MSE. Now noting that $\tilde{\mu}(\hat{\theta}_{BC}) - \mu_0 = O_p(T^{-3/2})$,[7] and $\hat{\mu}(\hat{\theta}_{ML}) = O_p(1)$. It follows that the denominator of the second term of the right-hand side of (2.33) is $O(1)$, since $E[\hat{\mu}(\hat{\theta}_{ML})]^2$ dominates. The numerator is $O(T^{-3})$. Hence, the entire term is $O(T^{-3})$.

A plug-in estimator of $\lambda_{\mu_2,nT}$ is needed to obtain a feasible version of the shrinkage estimator, $\tilde{\mu}_{SH,2}$. What we need here is the asymptotic distribution of the following vector:

$$\sqrt{nT}\begin{pmatrix} \hat{\mu}(\hat{\theta}_{ML}) - \mu_0 \\ \tilde{\mu}(\hat{\theta}_{BC}) - \mu_0 \end{pmatrix}. \tag{2.34}$$

In that case, there are two options. The first is to show the joint normality of the MLE and BCE of the common parameter,

$$\sqrt{nT}\begin{pmatrix} \hat{\theta}_{ML} - \theta_0 \\ \hat{\theta}_{BC} - \theta_0 \end{pmatrix} = \sqrt{nT}\begin{pmatrix} \hat{\theta}_{ML} - \theta_0 \\ \hat{\theta}_{ML} - \frac{\bar{B}(\hat{\theta}_{ML})}{T} \end{pmatrix}. \tag{2.35}$$

[7] This follows from the fact that $\tilde{\mu}(\hat{\theta}_{BC})$ has zero first-order bias

which follows from the joint normality of $(\hat{\theta}_{ML}, \bar{B}(\hat{\theta}_{ML}))^{T}$.[8] One can then apply the δ-method to the expected quantities of the vector in (2.34) to show its asymptotic normality.

As for the second option, one can expand $\tilde{\mu}_{BC}$ around $\hat{\theta}_{ML}$ and show the asymptotic normality of (2.34) based on the asymptotic normality of $\hat{\theta}_{ML}$. The former step proceeds as follows:

Let

$$\begin{aligned}
\mu(\theta) &= \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} E[m(z_{it}, \theta, \hat{\alpha}_i(\theta)], \\
\Delta(\theta) &= \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} E[m_{\alpha}(z_{it}, \theta, \hat{\alpha}_i(\theta))[\beta_i(\theta) + \psi_{it}(\theta)] + m_{\alpha\alpha}(z_{it}, \theta, \hat{\alpha}_i)\sigma_i^2(\theta)/2]
\end{aligned}$$

and let μ_{θ} and Δ_{θ} denote the first-order derivative of μ and Δ with respect to θ, respectively.[9]

The shrinkage estimator for the APE can be written as follows:

$$\tilde{\mu}_{SH,2} = \lambda_{\mu_2,nT}\tilde{\mu}(\hat{\theta}_{BC}) + (1 - \lambda_{\mu_2,nT})\hat{\mu}(\hat{\theta}_{ML}) \tag{2.36}$$

where $\tilde{\mu}(\hat{\theta}_{BC}) = \mu(\hat{\theta}_{BC}) - \Delta(\hat{\theta}_{BC})$. Now to get the joint normality of $\tilde{\mu}_0(\hat{\theta}_{BC})$ and $\mu(\hat{\theta}_{ML})$, we need a first-order Taylor series expansion of $\mu(\hat{\theta}_{BC})$ and $\Delta(\hat{\theta}_{BC})$ around $\hat{\theta}_{ML}$.

$$\begin{aligned}
\mu(\hat{\theta}_{BC}) &= \mu(\hat{\theta}_{ML}) + \mu_{\theta}(\hat{\theta}_{ML})\frac{B(\hat{\theta}_{ML})}{T} + O(T^{-2}), \\
\Delta(\hat{\theta}_{BC}) &= \Delta(\hat{\theta}_{ML}) + \Delta_{\theta}(\hat{\theta}_{ML})\frac{B(\hat{\theta}_{ML})}{T} + O(T^{-2}).
\end{aligned}$$

[8] Note that this is already shown to derive the shrinkage estimator of the common parameter.
[9] Note that we use the chain rule and implicit function theorem to derive $\partial\hat{\alpha}_i(\theta)/\partial\theta$.

Now let $g(\theta) \equiv (\mu(\theta), \mu_\theta(\theta)B(\theta), \Delta(\theta), \Delta_\theta(\theta)B(\theta))^T$. We can now apply the δ-method to g.

$$\sqrt{nT}(g(\hat{\theta}_{ML}) - g(\tilde{\theta}_0)) \xrightarrow{d} N(0, g'(\tilde{\theta}_0)[\bar{E}[\mathcal{I}_i]]^{-1} g'(\tilde{\theta}_0)^T)$$

where $\tilde{\theta}_0 = \theta_0 + B\sqrt{\rho}$.

Now the vector $(\mu(\hat{\theta}_{ML}), \tilde{\mu}(\hat{\theta}_{BC}))^T \equiv D^T g(\hat{\theta}_{ML})$, where

$$D = \begin{pmatrix} 1 & 1 \\ 0 & 1/T \\ 0 & -1/T \\ 0 & -1/T^2 \end{pmatrix}$$

The infeasible versions of the shrinkage estimators, $\tilde{\mu}_{SH,1}(\lambda_{\mu_1,nT})$ and $\tilde{\mu}_{SH,2}(\lambda_{\mu_2,nT})$, lead to different MSEs, as in (2.31) and (2.33), respectively. Both lead to $O(T^{-3})$ MSE improvements. The plug-in estimator of $\tilde{\mu}_{SH,1}(\lambda_{\mu_1,nT})$ is less computationally intensive. Hence, it will be chosen for the Monte Carlo study.

2.3.3 Monte Carlo Study

Numerical Results for the Common Parameter θ

The estimator is tested using Monte Carlo simulations. The following estimators will be compared in terms of global properties and sample-by-sample

comparison:

$$\begin{aligned}
\hat{\theta}_{ML} &\equiv \arg\max \sum_{i=1}^{n}\sum_{t=1}^{T} \log f(z_{it}, \theta, \hat{\alpha}_i) \\
\hat{\theta}_{BC} &\equiv \hat{\theta}_{ML} - \frac{\bar{B}(\hat{\theta}_{ML})}{T} \\
\hat{\theta}_{SH} &\equiv \hat{\theta}_{ML} - \frac{\hat{\Lambda}_{nT}\bar{B}(\hat{\theta}_{ML})}{T}
\end{aligned} \tag{2.37}$$

The Monte Carlo setup is the same as in HN04, which is taken from Heckman (1981). The following equations describe how $z_{it}^{'} \equiv (y_{it}, x_{it}^{'})$ is generated.

$$\begin{aligned}
y_{it} &= 1_{\{x_{it}\theta_0+\alpha_i+\epsilon_{it}>0\}}, \alpha_i \sim N(0,1), \epsilon_{it} \sim N(0,1), \\
x_{it} &= t/10 + x_{i,t-1}/2 + u_{it}, z_{i0} = u_{i0}, u_{it} \sim U(-1/2, 1/2), \\
\text{where} \quad & \theta_0 = 1, \quad N = 100, T = 4, 8.
\end{aligned} \tag{2.38}$$

Tables A.1 and A.2 report the Monte Carlo results for the shrinkage estimator and compares it to the MLE and the analytical BCE. The results show a significant improvement in terms of reducing the bias, especially when individual heterogeneity is large and $T = 4$.

Numerical Results for the APE

Using the same Monte Carlo setup as for the common parameter, we report results for the Overall APE[10] as well as the APE at the Average. For the Overall APE we compare the MLE, analytical BCE and the shrinkage BCE, given as

[10]To my knowledge, this term was first defined in Kim and Sun (2009).

follows:

$$\begin{aligned}
\hat{\mu}(\hat{\theta}_{ML}) &= \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\hat{\theta}_{ML}\phi(x_{it};\hat{\theta}_{ML},\hat{\alpha}_i(\hat{\theta}_{ML})),\\
\tilde{\mu}(\hat{\theta}_{BC}) &= \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\hat{\theta}_{BC}\phi(x_{it};\hat{\theta}_{BC},\hat{\alpha}_i(\hat{\theta}_{BC})) - \frac{\hat{\Delta}(\hat{\theta}_{BC})}{T},\\
\tilde{\mu}_{SH,1} &= \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\hat{\theta}_{BC}\phi(x_{it};\hat{\theta}_{BC},\hat{\alpha}_i(\hat{\theta}_{BC})) - \hat{\lambda}_{\mu_1}\frac{\hat{\Delta}(\hat{\theta}_{BC})}{T}.
\end{aligned}$$

The respective quantities for the APE at the Average are

$$\begin{aligned}
\hat{\mu}(\hat{\theta}_{ML}) &= \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{ML}\phi(\bar{x};\hat{\theta}_{ML},\hat{\alpha}_i(\hat{\theta}_{ML})),\\
\tilde{\mu}(\hat{\theta}_{BC}) &= \frac{1}{nT}\sum_{i=1}^{n}\hat{\theta}_{BC}\phi(\bar{x};\hat{\theta}_{BC},\hat{\alpha}_i(\hat{\theta}_{BC})) - \frac{\hat{\Delta}(\hat{\theta}_{BC})}{T},\\
\hat{\mu}_{SH,1} &= \frac{1}{nT}\sum_{i=1}^{n}\hat{\theta}_{BC}\phi(\bar{x};\hat{\theta}_{BC},\hat{\alpha}_i(\hat{\theta}_{BC})) - \hat{\lambda}_{\mu_1}\frac{\hat{\Delta}(\hat{\theta}_{BC})}{T}.
\end{aligned}$$

Note that for the APE at the Average $\bar{\psi}_{it}$ is omitted from $\hat{\Delta}$.

The simulation results for the Overall APE and the APE at the Average are reported in Tables 3-4 and Tables 5-6, respectively. In interpreting the results for both types of APEs, it is important to note that as Fernandez-Val (2009) shows, the APE in a static nonlinear panel data model is not biased. Hence, bias correcting is not necessary in the first place and as we can see it generally leads to marginal improvement -if any- in mean or median. A similar conclusion applies to the shrinkage estimator of the Overall APE, when $T = 8$. Otherwise, the shrinkage estimator performs quite poorly, especially for the APE at the Average when $T = 4$. A possible explanation of this performance is that the shrinkage estimator relies heavily on the asymptotic distribution of $\hat{\mu}$ and $\hat{\Delta}$. The simulation results for

the analytical BCE show that applying the δ-method to $\hat{\mu}$ leads to a very poor approximation of the finite-sample distribution, which is evident from the coverage probabilities and the ratio of the standard error to the standard deviation. This leads us to conjecture that the shrinkage estimator for the APE performs poorly due to its heavy reliance on this asymptotic distribution.

2.4 Concluding Remarks

The paper at hand proposes a shrinkage approach to bias correction of fixed effects estimation of nonlinear panel data models. The paper has focused on shrinkage estimators based on analytical BCEs proposed in HN04. However, the approach outlined here can be applied in a straightforward manner to other analytical BCEs, such as those proposed in Hahn and Kuersteiner (2011) and Fernandez-Val (2009). The paper illustrates how shrinkage can lead to higher order improvements in the MSE and how it can reduce the uncertainty from estimating the bias. Simulation results show that the shrinkage estimator works well for the common parameter. However, for the APE in the static model, there is no gain from using shrinkage, since the APE is not biased and the asymptotic distribution on which the estimator is based is a poor approximation of the finite-sample distribution.

2.5 Mathematical Appendix

2.5.1 Notation, Definitions and Derivations

Expectation Operators

$$\bar{E}[m(Z_{iT},\theta,\alpha_i)] \equiv \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} E[m(Z_{iT},\theta,\alpha_i)], \tag{2.39}$$

$$\hat{E}_n[m(Z_{iT},\theta,\alpha_i)] \equiv \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} m(Z_{it},\hat{\theta}_{ML},\hat{\alpha}_i(\hat{\theta}_{ML})). \tag{2.40}$$

2.5.2 Bias Estimator: $\bar{B}(\theta)$

Now let $\mathrm{f}(z_{it};\theta,\alpha)$ be the contribution to the likelihood of individual i at time t.

$$u_{it}(\theta,\alpha) \equiv \frac{\partial}{\partial\theta}\log \mathrm{f}(z_{it};\theta,\alpha), \tag{2.41}$$

$$v_{it}(\theta,\alpha) \equiv \frac{\partial}{\partial\alpha_i}\log \mathrm{f}(z_{it};\theta,\alpha), \tag{2.42}$$

$$v_{it\alpha}(\theta,\alpha) \equiv \frac{\partial v_{it}(\theta,\alpha)}{\partial\alpha}. \tag{2.43}$$

Let g_{it} denote $g_{it}(\theta_0,\alpha_0)$, and $g_{it}(\theta)$ denote $g_{it}(\theta,\hat{\alpha}_i(\theta))$, where $\hat{\alpha}_i(\theta)$ is the concentrated individual effect.

$$U_{it}(\theta) \equiv u_{it}(\theta) - v_{it}(\theta)\frac{\sum_{s=1}^{T} u_{it}(\theta)v_{it}(\theta)}{\sum_{s=1}^{T} v_{is}(\theta)^2}, \tag{2.44}$$

$$V_{2it}(\theta) \equiv v_{it}(\theta)^2 + v_{it\alpha}(\theta), \tag{2.45}$$

Now the bias estimator based on the Bartlett identities is defines as follows:

$$\begin{aligned}
\bar{B}(\theta) &= -\bar{H}(\theta)^{-1}\bar{b}(\theta), && (2.46)\\
\bar{H}(\theta) &= -\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} U_{it}(\theta)U_{it}(\theta)^T, && (2.47)\\
\bar{b}(\theta) &= -\frac{1}{2n}\sum_{i=1}^{n}\sum_{t=1}^{T} U_{it}(\theta)V_{2it}(\theta)/(\sum_{t=1}^{T} v_{is}(\theta)^2). && (2.48)
\end{aligned}$$

Asymptotic Bias and Variance of the MLE

$$\begin{aligned}
B &\equiv -\frac{1}{2}[\bar{E}[\mathcal{I}_i]]^{-1}\bar{E}[E[U_{it}(\theta_0)V_{2it}(\theta_0)]/E[v_{it}(\theta_0)^2]] && (2.49)\\
\bar{E}[\mathcal{I}_i] &\equiv \bar{E}[U_{it}U_{it}^T] && (2.50)
\end{aligned}$$

2.5.3 Derivation of Shrinkage Weighting Matrix and Equation (2.23)

$$\begin{aligned}
\Lambda_{nT} &= \arg\min E\|\hat{\theta}_{SH}(\Lambda) - \theta_0\|_T\\
&= \arg\min E\left\|\hat{\theta}_{ML} - \Lambda\frac{\bar{B}(\hat{\theta}_{ML})}{T} - \theta_0\right\|_T\\
&= \arg\min\left(E\|\hat{\theta}_{ML} - \theta_0\|_T - E\left[(\hat{\theta}_{ML} - \theta_0)\frac{\bar{B}(\hat{\theta}_{ML})}{T}\Lambda^T\right]\right.\\
&\quad \left. -E\left[\Lambda\frac{\bar{B}(\hat{\theta}_{ML})}{T}(\hat{\theta}_{ML} - \theta_0)^T\right] + E\left\|\Lambda\frac{\bar{B}(\hat{\theta}_{ML})}{T}\right\|_T\right)
\end{aligned}$$

The first-order conditions imply

$$-E\left[(\hat{\theta}_{ML} - \theta_0)\frac{\bar{B}(\hat{\theta}_{ML})^T}{T}\right] + \Lambda E\left\|\frac{\bar{B}(\hat{\theta}_{ML})}{T}\right\|_T = 0.$$

$$\Lambda_{nT} = E\left[(\hat{\theta}_{ML}-\theta_0)\frac{\bar{B}(\hat{\theta}_{ML})^T}{T}\right]\left\{E\left\|\frac{\bar{B}(\hat{\theta}_{ML})}{T}\right\|_T\right\}^{-1}. \tag{2.51}$$

$$\begin{aligned}
& E\|\hat{\theta}_{SH}(\Lambda_{nT})-\theta_0\|_T \\
= \; & E\|\hat{\theta}_{ML}-\theta_0\|_T \\
- \; & E\left[(\hat{\theta}_{ML}-\theta_0)\frac{\bar{B}(\hat{\theta}_{ML})^T}{T}\right]\left\{E\left\|\frac{\bar{B}(\hat{\theta}_{ML})}{T}\right\|_T\right\}^{-1} E\left[\frac{\bar{B}(\hat{\theta}_{ML})}{T}(\hat{\theta}_{ML}-\theta_0)^T\right] \\
- \; & E\left[(\hat{\theta}_{ML}-\theta_0)\frac{\bar{B}(\hat{\theta}_{ML})^T}{T}\right]\left\{E\left\|\frac{\bar{B}(\hat{\theta}_{ML})}{T}\right\|_T\right\}^{-1} E\left[\frac{\bar{B}(\hat{\theta}_{ML})}{T}(\hat{\theta}_{ML}-\theta_0)^T\right] \\
+ \; & E\left[(\hat{\theta}_{ML}-\theta_0)\frac{\bar{B}(\hat{\theta}_{ML})^T}{T}\right]\left\{E\left\|\frac{\bar{B}(\hat{\theta}_{ML})}{T}\right\|_T\right\}^{-1} E\left[\frac{\bar{B}(\hat{\theta}_{ML})}{T}(\hat{\theta}_{ML}-\theta_0)^T\right]
\end{aligned}$$

$$\begin{aligned}
= \; & E\|\hat{\theta}_{ML}-\theta_0\|_T \\
- \; & E\left[(\hat{\theta}_{ML}-\theta_0)\frac{\bar{B}(\hat{\theta}_{ML})^T}{T}\right]\left\{E\left\|\frac{\bar{B}(\hat{\theta}_{ML})}{T}\right\|_T\right\}^{-1} E\left[\frac{\bar{B}(\hat{\theta}_{ML})}{T}(\hat{\theta}_{ML}-\theta_0)^T\right]
\end{aligned}$$

Derivation of $MSE(\tilde{\mu}_{SH,2}(\lambda_{\mu_2,nT}))$

$$\begin{aligned}
& MSE(\tilde{\mu}_{SH,2}(\lambda_{\mu_2,nT})) \\
= \; & \frac{\{E[\hat{\mu}(\hat{\theta}_{ML})]^2 - E[(\tilde{\mu}(\hat{\theta}_{BC})-\mu_0)\hat{\mu}(\hat{\theta}_{ML})]\}^2 E[\tilde{\mu}(\hat{\theta}_{BC})-\mu_0]^2}{\{E[\tilde{\mu}(\hat{\theta}_{BC})-\mu_0]^2 - 2E[(\tilde{\mu}(\hat{\theta}_{BC})-\mu_0)\hat{\mu}(\hat{\theta}_{ML})] + E[\hat{\mu}(\hat{\theta}_{ML})]^2\}^2} \\
+ \; & 2\frac{\{E[\hat{\mu}(\hat{\theta}_{ML})]^2 - E[(\tilde{\mu}(\hat{\theta}_{BC})-\mu_0)\hat{\mu}(\hat{\theta}_{ML})]\}}{\{E[\tilde{\mu}(\hat{\theta}_{BC})-\mu_0]^2 - 2E[(\tilde{\mu}(\hat{\theta}_{BC})-\mu_0)\hat{\mu}(\hat{\theta}_{ML})] + E[\hat{\mu}(\hat{\theta}_{ML})]^2\}^2} \\
\times \; & \{E[\tilde{\mu}(\hat{\theta}_{BC})-\mu_0]^2 - E[(\tilde{\mu}(\hat{\theta}_{BC})-\mu_0)\hat{\mu}(\hat{\theta}_{ML})]\}E[(\tilde{\mu}(\hat{\theta}_{BC})-\mu_0)\hat{\mu}(\hat{\theta}_{ML})] \\
+ \; & \left\{\frac{E[\tilde{\mu}(\hat{\theta}_{BC})-\mu_0]^2 - E[(\tilde{\mu}(\hat{\theta}_{BC})-\mu_0)\hat{\mu}(\hat{\theta}_{ML})]}{E[\tilde{\mu}(\hat{\theta}_{BC})-\mu_0]^2 - 2E[(\tilde{\mu}(\hat{\theta}_{BC})-\mu_0)\hat{\mu}(\hat{\theta}_{ML})] + E[\hat{\mu}(\hat{\theta}_{ML})]^2}\right\}^2 E[\hat{\mu}(\hat{\theta}_{ML})]^2
\end{aligned} \tag{2.52}$$

Let $a_1 \equiv E[\tilde{\mu}(\hat{\theta}_{BC}) - \mu_0]^2$, $a_2 \equiv E[(\tilde{\mu}(\hat{\theta}_{BC}) - \mu_0)\hat{\mu}(\hat{\theta}_{ML})]$, and $a_3 \equiv E[\hat{\mu}(\hat{\theta}_{ML})]^2$. Now the numerator of the right hand side of the last equation can be written as:

$$\begin{aligned}
& (a_3 - a_2)^2 a_1 + 2(a_3 - a_2)(a_1 - a_2)a_2 + (a_1 - a_2)^2 a_3 \\
= \; & (a_3^2 - 2a_3a_2 + a_2^2)a_1 + 2(a_3a_1 - a_3a_2 - a_2a_1 + a_2^2)a_2 \\
& + (a_1^2 - 2a_1a_2 + a_2^2)a_3 \\
= \; & a_1a_3^2 - 2a_1a_2a_3 + a_1a_2^2 + 2a_1a_2a_3 - 2a_2^2a_3 - 2a_1a_2^2 \\
& + 2a_2^3 + a_1^2a_3 - 2a_1a_2a_3 + a_2^2a_3 \\
= \; & a_1a_3^2 - 2a_1a_2a_3 - a_1a_2^2 - a_2^2a_3 + 2a_2^3 + a_1^2a_3 \\
= \; & a_1a_3(a_1 - 2a_2 + a_3) - a_2^2(a_1 - 2a_2 + a_3) \qquad (2.53)
\end{aligned}$$

Noting that $a_1 - 2a2 + a_3$ is the denominator, (2.52) simplifies to the following:

$$\begin{aligned}
& MSE(\tilde{\mu}_{SH,2}(\lambda_{\mu_2,nT})) \\
= \; & \frac{E[\tilde{\mu}(\hat{\theta}_{BC}) - \mu_0]^2 E[\hat{\mu}(\hat{\theta}_{ML})]^2 - \{E[(\tilde{\mu}(\hat{\theta}_{BC}) - \mu_0)\hat{\mu}(\hat{\theta}_{ML})]\}^2}{E[\tilde{\mu}(\hat{\theta}_{BC}) - \mu_0]^2 - 2E[(\tilde{\mu}(\hat{\theta}_{BC}) - \mu_0)\hat{\mu}(\hat{\theta}_{ML})] + E[\hat{\mu}(\hat{\theta}_{ML})]^2} \qquad (2.54)
\end{aligned}$$

We can rewrite (2.54) in terms of a_1, a_2, and a_3, then adding and subtracting a_1 leads to the following.

$$\begin{aligned}
& \frac{a_1a_3 - a_2^2}{a_1 - 2a_2 + a_3} \\
= \; & a_1 + \frac{a_1a_3 - a_2^2 - a_1(a_1 - 2a_2 + a_3)}{a_1 - 2a_2 + a_3} \\
= \; & a_1 - \frac{(a_1 - a_2)^2}{a_1 - 2a_2 + a_3}. \qquad (2.55)
\end{aligned}$$

2.5.4 Lemmas and Proofs

Proof (Proof of Proposition 2.3.1) First, we derive the relationship between the MSE of $\hat{\theta}_{SH}(\Lambda_{nT})$ and $\hat{\theta}_{BC}$, assuming the MSE exists to perform the following derivation.[11]

$$
\begin{aligned}
& E\|\hat{\theta}_{SH}(\Lambda_{nT}) - \theta_0\|_T \\
= \; & E\|\hat{\theta}_{BC} - \theta_0\|_T + E\left[(\hat{\theta}_{ML} - \theta_0)\frac{\bar{B}(\hat{\theta}_{ML})^T}{T}\right] \\
& + E\left[\frac{\bar{B}(\hat{\theta}_{ML})}{T}(\hat{\theta}_{ML} - \theta_0)^T\right] - E\left\|\frac{\bar{B}(\hat{\theta}_{ML})}{T}\right\|_T \\
- \; & E\left[(\hat{\theta}_{ML} - \theta_0)\frac{\bar{B}(\hat{\theta}_{ML})^T}{T}\right]\left\{E\left\|\frac{\bar{B}(\hat{\theta}_{ML})}{T}\right\|_T\right\}^{-1} E\left[\frac{\bar{B}(\hat{\theta}_{ML})}{T}(\hat{\theta}_{ML} - \theta_0)^T\right] \\
= \; & E\|\hat{\theta}_{BC} - \theta_0\|_T \\
& + E\left[(\hat{\theta}_{ML} - \theta_0)\frac{\bar{B}(\hat{\theta}_{ML})^T}{T}\right]\left\{E\left\|\frac{\bar{B}(\hat{\theta}_{ML})}{T}\right\|_T\right\}^{-1} E\left\|\frac{\bar{B}(\hat{\theta}_{ML})}{T}\right\|_T \\
+ \; & E\left\|\frac{\bar{B}(\hat{\theta}_{ML})}{T}\right\|_T\left\{E\left\|\frac{\bar{B}(\hat{\theta}_{ML})}{T}\right\|_T\right\}^{-1} E\left[\frac{\bar{B}(\hat{\theta}_{ML})}{T}(\hat{\theta}_{ML} - \theta_0)^T\right] \\
- \; & E\left\|\frac{\bar{B}(\hat{\theta}_{ML})}{T}\right\|_T\left\{E\left\|\frac{\bar{B}(\hat{\theta}_{ML})}{T}\right\|_T\right\}^{-1} E\left\|\frac{\bar{B}(\hat{\theta}_{ML})}{T}\right\|_T \\
- \; & E\left[(\hat{\theta}_{ML} - \theta_0)\frac{\bar{B}(\hat{\theta}_{ML})^T}{T}\right]\left\{E\left\|\frac{\bar{B}(\hat{\theta}_{ML})}{T}\right\|_T\right\}^{-1} E\left[\frac{\bar{B}(\hat{\theta}_{ML})}{T}(\hat{\theta}_{ML} - \theta_0)^T\right] \\
= \; & E\|\hat{\theta}_{BC} - \theta_0\|_T \\
- \; & \left\{E\left\|\frac{\bar{B}(\hat{\theta}_{ML})}{T}\right\|_T - E\left[(\hat{\theta}_{ML} - \theta_0)\frac{\bar{B}(\hat{\theta}_{ML})^T}{T}\right]\right\}\left\{E\left\|\frac{\bar{B}(\hat{\theta}_{ML})}{T}\right\|_T\right\}^{-1} \\
\times \; & \left\{E\left\|\frac{\bar{B}(\hat{\theta}_{ML})}{T}\right\|_T - E\left[(\hat{\theta}_{ML} - \theta_0)\frac{\bar{B}(\hat{\theta}_{ML})^T}{T}\right]\right\}^T
\end{aligned}
\tag{2.56}
$$

[11] This assumption will be dropped later.

Now in order to find the order of magnitude of the last term in the above equation, we will focus on the T dimension and will introduce some notation. Let $\theta_{ML,T} \equiv \lim_{n\to\infty} \hat{\theta}_{ML}$, $\theta_{BC,T} \equiv \lim_{n\to\infty} \hat{\theta}_{BC}$, $\theta_{SH,T} \equiv \lim_{n\to\infty} \hat{\theta}_{SH}$, $\overline{\bar{B}(\hat{\theta}_{ML})} = \lim_{n\to\infty} \bar{B}(\hat{\theta}_{ML})$ and $\Lambda_T \equiv \lim_{n\to\infty} \Lambda_{nT}$.

$$\begin{aligned} & E\left\|\frac{\overline{\bar{B}(\hat{\theta}_{ML})}}{T}\right\|_T - E\left[(\theta_{ML,T}-\theta_0)\frac{\overline{\bar{B}(\hat{\theta}_{ML})}^T}{T}\right] \\ = \; & -E\left[\left(\theta_{ML,T}-\theta_0-\frac{\overline{\bar{B}(\hat{\theta}_{ML})}}{T}\right)\frac{\overline{\bar{B}(\hat{\theta}_{ML})}^T}{T}\right] \\ = \; & -E\left[\left(\frac{B}{T}+O_p(T^{-2})-\frac{\overline{\bar{B}(\hat{\theta}_{ML})}}{T}\right)\frac{\overline{\bar{B}(\hat{\theta}_{ML})^T}}{T}\right] \\ = \; & O(T^{-5/2}) \end{aligned} \tag{2.57}$$

Given that the assumptions of Lemma A.2 from Rilstone, Srivastava, and Ullah (1996) are fulfilled here, it follows that the MSE of $\theta_{ML,T}$, $\theta_{BC,T}$ and $\theta_{SH,T}(\Lambda_T)$ exist up to order ($O(T^{-3})$

$$MSE(\theta_{SH,T}(\Lambda_T)) = MSE(\theta_{BC,T}) - \frac{\Delta_1}{T^3},$$

where Δ_1 is positive definite.

Now we manipulate the MSE of $\theta_{BC,T}$ similarly.

$$\begin{aligned} E\|\theta_{BC,T}-\theta_0\|_T &= E\|\theta_{ML,T}-\theta_0\|_T - \left\|\frac{B}{T}\right\|_T + E\left\|\frac{\overline{\bar{B}(\hat{\theta}_{ML})}-B}{T}\right\|_T \\ &= E\|\theta_{ML,T}-\theta_0\|_T - \left\|\frac{B}{T}\right\|_T + \frac{\Delta_2}{T^3} \end{aligned} \tag{2.58}$$

where Δ_2 is positive definite.

$$
\begin{aligned}
& \frac{\Delta_2}{T^3} - \frac{\Delta_1}{T^3} \\
= \; & E\left\| \overline{\frac{\bar{B}(\hat{\theta}_{ML})}{T}} - \frac{B}{T} \right\|_T - E\left[\left(\overline{\frac{\bar{B}(\hat{\theta}_{ML})}{T}} - \frac{B}{T} \right) \overline{\frac{\bar{B}(\hat{\theta}_{ML})^T}{T}} \right] \left\{ E\left\| \overline{\frac{\bar{B}(\hat{\theta}_{ML})}{T}} \right\|_T \right\}^{-1} \\
\times \; & \left\{ E\left[\left(\overline{\frac{\bar{B}(\hat{\theta}_{ML})}{T}} - \frac{B}{T} \right) \overline{\frac{\bar{B}(\hat{\theta}_{ML})^T}{T}} \right] \right\}^T - O(T^{-4})
\end{aligned}
$$

The second term on the right-hand side can be simplified as follows:

$$
\begin{aligned}
& E\left[\left(\overline{\frac{\bar{B}(\hat{\theta}_{ML})}{T}} - \frac{B}{T} \right) \overline{\frac{\bar{B}(\hat{\theta}_{ML})^T}{T}} \right] \\
& \times \left\{ E\left\| \overline{\frac{\bar{B}(\hat{\theta}_{ML})}{T}} \right\|_T \right\}^{-1} \left\{ E\left[\left(\overline{\frac{\bar{B}(\hat{\theta}_{ML})}{T}} - \frac{B}{T} \right) \overline{\frac{\bar{B}(\hat{\theta}_{ML})^T}{T}} \right] \right\}^T \\
= \; & E\left[\left(\overline{\frac{\bar{B}(\hat{\theta}_{ML})}{T}} - \frac{B}{T} \right) \overline{\frac{\bar{B}(\hat{\theta}_{ML})^T}{T}} - \frac{B^T}{T} + \frac{B^T}{T} \right] \left\{ E\left\| \overline{\frac{\bar{B}(\hat{\theta}_{ML})}{T}} \right\|_T \right\}^{-1} \\
\times \; & \left\{ E\left[\left(\overline{\frac{\bar{B}(\hat{\theta}_{ML})}{T}} - \frac{B}{T} + \frac{B}{T} - \frac{B}{T} \right) \overline{\frac{\bar{B}(\hat{\theta}_{ML})^T}{T}} \right] \right\}^T \\
= \; & \left\{ E\left\| \overline{\frac{\bar{B}(\hat{\theta}_{ML})}{T}} - \frac{B}{T} \right\|_T + E\left[\overline{\frac{\bar{B}(\hat{\theta}_{ML})}{T}} - \frac{B}{T} \right] \frac{B^T}{T} \right\} \left\{ E\left\| \overline{\frac{\bar{B}(\hat{\theta}_{ML})}{T}} \right\|_T \right\}^{-1} \\
\times \; & \left\{ E\left\| \overline{\frac{\bar{B}(\hat{\theta}_{ML})}{T}} \right\|_T - \frac{B}{T} E\left[\overline{\frac{\bar{B}(\hat{\theta}_{ML})^T}{T}} \right] \right\}^T \\
= \; & E\left\| \overline{\frac{\bar{B}(\hat{\theta}_{ML})}{T}} - \frac{B}{T} \right\|_T \\
+ \; & E\left[\overline{\frac{\bar{B}(\hat{\theta}_{ML})}{T}} - \frac{B}{T} \right] \frac{B^T}{T} \left\{ E\left\| \overline{\frac{\bar{B}(\hat{\theta}_{ML})}{T}} \right\|_T \right\}^{-1} E\left\| \overline{\frac{\bar{B}(\hat{\theta}_{ML})}{T}} \right\|_T
\end{aligned}
$$

$$
\begin{aligned}
&- E\left[\frac{\overline{\bar{B}(\hat{\theta}_{ML})} - B}{T}\right] \frac{B^T}{T} \left\{ E \left\| \frac{\overline{\bar{B}(\hat{\theta}_{ML})}}{T} \right\|_T \right\}^{-1} E\left[\frac{\overline{\bar{B}(\hat{\theta}_{ML})}}{T}\right] \frac{B^T}{T} \\
&- E\left\| \frac{\overline{\bar{B}(\hat{\theta}_{ML})} - B}{T} \right\|_T \left\{ E \left\| \frac{\overline{\bar{B}(\hat{\theta}_{ML})}}{T} \right\|_T \right\}^{-1} E\left[\frac{\overline{\bar{B}(\hat{\theta}_{ML})}}{T}\right] \frac{B^T}{T} \\
&= E\left\| \frac{\overline{\bar{B}(\hat{\theta}_{ML})} - B}{T} \right\|_T \\
&+ E\left[\frac{\overline{\bar{B}(\hat{\theta}_{ML})} - B}{T}\right] \frac{B^T}{T} \left\{ E \left\| \frac{\overline{\bar{B}(\hat{\theta}_{ML})}}{T} \right\|_T \right\}^{-1} E\left[\frac{\overline{\bar{B}(\hat{\theta}_{ML})}}{T} \frac{(\overline{\bar{B}(\hat{\theta}_{ML})} - B)^T}{T}\right] \\
&- E\left\| \frac{\overline{\bar{B}(\hat{\theta}_{ML})} - B}{T} \right\|_T \left\{ E \left\| \frac{\overline{\bar{B}(\hat{\theta}_{ML})}}{T} \right\|_T \right\}^{-1} \frac{BB^T}{T^2} + o(T^{-3})
\end{aligned}
$$

where the last equality follows from substituting $B + O_p(T^{-1/2})$ for $\overline{\bar{B}(\hat{\theta}_{ML})}$ in the last term term on the right-hand side of the penultimate equality.

$$
\begin{aligned}
&\frac{\Delta_2}{T} - \frac{\Delta_1}{T} \\
&= -E\left[\frac{\overline{\bar{B}(\hat{\theta}_{ML})} - B}{T}\right] \frac{B^T}{T} \left\{ E \left\| \frac{\overline{\bar{B}(\hat{\theta}_{ML})}}{T} \right\|_T \right\}^{-1} E\left[\frac{\overline{\bar{B}(\hat{\theta}_{ML})}}{T} \frac{(\overline{\bar{B}(\hat{\theta}_{ML})} - B)^T}{T}\right] \\
&+ E\left\| \frac{\overline{\bar{B}(\hat{\theta}_{ML})} - B}{T} \right\|_T \left\{ E \left\| \frac{\overline{\bar{B}(\hat{\theta}_{ML})}}{T} \right\|_T \right\}^{-1} \frac{BB^T}{T^2} + o(T^{-3}) \\
&= -E\left[\frac{\overline{\bar{B}(\hat{\theta}_{ML})} - B}{T}\right] \frac{B^T}{T} \left\{ E \left\| \frac{\overline{\bar{B}(\hat{\theta}_{ML})}}{T} \right\|_T \right\}^{-1} \\
&\quad \times E\left[\frac{(\overline{\bar{B}(\hat{\theta}_{ML})} - B)}{T} \frac{(\overline{\bar{B}(\hat{\theta}_{ML})} - B)^T}{T}\right] \\
&- E\left[\frac{\overline{\bar{B}(\hat{\theta}_{ML})} - B}{T}\right] \frac{B^T}{T} \left\{ E \left\| \frac{\overline{\bar{B}(\hat{\theta}_{ML})}}{T} \right\|_T \right\}^{-1} \frac{B}{T} E\left[\frac{(\overline{\bar{B}(\hat{\theta}_{ML})} - B)^T}{T}\right] \\
&+ E\left\| \frac{\overline{\bar{B}(\hat{\theta}_{ML})} - B}{T} \right\|_T \left\{ E \left\| \frac{\overline{\bar{B}(\hat{\theta}_{ML})}}{T} \right\|_T \right\}^{-1} \frac{BB^T}{T^2} + o(T^{-3})
\end{aligned}
$$

$$
\begin{aligned}
&= -E\left[\frac{\overline{\bar{B}(\hat{\theta}_{ML})} - B}{T}\right]\frac{B^T}{T}\left\{E\left\|\frac{\overline{\bar{B}(\hat{\theta}_{ML})}}{T}\right\|_T\right\}^{-1}\frac{B}{T}E\left[\frac{(\overline{\bar{B}(\hat{\theta}_{ML})} - B)^T}{T}\right] \\
&+ E\left\|\frac{\overline{\bar{B}(\hat{\theta}_{ML})} - B}{T}\right\|_T\left\{E\left\|\frac{\overline{\bar{B}(\hat{\theta}_{ML})}}{T}\right\|_T\right\}^{-1}\frac{BB^T}{T^2} + o(T^{-3}) \\
&= Var\left(\frac{\overline{\bar{B}(\hat{\theta}_{ML})} - B}{T}\right)\left\{E\left\|\frac{\overline{\bar{B}(\hat{\theta}_{ML})}}{T}\right\|_T\right\}^{-1}\frac{BB^T}{T^2} + o(T^{-3})
\end{aligned}
$$

where the last equality follows from the first-order unbiasedness of $\overline{\bar{B}(\hat{\theta}_{ML})}$ for B. Hence, $|\Delta_2 - \Delta_1| > 0$. □

2.5.5 Lemmas

Lemma 2.5.1 *Let* $f(\theta) \equiv (\theta^T, B(\theta)^T)^T$ *and* $\tilde{\theta}_0 \equiv \theta_0 + B/T$. *Under Conditions 1-4,*

$$\sqrt{nT}(f(\hat{\theta}_{ML}) - f(\tilde{\theta}_0)) \xrightarrow{d} N\left(0, \frac{\partial f(\theta_0)}{\partial \theta}[\bar{E}[\mathcal{I}_i]]^{-1}\frac{\partial f(\theta_0)}{\partial \theta^T}\right)$$

Proof Hahn and Newey (2004), hereinafter HN04, show the asymptotic normality of $\sqrt{nT}(\hat{\theta}_{ML} - \theta_0)$ under conditions 1-4.[12] The result follows from the δ-method.□

Lemma 2.5.2 *Under Conditions 1, 2 and 4,*

$$\sup_{\theta\in\Theta}|\bar{B}(\theta) - B(\theta)| = O_p((nT)^{-1/2}) \tag{2.59}$$

[12]Note that $B(\tilde{\theta}_0)$ converges to B, because $\sqrt{\frac{n}{T}}\bar{B}(\hat{\theta}_{ML})$ is consistent for $\sqrt{\rho}B$ as $n/T \to \rho$ by Theorem 2 in HN04, and by the Continuous Mapping Theorem, $\sqrt{\frac{n}{T}}\bar{B}(\hat{\theta}_{ML}) \xrightarrow{p} B(\tilde{\theta}_0)\sqrt{\rho}$. Thus, $\bar{B}(\tilde{\theta}_0) \xrightarrow{p} B$.

Proof

$$\begin{aligned} |\bar{B}(\theta) - B(\theta)| &\leq |(\bar{H}(\theta)^{-1} - H(\theta)^{-1})\bar{b}(\theta)| + |H(\theta)^{-1}(\bar{b}(\theta) - b(\theta))| \\ &= |\bar{H}(\theta)^{-1} - H(\theta)^{-1}|O_p(1) + O(1)|\bar{b}(\theta) - b(\theta)| \end{aligned} \tag{2.60}$$

To find the rate of convergence for (2.60), the conditions for the Liapunov Central Limit Theorem will be checked for the terms in (2.60).

First, we address the first term, $\sqrt{nT}(\bar{H}(\theta) - H(\theta))$. Let $\boldsymbol{1}_k$ denote a k-dimensional vector of ones.

$$\begin{aligned} |U_{it}(\theta)| &= |u_{it}(\theta) - v_{it}(\theta)\frac{\sum_{s=1}^{T} u_{it}(\theta)v_{it}(\theta)}{\sum_{s=1}^{T} v_{is}(\theta)^2}| \leq |u_{it}(\theta)| + |v_{it}(\theta)\|\frac{\sum_{s=1}^{T} u_{it}(\theta)v_{it}(\theta)}{\sum_{s=1}^{T} v_{is}(\theta)^2}| \\ &\leq \left\{ M(z_{it}) + M(z_{it})\frac{\sum_{s=1}^{T} M(z_{is})^2}{\sum_{s=1}^{T} m(z_{is})^2} \right\} \boldsymbol{1}_k \equiv \mathcal{A}_{it} \end{aligned} \tag{2.61}$$

Now

$$\begin{aligned} & E|U_{it}(\theta)U_{it}(\theta)^T|^{2+\delta} \\ \leq\ & \{(E|M(z_{it})^2 \boldsymbol{1}_k \boldsymbol{1}_k^T|^{2+\delta})^{1/(2+\delta)} \\ +\ & 2(E|M(z_{it})^2 \frac{\sum_{s=1}^{T} M(z_{is})^2}{\sum_{s=1}^{T} m(z_{is})^2} \boldsymbol{1}_k \boldsymbol{1}_k^T|^{2+\delta})^{1/(2+\delta)} \\ +\ & (E|M(z_{it}) \frac{\sum_{s=1}^{T} M(z_{is})^2}{\sum_{s=1}^{T} m(z_{is})^2} M(z_{it}) \frac{\sum_{s=1}^{T} M(z_{is})^2}{\sum_{s=1}^{T} m(z_{is})^2} \boldsymbol{1}_k \boldsymbol{1}_k^T|^{2+\delta})^{1/(2+\delta)}\}^{2+\delta} \\ \leq\ & 4^{1+\delta}|\boldsymbol{1}_k \boldsymbol{1}_k^T|^{2+\delta}\{E|M(z_{it})^2|^{2+\delta} + 2E|M(z_{it})^2 \frac{\sum_{s=1}^{T} M(z_{is})^2}{\sum_{s=1}^{T} m(z_{is})^2}|^{2+\delta} \\ +\ & E|M(z_{it})^2 \frac{\{\sum_{s=1}^{T} M(z_{is})^2\}^2}{\{\sum_{s=1}^{T} m(z_{is})^2\}^2}|^{2+\delta}\} \end{aligned}$$

$$
\begin{aligned}
&\leq 4^{1+\delta}|\boldsymbol{1}_k\boldsymbol{1}_k^T|^{2+\delta}\{\sup_i E|M(z_{it})^2|^{2+\delta} \\
&+ 2\sup_i E|M(z_{it})^2\left\{\frac{E[M(z_{is})^2]}{E[m(z_{is})^2]}+o_p(1)\right\}|^{2+\delta} \\
&+ \sup_i E|M(z_{it})^2\left\{\frac{\{E[M(z_{is})^2]\}^2}{\{E[m(z_{is})^2]\}^2}+o_p(1)\right\}|^{2+\delta}\} < \infty, \qquad (2.62)
\end{aligned}
$$

where the last inequality follows from Condition 4. It hence follows that

$$
\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}U_{it}(\theta)U_{it}(\theta)^T = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}E[U_{it}(\theta)U_{it}(\theta)^T]+O_p((nT)^{-1/2})
$$

Similarly,

$$
\begin{aligned}
&|\bar{b}(\theta)-b(\theta)| \\
=\ &|\frac{1}{2n}\sum_{i=1}^{n}\frac{\sum_{t=1}^{T}U_{it}(\theta)V_{2it}(\theta)}{\sum_{s=1}^{T}v_{is}(\theta)^2}-\frac{1}{2}\bar{E}\left[\frac{E[U_{it}(\theta)V_{2it}(\theta)]}{E[v_{it}(\theta)^2]}\right]|
\end{aligned}
$$

$$
\begin{aligned}
&\frac{1}{2n}\sum_{i=1}^{n}\frac{\sum_{t=1}^{T}U_{it}(\theta)V_{2it}(\theta)}{\sum_{s=1}^{T}v_{is}(\theta)^2}-\frac{1}{2n}\sum_{i=1}^{n}\frac{E[U_{it}(\theta)V_{2it}(\theta)]}{E[v_{it}(\theta)^2]}| \\
\leq\ &\frac{1}{2n}\sum_{i=1}^{n}|\frac{1}{T}\sum_{t=1}^{T}U_{it}(\theta)V_{2it}(\theta)-E[U_{it}(\theta)V_{2it}(\theta)]||\frac{1}{T}\sum_{s=1}^{T}v_{it}(\theta)^2| \\
+\ &\frac{1}{2n}\sum_{i=1}^{n}|E[U_{it}(\theta)V_{2it}(\theta)]||[\frac{1}{T}\sum_{s=1}^{T}v_{it}(\theta)^2]^{-1}-E[v_{it}(\theta)^2]^{-1}| \\
=\ &\frac{1}{2n}\sum_{i=1}^{n}|\frac{1}{T}\sum_{t=1}^{T}U_{it}(\theta)V_{2it}(\theta)-E[U_{it}(\theta)V_{2it}(\theta)]|O_p(1) \\
+\ &\frac{1}{2n}\sum_{i=1}^{n}O(1)|[\frac{1}{T}\sum_{s=1}^{T}v_{it}(\theta)^2]^{-1}-E[v_{it}(\theta)^2]^{-1}| \qquad (2.63)
\end{aligned}
$$

In the following, we show that both quantities, $|\frac{1}{T}\sum_{t=1}^{T}U_{it}(\theta)V_{2it}(\theta)-E[U_{it}(\theta)V_{2it}(\theta)]|$ and $|[\frac{1}{T}\sum_{s=1}^{T}v_{it}(\theta)^2]-E[v_{it}(\theta)^2]|$ satisfy the moment condition for the

Liapunov Central Limit Theorem.

$$|V_{2it}(\theta)| \leq |v_{it}(\theta)^2| + |v_{it\alpha}(\theta)| \leq |M(z_{it})^2| + |M(z_{it})| \equiv \mathcal{B}_{it}$$

Hence,

$$\begin{aligned} E|U_{it}(\theta)V_{2it}(\theta)|^{2+\delta} &\leq 4^{1+\delta}\{E|M(z_{it})^3 \boldsymbol{1}_k|^{2+\delta} + E|M(z_{it})^3 \frac{\sum_{s=1}^T M(z_{is})^2}{\sum_{s=1}^T m(z_{is})^2} \boldsymbol{1}_k|^{2+\delta} \\ &+ E|M(z_{it})^2 \boldsymbol{1}_k|^{2+\delta} + E|M(z_{it})^2 \frac{\sum_{s=1}^T M(z_{is})^2}{\{\sum_{s=1}^T m(z_{is})^2\}^2} \boldsymbol{1}_k|^{2+\delta}\} \end{aligned}$$

$$\begin{aligned} &\leq \sup_i 4^{1+\delta}\{E|M(z_{it})^3 \boldsymbol{1}_k|^{2+\delta} + \sup_i E|M(z_{it})^3 \left\{ \frac{E[M(z_{is})^2]}{E[m(z_{is})^2]} + o_p(1) \right\} \boldsymbol{1}_k|^{2+\delta} \\ &+ \sup_i E|M(z_{it})^2 \boldsymbol{1}_k|^{2+\delta} + E|M(z_{it})^2 \left\{ \frac{E[M(z_{is})^2]}{E[m(z_{is})^2]} + o_p(1) \right\} \boldsymbol{1}_k|^{2+\delta}\} < \infty \end{aligned}$$

where the last inequality follows from Condition 4.

$$E|v_{it}(\theta)^2|^{2+\delta} \leq E|M(z_{it})^2|^{2+\delta} \leq \sup_i E|M(z_{it})|^{2(2+\delta)} < \infty \tag{2.64}$$

By Condition 5 and applying the delta method, we have that $|[\frac{1}{T}\sum_{s=1}^T v_{it}(\theta)^2]^{-1} - \{E[v_{it}(\theta)^2]\}^{-1}| = O_p((nT)^{-1/2})$.

Thus, we have that

$$\frac{1}{2n}\sum_{i=1}^n \frac{\sum_{t=1}^T U_{it}(\theta)V_{2it}(\theta)}{\sum_{s=1}^T v_{is}(\theta)^2} + O_p((nT)^{-1/2}) = lim_{n\to\infty}\frac{1}{2}\sum_{i=1}^n \frac{E[U_{it}(\theta)V_{2it}(\theta)]}{E[v_{it}(\theta)^2]}.$$

It follows that $\sqrt{nT}|\bar{B}(\theta) - B(\theta)| \leq O_p(1)$.

Now it remains to show that $|\bar{B}(\theta) - B(\theta)|$ is stochastically equicontinuous to

ensure that a uniform law of large numbers applies. Let

$$\mathcal{K}_n(\theta) \equiv \bar{B}(\theta) - B(\theta) \tag{2.65}$$

and $\Theta \equiv \{\theta : (\theta, \alpha_i) \in \mathcal{Y} \text{ for } i = 1, ..., n\}$. We need to show that $\sup_{\theta \in \Theta} \sup_{\theta' \in B(\theta,\delta)} |\mathcal{K}_n(\theta) - \mathcal{K}_n(\theta')| \xrightarrow{p} 0$. By Lemma A.1 in Newey (1991), it suffices to show the stochastic equicontinuity of $\bar{B}(\theta) = \bar{H}^{-1}(\theta)\bar{b}(\theta)$.

$$\begin{aligned} & |\bar{H}^{-1}(\theta)\bar{b}(\theta) - \bar{H}^{-1}(\theta')\bar{b}(\theta')| \\ \leq & \left\{ |\bar{H}^{-1}(\theta)| \left| \frac{\partial \bar{b}(\theta)}{\partial \theta} \right| + |\bar{b}(\theta)^{'} \otimes I_k| \left| \frac{\partial \bar{H}^{-1}(\theta)}{\partial \theta} \right| \right\} \|\theta - \theta'\| \end{aligned}$$

Now it remains to show that $|\partial \bar{b}(\theta)/\partial\theta|$, $|\partial \bar{H}^{-1}(\theta)/\partial\theta|$, $|\bar{H}^{-1}(\theta)|$, and $|\bar{b}(\theta)|$ are bounded above by functions that only depend on z_{it}.

We follow the differentiation conventions in Abadir and Magnus (2005). The derivative of a $kx1$ vector, V, with respect to the $kx1$ vector, x, denoted by $\partial V(x)/\partial x$, yields a $k \times k$ matrix with elements $\partial V_i/\partial x_j$ in the i^{th} row and j^{th} column, $i = 1, 2, ..., k$, $j = 1, 2, ..., k$. The derivative of a $k \times k$ matrix $M(x)$ with respect to a $k \times k$ matrix, X, yields a $k^2 \times k^2$ matrix, $\partial vec(M(X))/\partial vec(X)^{'}$.

$$\begin{aligned} & \left| \frac{\partial \bar{b}(\theta)}{\partial \theta} \right| \\ = & \left| \frac{1}{2n} \sum_{i=1}^{n} \frac{(\sum_{s=1}^{T} v_{is}^2(\theta) \otimes I_k) \sum_{t=1}^{T} \frac{\partial U_{it}(\theta) V_{2it}(\theta)}{\partial \theta^T} - \sum_{t=1}^{T} U_{it} V_{2it} \sum_{s=1}^{T} \frac{\partial v_{is}^2(\theta)}{\partial \theta}}{\{\sum_{s=1}^{T} v_{is}(\theta)^2\}^2} \right| \\ \leq & \frac{1}{2n} \sum_{i=1}^{n} \frac{(\sum_{s=1}^{T} |v_{is}^2(\theta)| \otimes I_k) \sum_{t=1}^{T} |\frac{\partial U_{it}(\theta) V_{2it}(\theta)}{\partial \theta}| + \sum_{s=1}^{T} |\sum_{t=1}^{T} |U_{it} V_{2it}\| \frac{\partial v_{is}^2(\theta)}{\partial \theta}}{\{\sum_{s=1}^{T} v_{is}(\theta)^2\}^2} \end{aligned}$$

From the above as well as Condition 4, we have that

$$\begin{aligned}
|v_{is}^2(\theta)| &\geq m(z_{it})^2 \\
|v_{is}^2(\theta)| &\leq M(z_{it})^2 \\
|U_{it}(\theta)V_{2it}(\theta)| &\leq \{M(z_{it})^2 + M(z_{it})\}\boldsymbol{1}_k \left\{ M(z_{it}) + M(z_{it}) \frac{\sum_{s=1}^T M(z_{is})^2}{\sum_{s=1}^T m(z_{is})^2} \right\} \\
&\equiv \mathcal{A}_{it}\mathcal{B}_{it}
\end{aligned}$$

Now

$$\begin{aligned}
\left| \frac{\partial v_{is}^2(\theta)}{\partial \theta} \right| &= \left| 2v_{is}(\theta) \left\{ v_{is\theta}(\theta) - v_{is\alpha}(\theta) \frac{\sum_{q=1}^T u_{iq}(\theta) v_{iq}(\theta)}{\sum_{q=1}^T v_{iq}(\theta)^2} \right\} \right| \\
&\leq 2M(z_{is})^2 \left\{ 1 + \frac{\sum_{q=1}^T M(z_{iq})^2}{\sum_{q=1}^T m(z_{iq})^2} \right\} \boldsymbol{1}_k \equiv \mathcal{C}_{it} \qquad (2.66) \\
\left| \frac{\partial U_{it}(\theta)V_{2it}(\theta)}{\partial \theta} \right| &= \left| (V_{2it}(\theta) \otimes I_k) \frac{\partial U_{it}(\theta)}{\partial \theta} + U_{it}(\theta) \frac{\partial V_{2it}(\theta)}{\partial \theta} \right| \\
&\leq (|V_{2it}(\theta)| \otimes I_k) \left| \frac{\partial U_{it}(\theta)}{\partial \theta} \right| + |U_{it}(\theta)| \left| \frac{\partial V_{2it}(\theta)}{\partial \theta} \right| \qquad (2.67)
\end{aligned}$$

For the sake of brevity, we will suppress the θ argument of $g_{it}(\theta)$ such that $g_{it} \equiv g_{it}(\theta)$.

$$
\begin{aligned}
& \left|\frac{\partial U_{it}}{\partial \theta}\right| \\
= \; & |u_{it\theta} - u_{it\alpha}\frac{\sum_{s=1}^{T} u_{is}^{'} v_{is}}{\sum_{s=1}^{T} v_{is}^{2}} - \frac{\sum_{s=1}^{T} u_{is} v_{is}}{\sum_{s=1}^{T} v_{is}^{2}}\left(v_{it\theta} - v_{it\alpha}\frac{\sum_{s=1}^{T} u_{is}^{'} v_{is}}{\sum_{s=1}^{T} v_{is}^{2}}\right) \\
- \; & v_{it}\frac{\left\{\sum_{s=1}^{T}\left(u_{is\theta} - u_{is\alpha}\frac{\sum_{q=1}^{T} u_{iq}^{'} v_{iq}}{\sum_{q=1}^{T} v_{iq}^{2}}\right) v_{is}\right\}\sum_{s=1}^{T} v_{is}^{2}}{\{\sum_{s=1}^{T} v_{is}^{2}\}^{2}} \\
- \; & v_{it}\frac{\left\{\sum_{s=1}^{T} u_{is}\left(v_{is\theta} - v_{is\alpha}\frac{\sum_{q=1}^{T} u_{iq}^{'} v_{iq}}{\sum_{q=1}^{T} v_{iq}^{2}}\right)\right\}\sum_{s=1}^{T} v_{is}^{2}}{\{\sum_{s=1}^{T} v_{is}^{2}\}^{2}} \\
+ \; & v_{it}\frac{\sum_{s=1}^{T} u_{is} v_{is}\left\{\sum_{s=1}^{T} 2 v_{it}\left(v_{it\theta} - v_{it\alpha}\frac{\sum_{q=1}^{T} u_{iq}^{'} v_{iq}}{\sum_{q=1}^{T} v_{iq}^{2}}\right)\right\}}{\{\sum_{s=1}^{T} v_{is}^{2}\}^{2}}| \\
\leq \; & M(z_{it})\boldsymbol{1}_k \boldsymbol{1}_k^{'} + M(z_{it})\frac{\sum_{s=1}^{T} M(z_{is})^2}{\sum_{s=1}^{T} m(z_{is})^2}\boldsymbol{1}_k \boldsymbol{1}_k^{'} \\
+ \; & \left\{M(z_{it}) + M(z_{it})\frac{\sum_{s=1}^{T} M(z_{is})^2}{\sum_{s=1}^{T} m(z_{is})^2}\right\}\frac{\sum_{s=1}^{T} M(z_{is})^2}{\sum_{s=1}^{T} m(z_{is})^2}\boldsymbol{1}_k \boldsymbol{1}_k^{'} \\
+ \; & M(z_{it})\frac{\left\{\sum_{s=1}^{T} M(z_{is})^2 + M(z_{is})^2\frac{\sum_{q=1}^{T} M(z_{iq})^2}{\sum_{q=1}^{T} m(z_{iq})^2}\right\}\sum_{s=1}^{T} M(z_{is})^2}{\{\sum_{s=1}^{T} m(z_{is})^2\}^2}\boldsymbol{1}_k \boldsymbol{1}_k^{T} \\
+ \; & M(z_{it})\frac{\left\{\sum_{s=1}^{T} M(z_{is})^2 + M(z_{is})^2\frac{\sum_{q=1}^{T} M(z_{iq})^2}{\sum_{q=1}^{T} m(z_{iq})^2}\right\}\sum_{s=1}^{T} M(z_{is})^2}{\{\sum_{s=1}^{T} m(z_{is})^2\}^2}\boldsymbol{1}_k \boldsymbol{1}_k^{T} \\
+ \; & M(z_{it})\frac{\sum_{s=1}^{T} 2M(z_{is})^2\left(1 + \frac{\sum_{q=1}^{T} M(z_{iq})^2}{\sum_{q=1}^{T} m(z_{iq})^2}\right)\sum_{s=1}^{T} M(z_{is})^2}{\{\sum_{s=1}^{T} m(z_{is})^2\}^2}\boldsymbol{1}_k \boldsymbol{1}_k^{'} \equiv \mathcal{D}_{it}
\end{aligned}
$$

$$
\begin{aligned}
& \left|\frac{\partial V_{2it}}{\partial \theta}\right| \\
= & \left| v_{it\theta} - v_{it\alpha}\frac{\sum_{s=1}^{T} u_{is}^{'} v_{is}}{\sum_{s=1}^{T} v_{is}^{2}} + v_{it\alpha\theta} - v_{it\alpha\alpha}\frac{\sum_{s=1}^{T} u_{is}^{'} v_{is}}{\sum_{s=1}^{T} v_{is}^{2}} \right| \\
\leq & \left\{ M(z_{it}) + M(z_{it})\frac{\sum_{s=1}^{T} M(z_{is})^2}{\sum_{s=1}^{T} m(z_{is})^2} + M(z_{it}) + M(z_{it})\frac{\sum_{s=1}^{T} M(z_{is})^2}{\sum_{s=1}^{T} m(z_{is})^2} \right\} \boldsymbol{1}_k \\
\equiv & \mathcal{E}_{it}
\end{aligned}
$$

$$
\begin{aligned}
\left|\frac{\partial \bar{H}^{-1}(\theta)}{\partial \theta}\right| &= \left| (\bar{H}^{-1}(\theta) \otimes \bar{H}^{-1}(\theta)) \frac{1}{nT} \sum_{i=1}^{n}\sum_{t=1}^{T} 2(U_{it}(\theta) \otimes I_k) \frac{\partial U_{it}}{\partial \theta} \right| \\
&\leq |\bar{H}^{-1}(\theta) \otimes \bar{H}^{-1}(\theta)| \left| \frac{1}{nT} \sum_{i=1}^{n}\sum_{t=1}^{T} 2(U_{it}(\theta) \otimes I_k) \frac{\partial U_{it}(\theta)}{\partial \theta} \right| \\
&\leq |\bar{H}^{-1}(\theta)| \otimes |\bar{H}^{-1}(\theta)| \frac{1}{nT} \sum_{i=1}^{n}\sum_{t=1}^{T} 2(|U_{it}(\theta)| \otimes I_k) \left| \frac{\partial U_{it}(\theta)}{\partial \theta} \right|
\end{aligned}
$$

From the above, we have bounds for $|U_{it}(\theta)|$ and $|\partial U_{it}(\theta)/\partial\theta|$.

As for $|\bar{H}^{-1}(\theta)|$,

$$
\begin{aligned}
|\bar{H}^{-1}(\theta)| &\leq \left\{ \frac{1}{nT} \sum_{i=1}^{n}\sum_{t=1}^{T} \left(m(z_{it}) - M(z_{it})\frac{\sum_{s=1}^{T} M(z_{is})^2}{\sum_{s=1}^{T} m(z_{is})^2} \right)^2 \boldsymbol{1}_k \boldsymbol{1}_k^{'} \right\}^{-1} \\
&\equiv \mathcal{F}_{nT}^{-1}
\end{aligned}
$$

Hence,

$$
\begin{aligned}
& |\bar{H}^{-1}(\theta)\bar{b}(\theta) - \bar{H}^{-1}(\theta')\bar{b}(\theta')| \\
\leq\ & \mathcal{F}_{nT}^{-1}\frac{1}{2n}\sum_{i=1}^{n}\frac{\sum_{t=1}^{T}\{\mathcal{D}_{it}\mathcal{B}_{it}+\mathcal{E}_{it}\mathcal{A}_{it}\}\sum_{t=1}^{T}M(z_{it})^2}{\{\sum_{t=1}^{T}m(z_{it})^2\}^2}\|\theta-\theta'\| \\
+\ & \mathcal{F}_{nT}^{-1}\frac{1}{2n}\sum_{i=1}^{n}\frac{\sum_{t=1}^{T}\mathcal{C}_{it}\sum_{t=1}^{T}\mathcal{A}_{it}\mathcal{B}_{it}}{\{\sum_{t=1}^{T}m(z_{it})^2\}^2}\|\theta-\theta'\| \\
+\ & \left\{\left(\frac{1}{2n}\sum_{i=1}^{n}\frac{\sum_{t=1}^{T}\mathcal{A}_{it}\mathcal{B}_{it}}{\sum_{t=1}^{T}m(z_{it})^2}\right)^{T}\otimes I_k\right\}\{\mathcal{F}_{nT}^{-1}\otimes\mathcal{F}_{nT}^{-1}\} \\
& \times\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}2(\mathcal{A}_{it}\otimes I_k)\mathcal{D}_{it}\|\theta-\theta'\| \\
\equiv\ & \mathcal{Q}_{nT}\|\theta-\theta'\| \qquad (2.68)
\end{aligned}
$$

Let $B(\theta,\delta)$ denote a δ-ball around θ. It follows that

$$
\sup_{\theta\in\Theta}\sup_{\theta'\in B(\theta,\delta)}|\bar{H}^{-1}(\theta)\bar{b}(\theta)-\bar{H}^{-1}(\theta')\bar{b}(\theta')| \ < \ \mathcal{Q}_{nT}\delta = O_p(1)\delta,
$$

By the arbitrariness of δ, it follows that that $\bar{B}(\theta)$ is stochastically equicontinuous. □

Lemma 2.5.3 *Under Conditions 1, 2 and 4,*

$$
\sup_{\theta\in\Theta}\left\|\frac{\partial\bar{B}(\theta)}{\partial\theta}-\frac{\partial B(\theta)}{\partial\theta}\right\| = o_p(1)
$$

Proof Pointwise convergence follows by the law of large numbers. It remains to show stochastic equicontinuity. Hence, we need to find a dominating function for the second derivative of $\bar{B}(\theta)$ that only depends on z_{it} and is $O_p(1)$. In the following, we will use dominating function to mean a dominating function that

only depends on z_{it} and is $O_p(1)$.

$$\begin{aligned}
& \left|\frac{\partial^2 \bar{B}(\theta)}{\partial\theta\partial\theta^T}\right| \\
\leq\ & \left|\frac{\partial\{\bar{H}(\theta)^{-1}\partial\bar{b}(\theta)/\partial\theta\}}{\partial\theta^T}\right| + \left|\frac{\partial\{(\bar{b}(\theta)^T \otimes I_k)\partial\bar{H}^{-1}(\theta)/\partial\theta\}}{\partial\theta^T}\right| \qquad (2.69)
\end{aligned}$$

$$\begin{aligned}
& \left|\frac{\partial\{\bar{H}^{-1}(\theta)\partial\bar{b}(\theta)/\partial\theta\}}{\partial\theta^T}\right| \\
\leq\ & \left|\left(\left\{\frac{\partial\bar{b}(\theta)}{\partial\theta}\right\}^T \otimes I_k\right)\right| \left|\frac{\partial\bar{H}^{-1}(\theta)}{\partial\theta^T}\right| + \left|\bar{H}^{-1}(\theta)\otimes I_k\right| \left|\frac{\partial^2\bar{b}(\theta)}{\partial\theta\partial\theta^T}\right| \qquad (2.70)
\end{aligned}$$

$$\begin{aligned}
& \left|\frac{\partial\{(\bar{b}(\theta)^T \otimes I_k)\partial\bar{H}^{-1}(\theta)/\partial\theta\}}{\partial\theta^T}\right| \\
\leq\ & \left|\left(\left\{\frac{\partial\bar{H}^{-1}(\theta)}{\partial\theta}\right\}^T \otimes I_k\right)\frac{\partial\{\bar{b}(\theta)^T\otimes I_k\}}{\partial\theta^T}\right| \\
+\ & \left|(I_k\otimes\bar{b}(\theta)^T\otimes I_k)\frac{\partial^2\bar{H}^{-1}(\theta)}{\partial\theta\partial\theta^T}\right| \qquad (2.71)
\end{aligned}$$

The proof of Lemma 2.5.2 gives dominating functions for the above expressions except for the second-order derivatives of $\bar{H}(\theta)^{-1}$ and $\bar{b}(\theta)$. So it remains to show $O_p(1)$ dominating functions for both derivatives.

$$\begin{aligned}
& \left|\frac{\partial^2\bar{H}^{-1}(\theta)}{\partial\theta\partial\theta^T}\right| \\
=\ & \left|\frac{\partial\{\partial\bar{H}^{-1}(\theta)/\partial\theta\}}{\partial\theta^T}\right| \\
=\ & \left|\frac{\partial\left\{(\bar{H}^{-1}(\theta)\otimes\bar{H}^{-1}(\theta))\frac{1}{nT}\sum_{i=1}^n\sum_{t=1}^T 2(U_{it}(\theta)\otimes I_k)\frac{\partial U_{it}}{\partial\theta}\right\}}{\partial\theta^T}\right|
\end{aligned}$$

$$\leq \left| \left\{ \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} 2(U_{it}(\theta) \otimes I_k) \frac{\partial U_{it}}{\partial \theta} \right\}^T \otimes I_{k^2} \right| \left| \frac{\partial \bar{H}^{-1}(\theta) \otimes \bar{H}^{-1}(\theta)}{\partial \theta^T} \right|$$
$$+ \quad |I_k \otimes \bar{H}^{-1}(\theta) \otimes \bar{H}^{-1}(\theta)| \left| \frac{\partial \left\{ \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} 2(U_{it}(\theta) \otimes I_k) \frac{\partial U_{it}}{\partial \theta} \right\}}{\partial \theta^T} \right| \quad (2.72)$$

The proof of Lemma 2.5.2 gives dominating functions for the expressions in (2.72) except for the derivatives of $\bar{H}^{-1}(\theta) \otimes \bar{H}^{-1}(\theta)$ and $\sum_{i=1}^{n} \sum_{t=1}^{T} 2(U_{it}(\theta) \otimes I_k) \partial U_{it} / \partial \theta$ $/nT$.

$$\left| \frac{\partial \bar{H}^{-1}(\theta) \otimes \bar{H}^{-1}(\theta)}{\partial \theta^T} \right|$$
$$= \left| \frac{\partial (\bar{H}^{-1}(\theta) \otimes I_k)(I_k \otimes \bar{H}^{-1}(\theta))}{\partial \theta^T} \right|$$
$$\leq \left| ((I_k \otimes \bar{H}^{-1}(\theta))^T \otimes I_{k^2}) \frac{\partial \bar{H}^{-1}(\theta) \otimes I_k}{\partial \theta^T} \right| + \left| (I_{k^2} \otimes \bar{H}^{-1}(\theta) \otimes I_k) \frac{\partial I_k \otimes \bar{H}^{-1}(\theta)}{\partial \theta^T} \right|$$

Since we have a dominating function for $\bar{H}^{-1}(\theta)$ and $\partial \bar{H}^{-1}(\theta) / \partial \theta$ in the proof of Lemma 2.5.2, it follows that the above term also has a dominating function.

$$\left| \frac{\partial \left\{ \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} 2(U_{it}(\theta) \otimes I_k) \frac{\partial U_{it}}{\partial \theta} \right\}}{\partial \theta^T} \right|$$
$$\leq \quad \frac{2}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \left| \left(\left\{ \frac{\partial U_{it}}{\partial \theta} \right\}^T \otimes I_{k^2} \right) \frac{\partial U_{it}(\theta) \otimes I_k}{\partial \theta^T} \right|$$
$$+ \quad \left| (I_k \otimes U_{it}(\theta) \otimes I_k) \frac{\partial^2 U_{it}(\theta)}{\partial \theta \partial \theta^T} \right| \quad (2.73)$$

Except for $\partial^2 U_{it}(\theta) / \partial \theta \partial \theta^T$, there are dominating functions for all other terms given

in the proof of Lemma 2.5.2.

$$
\begin{aligned}
&\left|\frac{\partial^2 \bar{b}(\theta)}{\partial\theta\partial\theta^T}\right| \\
= &\left|\frac{\partial\left(\frac{1}{2n}\sum_{i=1}^{n}\frac{(\sum_{s=1}^{T}v_{is}^2(\theta)\otimes I_k)\sum_{t=1}^{T}\frac{\partial U_{it}(\theta)V_{2it}(\theta)}{\partial\theta}+\sum_{t=1}^{T}|U_{it}V_{2it}\sum_{s=1}^{T}\frac{\partial v_{is}(\theta)^2}{\partial\theta}}{\{\sum_{s=1}^{T}v_{is}(\theta)^2\}^2}\right)}{\partial\theta^T}\right| \\
\leq &\left|\frac{\partial\left(\frac{1}{2n}\sum_{i=1}^{n}\frac{(\sum_{s=1}^{T}v_{is}^2(\theta)\otimes I_k)\sum_{t=1}^{T}\frac{\partial U_{it}(\theta)V_{2it}(\theta)}{\partial\theta}}{\{\sum_{s=1}^{T}v_{is}(\theta)^2\}^2}\right)}{\partial\theta^T}\right| \qquad (2.74) \\
+ &\left|\frac{\partial\left(\frac{1}{2n}\sum_{i=1}^{n}\frac{\sum_{t=1}^{T}U_{it}V_{2it}\sum_{s=1}^{T}\frac{\partial v_{is}^2(\theta)}{\partial\theta}}{\{\sum_{s=1}^{T}v_{is}(\theta)^2\}^2}\right)}{\partial\theta^T}\right| \qquad (2.75)
\end{aligned}
$$

(2.75) has dominating functions given in the proof of Lemma 2.5.2. (2.74) can be manipulated as follows:

$$
\begin{aligned}
&\left|\partial\left(\frac{1}{2n}\sum_{i=1}^{n}\frac{(\sum_{s=1}^{T}v_{is}^2\otimes I_k)(\theta)\sum_{t=1}^{T}\frac{\partial U_{it}(\theta)V_{2it}(\theta)}{\partial\theta}}{\{\sum_{s=1}^{T}v_{is}(\theta)^2\}^2}\right)/\partial\theta^T\right| \\
= &\left|\partial\left(\frac{1}{2n}\sum_{i=1}^{n}\sum_{t=1}^{T}(\{\sum_{s=1}^{T}v_{is}(\theta)^2\}^{-1}\otimes I_k)\frac{\partial U_{it}(\theta)V_{2it}(\theta)}{\partial\theta}\right)/\partial\theta^T\right| \\
\leq &\frac{1}{2n}\sum_{i=1}^{n}\left|(\{\sum_{s=1}^{T}v_{is}(\theta)^2\}^{-1}\otimes I_{k^2})\right|\left|\frac{\partial\sum_{t=1}^{T}\partial U_{it}(\theta)V_{2it}(\theta)/\partial\theta}{\partial\theta^T}\right| \qquad (2.76) \\
+ &\frac{1}{2n}\sum_{i=1}^{n}\left|\left(\left\{\sum_{t=1}^{T}\frac{\partial U_{it}(\theta)V_{2it}(\theta)}{\partial\theta}\right\}^T\otimes I_k\right)\frac{\partial\{\sum_{s=1}^{T}v_{is}(\theta)^2\}^{-1}\otimes I_k}{\partial\theta}\right|
\end{aligned}
$$

The proof of Lemma 2.5.2 provides dominating functions for all of the above except

for the second term in the first summand (2.76).

$$\begin{aligned} & \left| \frac{\partial \{\sum_{t=1}^{T} \partial U_{it}(\theta) V_{2it}(\theta)/\partial\theta\}}{\partial \theta^T} \right| \\ \leq & \left| \frac{\partial \{\sum_{t=1}^{T} (V_{2it}(\theta) \otimes I_k) \frac{\partial U_{it}(\theta)}{\partial \theta} + U_{it}(\theta) \frac{\partial V_{2it}(\theta)}{\partial \theta}\}}{\partial \theta^T} \right| \\ \leq & \left| \sum_{t=1}^{T} \frac{\partial \left\{ (V_{2it}(\theta) \otimes I_k) \frac{\partial U_{it}(\theta)}{\partial \theta} \right\}}{\partial \theta^T} \right| + \left| \sum_{t=1}^{T} \frac{\partial \left\{ U_{it}(\theta) \frac{\partial V_{2it}(\theta)}{\partial \theta} \right\}}{\partial \theta^T} \right| \end{aligned} \tag{2.77}$$

Now manipulating each term in (2.77) separately,

$$\begin{aligned} & \left| \sum_{t=1}^{T} \frac{\partial \left\{ (V_{2it}(\theta) \otimes I_k) \frac{\partial U_{it}(\theta)}{\partial \theta} \right\}}{\partial \theta^T} \right| \\ \leq & \sum_{t=1}^{T} \left| (V_{2it}(\theta) \otimes I_{k^2}) \frac{\partial^2 U_{it}(\theta)}{\partial \theta \partial \theta^T} \right| + \sum_{t=1}^{T} \left| \left(\left\{ \frac{\partial U_{it}(\theta)}{\partial \theta} \right\}^T \otimes I_k \right) \frac{\partial V_{2it}(\theta)}{\partial \theta} \right| \end{aligned} \tag{2.78}$$

$$\begin{aligned} & \left| \sum_{t=1}^{T} \frac{\partial \left\{ U_{it}(\theta) \frac{\partial V_{2it}(\theta)}{\partial \theta} \right\}}{\partial \theta^T} \right| \\ \leq & \sum_{t=1}^{T} \left| \left\{ \frac{\partial V_{2it}(\theta)}{\partial \theta} \right\}^T \otimes I_k \right| \left| \frac{\partial U_{it}(\theta)}{\partial \theta^T} \right| + \sum_{t=1}^{T} |I_k \otimes U_{it}(\theta)| \left| \frac{\partial^2 V_{2it}(\theta)}{\partial \theta \partial \theta^T} \right| \end{aligned} \tag{2.79}$$

All terms in (2.78) and (2.79) have dominating functions given in the proof of B.2, except for $\partial^2 U_{it}(\theta)/\partial\theta\partial\theta^T$ and $\partial^2 V_{2it}(\theta)/\partial\theta\partial\theta^T$. For the sake of brevity, we will suppress the θ argument of $g_{it}(\theta)$ such that $g_{it} \equiv g_{it}(\theta)$. To simplify the derivation below, let

$$\begin{aligned} & \left| \frac{\partial \hat{\alpha}_i(\theta)}{\partial \theta} \right| \\ = & \left| -\frac{\sum_{s=1}^{T} v_{is} u_{is}^T}{\sum_{s=1}^{T} v_{is}^2} \right| \leq \frac{\sum_{s=1}^{T} M(z_{is})^2}{\sum_{s=1}^{T} m(z_{is})^2} \boldsymbol{1}_k^T \equiv \mathcal{G}_{1,i} \end{aligned}$$

$$
\begin{aligned}
&\left|\frac{\partial^2\hat{\alpha}_i(\theta)}{\partial\theta\partial\theta^T}\right| \\
= &\Bigg|-\frac{\left\{\sum_{s=1}^T(v_{is}\otimes I_k)\left(u_{is\theta}+u_{is\alpha}\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta}\right)^T+\sum_{s=1}^T u_{is}\left(v_{is\theta}+v_{is\alpha}\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta}\right)\right\}\sum_{s=1}^T v_{is}^2}{\{\sum_{s=1}^T v_{is}^2\}^2} \\
&+\frac{\sum_{s=1}^T u_{is}v_{is}\left\{\sum_{s=1}^T 2v_{is}\left(v_{is\theta}+v_{is\alpha}\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta}\right)\right\}}{\{\sum_{s=1}^T v_{is}^2\}^2}\Bigg| \\
\leq &\frac{\left\{\sum_{s=1}^T M(z_{is})^2+M(z_{is})^2\frac{\sum_{q=1}^T M(z_{iq})^2}{\sum_{q=1}^T m(z_{iq})^2}+\sum_{s=1}^T M(z_{is})^2+M(z_{is})^2\frac{\sum_{q=1}^T M(z_{iq})^2}{\sum_{q=1}^T m(z_{iq})^2}\right\}}{\sum_{s=1}^T m(z_{is})^2} \\
&\times\boldsymbol{1}_k\boldsymbol{1}_k^T \\
+ &\frac{\sum_{s=1}^T 2M(z_{is})^2\left(1+\frac{\sum_{q=1}^T M(z_{iq})^2}{\sum_{q=1}^T m(z_{iq})^2}\right)\sum_{s=1}^T M(z_{is})^2}{\{\sum_{s=1}^T m(z_{is})^2\}^2}\boldsymbol{1}_k\boldsymbol{1}_k^T\equiv\mathcal{G}_{2,i}
\end{aligned}
$$

$$
\begin{aligned}
&\left|\frac{\partial^3\hat{\alpha}_i(\theta)}{\partial\theta\partial\theta^T\partial\theta}\right| \\
= &\left|-\frac{\sum_{s=1}^T\left(u_{is\theta}+u_{is\alpha}\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta}\right)\otimes I_k\left\{\left(v_{is\theta}+v_{is\alpha}\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta}\right)\otimes vec(I_k)\right\}}{\sum_{s=1}^T v_{is}^2}\right| \\
- &\frac{\sum_{s=1}^T v_{is}\otimes I_{k^2}\left(u_{is\theta\theta}+u_{is\theta\alpha}\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta}+\left(u_{is\theta\alpha}+\left(I_k\otimes\left\{\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta}\right\}^T\right)u_{is\alpha\alpha}\right)\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta}\right)}{\sum_{s=1}^T v_{is}^2} \\
- &\frac{\sum_{s=1}^T v_{is}\otimes I_{k^2}(u_{is\alpha}\otimes I_k)\frac{\partial^2\hat{\alpha}_i(\theta)}{\partial\theta\partial\theta^T}}{\sum_{s=1}^T v_{is}^2} \\
- &\frac{\sum_{s=1}^T\left\{\left(v_{is\theta}+v_{is\alpha}\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta}\right)^T\otimes I_k\right\}\left(u_{is\theta}+u_{is\alpha}\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta}\right)}{\sum_{s=1}^T v_{is}^2} \\
- &\frac{\sum_{s=1}^T(I_k\otimes u_{is})\left(v_{is\theta\theta}+v_{is\theta\alpha}\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta}+\left\{\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta}\right\}^T(v_{is\alpha\theta}+v_{is\alpha\alpha}\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta})\right)}{\sum_{s=1}^T v_{is}^2} \\
- &\frac{\sum_{s=1}^T(I_k\otimes u_{is})(v_{is\alpha}\otimes I_k)\frac{\partial^2\hat{\alpha}_i(\theta)}{\partial\theta\partial\theta^T}}{\sum_{s=1}^T v_{is}^2}
\end{aligned}
$$

$$
\begin{aligned}
&+ \frac{\left\{\sum_{s=1}^{T}(v_{is}\otimes I_k)\left(u_{is\theta}+u_{is\alpha}\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta}\right)^{T}+\sum_{s=1}^{T}u_{is}\left(v_{is\theta}+v_{is\alpha}\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta}\right)\right\}\otimes I_k}{\{\sum_{s=1}^{T}v_{is}^2\}^2}\\
&\quad\times\left\{\sum_{s=1}^{T}2v_{is}\left(v_{is\theta}+v_{is\alpha}\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta}\right)\otimes vec(I_k)\right\}\\
&+ \frac{\sum_{s=1}^{T}2v_{is}\left\{\left(v_{is\theta}+v_{is\alpha}\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta}\right)^{T}\otimes I_k\right\}\sum_{s=1}^{T}\left(u_{is\theta}+u_{is\alpha}\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta}\right)v_{is}}{\{\sum_{s=1}^{T}v_{is}^2\}^2}\\
&+ \frac{\sum_{s=1}^{T}u_{is}\left(v_{is\theta}+v_{is\alpha}\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta}\right)}{\{\sum_{s=1}^{T}v_{is}^2\}^2}\\
&+ \frac{\left\{I_k\otimes\sum_{s=1}^{T}u_{is}v_{is}\right\}\sum_{s=1}^{T}2\left\{\left(v_{is\theta}+v_{is\alpha}\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta}\right)^{T}\left(v_{is\theta}+v_{is\alpha}\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta}\right)\right\}}{\{\sum_{s=1}^{T}v_{is}^2\}^2}\\
&+ \left\{I_k\otimes\sum_{s=1}^{T}u_{is}v_{is}\right\}\\
&\quad\times\frac{\sum_{s=1}^{T}2\left\{v_{is}\left(v_{is\theta\theta}+v_{is\theta\alpha}\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta}+\left\{\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta}\right\}^{T}(v_{is\alpha\theta}+v_{is\alpha\alpha}\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta})\right)\right\}}{\{\sum_{s=1}^{T}v_{is}^2\}^2}\\
&+ \left\{I_k\otimes\sum_{s=1}^{T}u_{is}v_{is}\right\}\frac{\sum_{s=1}^{T}2(v_{is\alpha}\otimes I_k)\frac{\partial^2\hat{\alpha}_i(\theta)}{\partial\theta\partial\theta^T}}{\{\sum_{s=1}^{T}v_{is}^2\}^2}\\
&- \left\{\sum_{s=1}^{T}u_{is}v_{is}\sum_{s=1}^{T}2v_{is}\left(v_{is\theta}+v_{is\alpha}\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta}\right)\right\}^{T}\otimes I_k\\
&\quad\times\frac{\left\{2\sum_{s=1}^{T}v_{is}^2(\theta)\sum_{s=1}^{T}2v_{is}\left(v_{it\theta}+v_{it\alpha}\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta}\right)\otimes vec(I_k)\right\}}{\{\sum_{s=1}^{T}v_{is}^2\}^4}|\\
&\leq \mathcal{G}_{3,i} \qquad (2.80)
\end{aligned}
$$

All of the quantities in the numerator are bounded above by products or sums of $M(z_{is})^2$, $\mathcal{G}_{1,i}$, or $\mathcal{G}_{2,i}$, which are all $O_p(1)$ by condition 5.i. The denominator can be bounded below by sums of $m(z_{is})^2$, the inverse of which is $O_p(1)$ by condition

5.ii. The last inequality hence follows, where $\sup_i \mathcal{G}_{3,i}$ is $O_p(1)$.

$$
\begin{aligned}
& \left|\frac{\partial^2 U_{it}}{\partial\theta\partial\theta^T}\right| \\
= & \left|\frac{\partial\{u_{it\theta}+u_{it\alpha}\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta}+(v_{it}\otimes I_k)\frac{\partial^2\hat{\alpha}_i(\theta)}{\partial\theta\partial\theta^T}+\left\{\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta}\right\}^T(v_{it\theta}+v_{it\alpha}\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta})\}}{\partial\theta}\right| \\
\leq & |u_{it\theta\theta}|+|u_{it\theta\alpha}\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta}|+\left|\left\{\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta}\right\}^T\otimes I_k\left\{u_{it\alpha\theta}+u_{it\alpha\alpha}\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta}\right\}\right| \\
+ & \left|(I_k\otimes u_{it\alpha})\frac{\partial^2\hat{\alpha}_i(\theta)}{\partial\theta\partial\theta^T}\right| \\
+ & \left|\left\{\frac{\partial^2\hat{\alpha}_i(\theta)}{\partial\theta\partial\theta^T}\right\}^T\otimes I_k\left\{\left(v_{it\theta}+v_{it\alpha}\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta}\right)\otimes vec(I_k)\right\}\right|+\left|(v_{it}\otimes I_{k^2})\frac{\partial^3\hat{\alpha}_i(\theta)}{\partial\theta\partial\theta^T\partial\theta}\right| \\
+ & \left|(v_{it\theta}+v_{it\alpha}\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta})^T\otimes I_k\left\{\frac{\partial^2\hat{\alpha}_i(\theta)}{\partial\theta\partial\theta^T}\right\}^T\right| \\
+ & \left|\left(I_k\otimes\left\{\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta}\right\}^T\right)\left(v_{it\theta\theta}+v_{it\theta\alpha}\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta}+\left\{\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta}\right\}^T\left(v_{it\alpha\theta}+v_{it\alpha\alpha}\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta}\right)\right)\right. \\
& \left.+\left(I_k\otimes\left\{\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta}\right\}^T\right)(v_{it\alpha}\otimes I_k)\frac{\partial^2\hat{\alpha}_i(\theta)}{\partial\theta\partial\theta^T}\right|
\end{aligned}
$$

$$
\begin{aligned}
\leq & M(z_{it})1_{k^2}\boldsymbol{1}_k^T+M(z_{it})1_{k^2}\mathcal{G}_{1,i}+(\mathcal{G}_{1,i}^T\otimes I_k)(M(z_{it})\boldsymbol{1}_k\boldsymbol{1}_k^T \\
+ & M(z_{it})\boldsymbol{1}_k\mathcal{G}_{1,i})+(I_k\otimes M(z_{it})\boldsymbol{1}_k)\mathcal{G}_{2,i}+\mathcal{G}_{2,i}^T\otimes I_k\{(M(z_{it})\boldsymbol{1}_k^T \\
+ & M(z_{it})\mathcal{G}_{1,i})\otimes vec(I_k)\}+(M(z_{it})\otimes I_{k^2})\mathcal{G}_{3,i}+\{(M(z_{it})\boldsymbol{1}_k^T+M(z_{it})\mathcal{G}_{1,i}^T)^T\otimes I_k\}\mathcal{G}_{2,i}^T \\
+ & I_k\otimes\mathcal{G}_{1,i}^T(M(z_{it})\boldsymbol{1}_k\boldsymbol{1}_k^T+M(z_{it})\boldsymbol{1}_k\mathcal{G}_{1,i}+\mathcal{G}_{1,i}^T(M(z_{it}\boldsymbol{1}_k^T \\
+ & M(z_{it})\mathcal{G}_{1,i})+(M(z_{it})\otimes I_k)\mathcal{G}_{2,i})
\end{aligned}
$$

$$
\begin{aligned}
&\left|\frac{\partial^2 V_{2it}(\theta)}{\partial\theta\partial\theta^T}\right| \\
\leq\ & |v_{it\theta\theta}| + \left|v_{it\theta\alpha}\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta}\right| + \left|\left(\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta}\right)^T\left(v_{it\alpha\theta} + v_{it\alpha\alpha}\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta}\right)\right| + \left|v_{it\alpha}\frac{\partial^2\hat{\alpha}_i(\theta)}{\partial\theta\partial\theta^T}\right| \\
+\ & |v_{it\alpha\theta\theta}| + \left|v_{it\alpha\theta\alpha}\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta}\right| + \left|\left(\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta}\right)^T\left(v_{it\alpha\alpha\theta} + v_{it\alpha\alpha\alpha}\frac{\partial\hat{\alpha}_i(\theta)}{\partial\theta}\right)\right| + \left|v_{it\alpha\alpha}\frac{\partial^2\hat{\alpha}_i(\theta)}{\partial\theta\partial\theta^T}\right| \\
\leq\ & M(z_{it})\boldsymbol{1}_k\boldsymbol{1}_k^T + M(z_{it})\boldsymbol{1}_k\mathcal{G}_{1,i} + \mathcal{G}_{1,i}^T(M(z_{it})\boldsymbol{1}_k^T + M(z_{it})\mathcal{G}_{2,i}) + M(z_{it})\mathcal{G}_{2,i} \\
+\ & M(z_{it})\boldsymbol{1}_k\boldsymbol{1}_k^T + M(z_{it})\boldsymbol{1}_k\mathcal{G}_{1,i} + \mathcal{G}_{1,i}^T(M(z_{it})\boldsymbol{1}_k^T \\
+\ & M(z_{it})\mathcal{G}_{1,i}) + M(z_{it})\mathcal{G}_{2,i}
\end{aligned}
$$

Back to (2.73), (2.78) and (2.79), we see that each summand has a dominating function which consists of sums and products of $M(z_{it})^2$, $\mathcal{G}_{1,i}$, $\mathcal{G}_{2,i}$, and $\mathcal{G}_{3,i}$. Plugging the dominating functions for (2.73), (2.78) and (2.79) into (2.72) and (2.76), respectively, we find that both (2.72) and (2.76) have dominating functions. Hence, the result follows. □

2.5.6 Proof of Proposition 2.3.2

Proof As $n, T \to \infty$ and $n/T \to \rho$, Theorem 1 and 2 in HN04 imply that

$$
\begin{aligned}
\sqrt{nT}(\hat{\theta}_{ML} - \theta_0) &\xrightarrow{d} Z_{ML} \sim N(B\sqrt{\rho}, [\bar{E}[\mathcal{I}_i]]^{-1}), && (2.81)\\
\sqrt{nT}(\hat{\theta}_{BC} - \theta_0) &\xrightarrow{d} Z_{BC} \sim N(0, [\bar{E}[\mathcal{I}_i]]^{-1}) && (2.82)\\
\frac{\sqrt{nT}\bar{B}(\hat{\theta}_{ML})}{T} &\xrightarrow{p} Z_B^\rho \equiv B\sqrt{\rho} && (2.83)
\end{aligned}
$$

where B and $[\bar{E}[\mathcal{I}_i]]^{-1}$ are given in (2.15) and (2.13), respectively.

The random variables on the right-hand side of (2.81), (2.82), and (2.83) characterize the limit experiment of the random variables on their left-hand side. First, we characterize the MSE-minimizing weighting matrix, Λ_0, in the limit experiment.

Let $Q(\Lambda) = E\|Z_{ML} - \Lambda Z_B^{\rho}\|_T$. The first-order conditions imply that

$$\Lambda_0 \;=\; \{E\|Z_B^{\rho}\|_T\}^{-1} E[Z_{ML}\{Z_B^{\rho}\}^T]. \tag{2.84}$$

By Le Cam's Limit of Experiments Theory (der Vaart, 1998), we can write the above in terms of the random variables in the sequence of experiments to obtain Λ_{nT}:

$$\Lambda_{nT} \;=\; \left\{ E \left\| \frac{\sqrt{nT}\bar{B}(\hat{\theta}_{ML})}{T} \right\|_T \right\}^{-1} E\left[\sqrt{nT}(\hat{\theta}_{ML} - \theta_0)\sqrt{nT}\bar{B}(\hat{\theta}_{ML})^T \right]. \tag{2.85}$$

The numerator of (2.85) can be simplified as follows:

$$E[\sqrt{nT}(\hat{\theta}_{ML} - \theta_0)] E\left[\sqrt{\frac{n}{T}} \bar{B}(\hat{\theta}_{ML})^T \right] + cov\left(\sqrt{nT}(\hat{\theta}_{ML} - \theta_0), \sqrt{\frac{n}{T}} B(\hat{\theta}_{ML})^T \right)$$

A similar manipulation of the denominator of (2.85) leads to

$$\begin{aligned} & E \left\| \sqrt{n}T\bar{B}(\hat{\theta}_{ML}) \right\|_T \\ = \;& \left\| E\left[\sqrt{n}T\bar{B}(\hat{\theta}_{ML}) \right] \right\|_T + var\left(\sqrt{nT} \frac{\bar{B}(\hat{\theta}_{ML})}{T} \right) \end{aligned}$$

By Lemma 2.5.1 and 2.5.2, the δ-method can be applied to derive $\hat{\Lambda}_{nT}$, a plug-in estimator of Λ_{nT}. The consistency of $\hat{\Lambda}_{nT}$ follows from the consistency of $\{\hat{E}[\mathcal{I}_i]\}^{-1}$ for $\{\bar{E}[\mathcal{I}_i]\}^{-1}$, the consistency $\partial\bar{B}(\hat{\theta}_{ML})/\partial\theta$ for $\partial B/\partial\theta$, implied by the uniform law of large numbers and Lemma 2.5.3, and the Continuous Mapping Theorem. Now note that as $n, T \to \infty$, $n/T \to \rho$, $\Lambda_{nT} \xrightarrow{p} \Lambda_0 \equiv I_k$, since $E[Z_{ML}\{Z_B^{\rho}\}^T] = E\|Z_B^{\rho}\|_T$ in the limit experiment. It follows that $\hat{\Lambda}_{nT} \xrightarrow{p} I_k$.

Let $Q_{nT}(\Lambda) \equiv E\|\sqrt{nT}(\hat{\theta}_{SH}(\Lambda) - \theta_0)\|_T$ and $\|.\|$ denote the Frobenius norm. It

remains to show that

$$\|Q_{nT}(\hat{\Lambda}_{nT}) - min_{\Lambda\in\mathcal{L}} Q_{nT}(\Lambda)\| = o_p(1) \tag{2.86}$$

$$\begin{aligned} & \|Q_{nT}(\hat{\Lambda}_{nT}) - min_{\Lambda} Q_{nT}(\Lambda)\| \\ \leq\ & \|Q_{nT}(\hat{\Lambda}_{nT}) - Q(\hat{\Lambda}_{nT})\| + \|Q(\hat{\Lambda}_{nT}) - Q(\Lambda_0)\| + \|Q(\Lambda_0) - min_{\Lambda} Q_{nT}(\Lambda)\| \\ \leq\ & \sup_{\Lambda} \|Q_{nT}(\Lambda) - Q(\Lambda)\| + \|Q(\Lambda_0) - min_{\Lambda} Q_{nT}(\Lambda)\| + o_p(1) \end{aligned} \tag{2.87}$$

where the second term of the penultimate inequality is $o_p(1)$ by the consistency of $\hat{\Lambda}_{nT}$ and the continuity of Q.

Now, we can find upper and lower bounds for the second term,

$$Q(\Lambda_0) - \min_{\Lambda} Q_{nT}(\Lambda) \leq Q(\Lambda_{nT}) - Q_{nT}(\Lambda_{nT}) \tag{2.88}$$

$$Q(\Lambda_0) - \min_{\Lambda\in\mathcal{L}} Q_{nT}(\Lambda) \geq Q(\Lambda_0) - Q_{nT}(\Lambda_0) \tag{2.89}$$

Hence, $|Q(\Lambda_0) - \min_{\Lambda} Q_{nT}(\Lambda)| \leq \sup_{\Lambda\in\mathcal{L}} |Q_{nT}(\Lambda) - Q(\Lambda)|$. Plugging this into (2.87) yields,

$$\begin{aligned} & |Q_{nT}(\hat{\Lambda}_{nT}) - \min_{\Lambda} Q_{nT}(\Lambda)| \\ \leq\ & 2\sup_{\Lambda} |Q_{nT}(\Lambda) - Q(\Lambda)| + o_p(1) \end{aligned} \tag{2.90}$$

Hence, we need to make sure that a uniform law of large numbers applies. To do that, we need weak convergence and stochastic equicontinuity of $Q_{nT}(\Lambda)$.

First, we show that $Q_{nT}(\Lambda) \xrightarrow{p} Q(\Lambda)$. Let $\hat{Z}^{nT}_{ML} \equiv \sqrt{nT}(\hat{\theta}_{ML} - \theta_0)$ and $\hat{Z}^{nT}_{B} \equiv$

$\sqrt{nT}\bar{B}(\hat{\theta}_{ML})/T$.

$$\begin{aligned} Q_{nT}(\Lambda) &= E\|\hat{Z}_{ML}^{nT} - \Lambda\hat{Z}_{B}^{nT}\|_{T} \\ &= E\|\hat{Z}_{ML}^{nT}\|_{T} - 2\Lambda E[\hat{Z}_{B}^{nT}\,\hat{Z}_{ML}^{nT}] + \Lambda E\|\hat{Z}_{B}^{nT}\|_{T}\Lambda^{T} \end{aligned} \tag{2.91}$$

The conditions that ensure $\hat{\Lambda}_{nT} \xrightarrow{p} \Lambda_0$ ensure that $Q_{nT}(\Lambda) \xrightarrow{p} Q(\Lambda)$. It remains to show that $(Q_{nT}(\Lambda) - Q(\Lambda))$ is stochastic equicontinuous in Λ.

$$\limsup_{n,T\to\infty} P(\sup_{\Lambda\in\Lambda}\sup_{\Lambda'\in B(\Lambda,\delta)} \|(Q_{nT}(\Lambda) - Q(\Lambda)) - (Q_{nT}(\Lambda') - Q(\Lambda'))\| \geq \epsilon) < \epsilon$$

$$\begin{aligned} &\sup_{\Lambda\in\Lambda}\sup_{\Lambda'\in B(\Lambda,\delta)} \|Q_{nT}(\Lambda) - Q_{nT}(\Lambda') - (Q(\Lambda) - Q(\Lambda'))\| \\ =\ &\sup_{\Lambda\in\Lambda}\sup_{\Lambda'\in B(\Lambda,\delta)} \|Q_{nT}(\Lambda) - Q_{nT}(\Lambda') && (2.92) \\ -\ &(Q(\Lambda) - Q(\Lambda'))\| && (2.93) \end{aligned}$$

For (2.92)

$$\begin{aligned} &E\|\hat{Z}_{ML}^{nT} - \Lambda\hat{Z}_{B}^{nT}\|_{T} - E\|\hat{Z}_{ML}^{nT} - \Lambda'\hat{Z}_{B}^{nT}\|_{T} \\ =\ &2(\Lambda' - \Lambda)E[\hat{Z}_{ML}^{nT}\{\hat{Z}_{B}^{nT}\}^{T}] + \Lambda E\|\hat{Z}_{B}^{nT}\|_{T}\Lambda^{T} - \Lambda' E\|\hat{Z}_{B}^{nT}\|_{T}\Lambda'^{T} \end{aligned} \tag{2.94}$$

Applying a similar manipulation to (2.93), we obtain the following:

$$\begin{aligned} &\sup_{\Lambda\in\Lambda}\sup_{\Lambda'\in B(\Lambda,\delta)} \|Q_{nT}(\Lambda) - Q_{nT}(\Lambda') - (Q(\Lambda) - Q(\Lambda'))\| \\ =\ &\sup_{\Lambda\in\Lambda}\sup_{\Lambda'\in B(\Lambda,\delta)} \|(\Lambda' - \Lambda)\{E[\hat{Z}_{ML}^{nT}\{\hat{Z}_{B}^{nT}\}^{T}] - E[Z_{ML}\{Z_{B}^{\rho}\}^{T}]\} \\ +\ &\Lambda\{E\|\hat{Z}_{B}^{nT}\|_{T} - E\|Z_{B}^{\rho}\|_{T}\}\Lambda^{T} - \Lambda'\}E\|\hat{Z}_{B}^{nT}\|_{T} - E\|Z_{B}^{\rho}\|_{T}\}\Lambda'^{T}\| \end{aligned} \tag{2.95}$$

$$\leq \sup_{\Lambda\in\Lambda}\sup_{\Lambda'\in B(\Lambda,\delta)} \|\Lambda'-\Lambda\|\|E[\hat{Z}_{ML}^{nT}\{\hat{Z}_B^{nT}\}^T]-E[Z_{ML}\{Z_B^{\rho}\}^T]\|$$

$$+ \quad \|\Lambda\|^2\|\|E\|\hat{Z}_B^{nT}\|_T - E\|Z_B^{\rho}\|_T\| + \|\Lambda'\|^2\|E\|\hat{Z}_B^{nT}\|_T - E\|Z_B^{\rho}\|_T\| \quad (2.96)$$

$$\leq \quad \delta k^2 o(1) + 2k^2M^2o(1) \quad (2.97)$$

Hence, we have shown stochastic equicontinuity. □

Chapter 3

Average Partial Effects in Nonseparable Panel Data Models: Identification and Testing

3.1 Introduction

In many empirical settings, the researcher's objective is to identify the *ceteris paribus* effect of a particular variable on an outcome of interest; for instance, the effect of a job training program on participants' employment or income, the effect of having children on female labor participation, and the effect of schooling on income. While economic theory gives us some direction about which variables to include in order to identify the *ceteris paribus* effect, the functional form governing the relationship between the variables is often left unspecified. This is where nonparametric identification becomes particularly useful, even when parametric models are used as an approximation to the unknown relationship for estimation purposes.

The paper at hand examines the identification of the average partial effect (APE) of a discrete regressor in a nonseparable panel data model, where the time dimension, T, is fixed. This reflects the situation in many microeconomic panel data settings, where we observe a large number of individuals over a short time horizon. In this setup, the APE is point-identified only for a subpopulation.[1]

This paper starts with an unrestricted data-generating process (DGP), where the outcome variable for individual i in period t, Y_{it}, is given by the following,

$$Y_{it} = \xi_t(X_{it}, \mathcal{A}_i, \mathcal{U}_{it}), \quad i = 1, 2, ..., n, t = 1, 2, ..., T \tag{3.1}$$

where X_{it} is a vector of regressors, $\mathcal{A}_i$ includes time-invariant, individual-specific variables that are unobservable,[2] and $\mathcal{U}_{it}$ are idiosyncratic shocks. This paper is

[1] For the purposes of this paper, we will use use identification and point-identification interchangeably. If *a priori* bounds exist for the outcome variable, then bounds on the APE can be constructed as in Chernozhukov et al. (2013). However, for the purposes of this paper, I am concerned with point-identification only.

[2] In linear models, $\mathcal{A}_i$ is referred to as an individual fixed effect. I refrain from this terminology,

concerned with static models. Hence, X_{it} does not include lags of the outcome variable as well as other regressors that may introduce feedback mechanisms from Y_{it} to X_{is} for $s \geq t$.[3] The structural function, $\xi_t(.)$, is assumed to be unknown and is allowed to vary over time.[4] Furthermore, observables and unobservables are allowed to interact in arbitrary ways in the structural function. Without further assumptions, the DGP reflects arbitrary individual and time heterogeneity, which I formally define in Section 3.3.

Restrictions on this general DGP achieve identification of the APE for a subpopulation.[5] There are three goals behind starting with a general DGP: (1) to formally characterize the trade-off between identifying assumptions that restrict the structural function, individual and/or time heterogeneity, (2) to present a menu of identifying assumptions, (3) to propose tests thereof.

From a theoretical viewpoint, the characterization of the trade-off contributes to the understanding of nonparametric identification of APEs in nonseparable panel data models. The menu of identifying assumptions includes existing methods in the literature as a special case. It also suggests new methods that fall under both the fixed-effects and correlated-random-effects categories, which are referred to in this paper as within-group and within-period identification, respectively. Finally, the testing problem addresses new issues for testing the equality of distributions in the two-sample problem, where samples are dependent and the data is possibly demeaned. Bootstrap-adjusted Kolmogorov-Smirnov (KS) and Cramer-von-Mises (CM) statistics are proposed and shown to be asymptotically

since it misrepresents what $\mathcal{A}_i$ stands for.

[3]This rules out dynamic selection.

[4]It is structural in the sense that $Y_{it}^x = \xi_t(x, \mathcal{A}_i, \mathcal{U}_{it})$, where Y_{it}^x is the outcome variable when X_{it} is fixed to x.

[5]The intuition here is similar to the distinction between average treatment effect (ATE) and Treatment on the Treated (TOT). For fixed T, we only observe a subpopulation with and without the treatment. In the presence of unobservable heterogeneity, we can only point-identify the APE for a subpopulation and cannot point-identify the APE for the entire population.

valid.

For practical purposes, the menu of identifying assumptions provides the empirical researcher with a set of alternative identification strategies to choose from. Since the identification strategies may not be justified *a priori*, the tests are tools to aid the empirical researcher in justifying the choice of a particular identification strategy. The tests proposed here include a nonparametric test of the fixed-effects assumption in the presence of time effects.

Angrist and Newey (1991) propose an over-identification test of the fixed-effects assumption in the linear model and find evidence against it for the human capital earningsc function using a subsample of the national longitudinal survey of youth (NLSY). We revisit the same dataset in the empirical illustration. When we apply our nonparametric test, we do not find evidence against the fixed-effects assumption.[6] We also estimate the APE. Our results indicate that implications of the linear model, such as constant APEs across time, are violated. We conclude that testing nonparametric identification may be useful to the empirical researcher, even when parametric models are used for estimation purposes to approximate the unknown structural relationship.

Related Literature. Athey and Imbens (2006),Chernozhukov et al. (2013) hereinafter, CFHN2013, and Hoderlein and White (2012) impose restrictions on time heterogeneity to achieve identification. We will refer to this category of restrictions as time homogeneity, but it is also referred to as time invariance and time stationarity in the literature.[7] The papers mentioned above are categorized as 'fixed effects', since they do not restrict individual heterogeneity. As CFHN2013

[6]The test in Angrist and Newey (1991) uses restrictions implied by the linear model in addition to restrictions implied by the fixed-effects assumptions. Hence, the rejection of the test may be either due to violations of the linear model or the fixed-effects assumption.

[7]CFHN2013 show that in the presence of location and scale time effects, one can still identify the effect of a discrete regressor using time homogeneity.

point out, time homogeneity is a very strong assumption, and its empirical content may be thought of as time being "randomly assigned." Bester and Hansen (2009) propose a correlated random effects approach where they impose restrictions on individual heterogeneity, but allow for time heterogeneity. Altonji and Matzkin (2005) discuss the issue of identifying average effects for cross-sectional and panel data models. Although they do not explicitly allow for time heterogeneity, they propose exchangeability restrictions that are similar in spirit to Bester and Hansen (2009)'s approach.

Similar to the paper at hand, the approach of the aforementioned papers is focused on the identification of a particular object of interest, the APE of a discrete or continuous regressor. The approach in this strand of the literature extends the intuition of differencing out individual heterogeneity in linear models to nonseparable models, which was termed by Magnac (2004) and Hoderlein and White (2012) as quasi-differencing. Hence, this literature is quite relevant for situations where the linear model is used to approximate an unknown, possibly nonseparable model. In the absence of separability, quasi-differencing here is simply averaging over unobservables in an "appropriate way."

The quasi-differencing approach originated in parametric binary choice models, such as conditional logit (Chamberlain, 1984, 2010), where the presence of a sufficient statistic for the individual effect allowed for the identification of the common parameter while treating individual heterogeneity nonparametrically. Magnac (2004) introduces the concept of quasi-differencing as the presence of a sufficient statistic for the individual effect. The term quasi-differencing has been more generally used to refer to identification strategies that extend the intuition of differencing in the linear model to more general settings, where the distribution of individual heterogeneity is left unrestricted. For semiparamteric binary choice models, Hon-

ore and Kyriazydou (2000) use the intuition of quasi-differencing to nonparametrically identify the common parameter. Hoderlein and White (2012) refer to their approach as quasi-differencing. For random coefficient models, Graham and Powell (2012) also use a differencing approach to identify the APE.

This paper further generalizes the concept of quasi-differencing to refer more generally to any approach where the APE is identified using average changes of the outcome variable across time or subpopulations that coincide with a change in the variable of interest.[8] This definition allows us to include works such as Altonji and Matzkin (2005) and Bester and Hansen (2009) under this category.

The key intuition behind the quasi-differencing approach is that if we seek to identify the APE of a regressor from average changes of the outcome variable across time, then we have to assume that the distribution of unobservables does not change over time. However, if we would like to use average changes of the outcome variable across subpopulations, then we have to assume that the distribution of unobservables is the same across subpopulations. Hence, some form of homogeneity assumption is required either across time periods or subpopulations. Time homogeneity assumptions are widely used in the literature as noted above. Homogeneity assumptions across subpopulations are less prevalent in the nonparametric identification literature. For continuous regressors, Bester and Hansen (2009) proposes assumptions that are closest to this in spirit.[9]

Another strand in the literature follows the classical identification approach, which seeks to identify all structural objects, i.e. the structural function and the conditional distribution of unobservables, which include Altonji and Matzkin

[8]It is important to note here that the generalization of quasi-differencing here is different from the generalization proposed in Evdokimov (2011), which is used to identify all structural objects. However, both generalizations allow us to include situations where the structural function is not assumed to be stationary across time.

[9]For cross-sectional setups, Angrist (2004) uses this type of homogeneity assumptions to relate local average treatment effect (LATE) to the average treatment effect (ATE).

(2005), Evdokimov (2010) and Evdokimov (2011). The key differences between this category and the quasi-differencing approach is that the former identifies all objects, whereas the latter focuses on a specific object and hence requires weaker conditions. In other words, the quasi-differencing literature answers the question: What object can one identify under minimal assumptions? The classical identification literature on the other hand answers a different question: What assumptions are sufficient to identify all structural objects? Both contribute immensely to our understanding of the possibilities and limits of identification of various objects of interest from panel data without parametric assumptions. The paper at hand falls under the quasi-differencing category. Its findings are related to the classical identification literature and attempts to point to the relative merits of both approaches.

3.2 Basic Identification Problem: Job Training Example

Now let our variable of interest, X, be a binary variable for the participation in a job training program and our outcome of interest, Y, be earnings. Our object of interest is the effect of participating in a job training program on earnings, and we observe individuals in two time periods, $t = 1, 2$. We assume that there are three subpopulations: (1) individuals that never participate in the job training program, $X_i = (0, 0)$, (2) individuals that participate in the first time period, but not in the second, $X_i = (1, 0)$, (3) individuals that participate in the second time period, but not in the first, $X_i = (0, 1)$.

Now there are several unobservable factors here that may confound our effect of interest if we examine average changes across time or subpopulations

that coincide with changes in job training status: (1) macroeconomic conditions, which may change over time, (2) time-invariant individual characteristics, such as innate ability, which makes different subpopulations incomparable, (3) time-varying unobservable individual characteristics, such as the development of new skills, which may change the probability of participation across time. If all of these confounding factors are present, then the effect of job training on earnings cannot be identified for any subpopulation. For instance, the average change of earnings across time for subpopulation $X_i = (0, 1)$ is partly due to the change in job training status and partly due to changing macroeconomic conditions and time-varying ability.

Now if one assumes that macroeconomic conditions are the same across both time periods and ability was only time-invariant, then one can identify the APE of job training for the subpopulations that participate in the program using average within-group changes across time, specifically, $E[Y_{i2} - Y_{i1}|X_i = (0, 1)]$ and $E[Y_{i1} - Y_{i2}|X_i = (1, 0)]$.

However, if macroeconomic conditions were changing over time and affect individuals' decision to participate in the job training program, within-group changes over time can no longer be interpreted as the APE of the job training program. The change over time will be confounded by the changing macroeconomic climate and are only partly due to changes in job training status. In this setup, one could examine within-period changes across groups. If one assumes that subpopulations $X_i = (0, 1)$ and $X_i = (1, 0)$ have the same unobservable tendencies, e.g. they have the same proportion of high-types to low-types, then comparing individuals with $X_i = (0, 1)$ and $X_i = (1, 0)$ within the same period would not confound our effect of interest. In this case, $\{E[Y_{i1}|X_i = (1, 0)] - E[Y_{i1}|X_i = (0, 1)]\}$ and $\{E[Y_{i2}|X_i = (0, 1)] - E[Y_{i2}|X_i = (1, 0)]\}$ identify the effect of job training for the

two subpopulations in the first and second period, respectively. Note that in the presence of time-varying unobservables, the APE is different in each period.

Finally, another possible strategy for within-period identification is assuming that job training status in the second period is random conditional on the first. This implies selection on observables, as introduced in Heckman and Robb (1985) in the second time period conditional on job training status in the first time period. In this case, one can identify the effect of job training on earnings from $\{E[Y_{i2}|X_i = (0,1)] - E[Y_{i2}|X_i = (0,0)]\}$.

One of the key points of the paper at hand is that neither of the above identification strategies is justified *a priori*. However, they all have clear testable implications. Hence, this paper proposes strategies to test for these identifying assumptions. We first proceed to a formal discussion of the identification problem.

3.3 Nonparametric Identification of APEs

3.3.1 DGP and Definitions

We begin with a general DGP exhibiting arbitrary individual and time heterogeneity, formally defined below. Let Y_{it} be the outcome variable of interest with support $\mathcal{Y} \subseteq \mathbb{R}$. For $i = 1, 2, ..., n$, $t = 1, 2$,

$$Y_{it} = \xi_t(X_{it}, \mathcal{A}_i, \mathcal{U}_{it}) \tag{3.2}$$

where X_{it} is a $d \times 1$ vector of discrete regressors, $\mathcal{A}_i$ denotes unobservable individual-specific, time-invariant factors, and $\mathcal{U}_{it}$ denotes individual-specific, time-varying unobservables. $\{Y_{it}, X_{it}\}$ are observables, whereas $\{\mathcal{A}_i, \mathcal{U}_{it}\}$ are unobservables that

may confound our effect of interest. For simplicity, we assume that $\mathcal{A}_i$ and $\mathcal{U}_{it}$ are scalar real-valued random variables for $t = 1, 2$. However, the results would hold for any finite-dimensional vectors $\mathcal{A}_i$ and $\mathcal{U}_{it}$.[10] Calligraphic letters are used to distinguish unobservable from observable factors. ξ_t is unknown and is referred to as the structural function in the sense that $Y_{it}(x) = \xi_t(x, \mathcal{A}_i, \mathcal{U}_{it})$.

Notation. Let $\mathcal{X}$ denote the support of X_{it}. x and x' denote elements in $\mathcal{X}$. Now $X_i \equiv (X_{i1}, X_{i2}, ..., X_{iT})$, a $d \times T$ matrix with support $\mathcal{X}^T \equiv \times_{t=1}^{T} \mathcal{X}$. $\underline{x}$ and $\underline{x}'$ denote elements in $\mathcal{X}^T$. Note that $\underline{x} = (\underline{x}_1, \underline{x}_2, ..., \underline{x}_T)$, hence $\underline{x}_t$ denotes the t^{th} column of $\underline{x}$. For random variables, W_{it} and Z_i, let $F_{W_{it}|Z_i}(.|.)$ denote the conditional distribution function of W_{it} given Z_i in period t. For a set S, $|S|$ of a set denotes the cardinality of S.

Assumption 3.3.1 *(General DGP)*

(i) $Y_{it} = \xi_t(X_{it}, \mathcal{A}_i, \mathcal{U}_{it})$, *where* $\xi_t : \mathcal{X} \times \mathbb{R}^2 \mapsto \mathcal{Y}$, *where* $\mathcal{Y} \subseteq \mathbb{R}$,

(ii) $\mathcal{X}$ *is finite,* $|\mathcal{X}| = K$,

(iii) $E[Y_{it}] < \infty$ *for all* $t = 1, 2, .., T$,

(iv) $P(X_i = \underline{x}) > 0$ *for all* $\underline{x} \in \mathcal{X}^T$.

The main content of the above assumption is the finite support of X_{it} in (ii).[11] Since $\mathcal{X}$ is finite and T is fixed, $\mathcal{X}^T$ is also a finite set. Assumption 3.3.1 (i) may be thought of as a 'correct specification' assumption. However, it is important to note that the choice of variables to include in X_{it} is so far not restrictive, since the assumption allows for an arbitrary, time-varying structural relationship and there

[10] The support of $\mathcal{A}_i$ and $\mathcal{U}_{it}$ could also be different from $\mathbb{R}$.

[11] It is important to note that the identification component of this paper applies to a discrete change of continuous variables as well.

are no assumptions on the distribution of unobservables.[12] It is also important to note that $\mathcal{A}_i$ and $\mathcal{U}_{it}$ may be any finite-dimensional vector, we assume that they are scalar real-valued random variables for simplicity. Assumption 3.3.1 (iii) and (iv) are regularity conditions that ensure that the APE exists for all elements in the support of X_i, which simplifies our analysis.

The general DGP does not impose any further restrictions on the structural functions ξ_t or the conditional distributions of unobservables $F_{\mathcal{A}_i,\mathcal{U}_{it}|X_i} = F_{\mathcal{U}_{it}|X_i,\mathcal{A}_i}F_{\mathcal{A}_i|X_i}$.[13] Now we formally define what we mean by a nonstationary structural function, arbitrary individual and time heterogeneity.

Definition 3.3.1 *(Nonstationary Structural Function)* $\xi_t(.)$ *may vary for* $t = 1, 2, ..., T$.

Definition 3.3.2 *(Arbitrary Individual Heterogeneity) A DGP exhibits arbitrary individual heterogeneity, if* $F_{\mathcal{A}_i|X_i}(.|.)$ *is unrestricted.*

The above is the fundamental definition of a 'fixed effect'. Arellano (2003) shows how the linear fixed effects estimator is equivalent to the maximum likelihood estimator where the distribution of individual heterogeneity is unrestricted. Fixed effects logit and Poisson are examples of nonlinear parametric models where the presence of a sufficient statistic for individual heterogeneity allows one to identify parameters of interest without assumptions on the distribution of individual heterogeneity.

Definition 3.3.3 *(Arbitrary Time Heterogeneity) A DGP exhibits arbitrary time heterogeneity, when* $F_{\mathcal{U}_{it}|X_i,\mathcal{A}_i}(.|.,.)$ *may vary for* $t = 1, 2, ..., T$.

[12] Once we impose identifying assumptions, the choice of X_{it} will be very important.

[13] Depending on the nature of Y_{it}, *a priori* bounds may be given as assumed in CFHN2013. It is important to note however that, unlike the case of set-identification, *a priori* bounds are irrelevant for the discussion of point-identification.

The key content in the above definition is that the distribution of time-varying unobservables may change over time. However, the definition does not impose that $F_{\mathcal{U}_{it}|X_i,\mathcal{A}_i}$ is unrestricted. This is a key distinction between the definition of arbitrary individual and time heterogeneity.

We will also use the terminology of movers and stayers introduced in Chamberlain (1982) to refer to subpopulations that change the regressor vector, X_{it}, over time and those who do not, respectively.[14] The following definition uses the realizations of X_i to define a subpopulation.

Definition 3.3.4 *(Subpopulation) A subpopulation is defined by its realization of* $X_i = \underline{x}$*, where* $\underline{x} \in \mathcal{X}^T$*.*

Since $|\mathcal{X}^T|$ is finite, we have finitely many subpopulations. It is important to note that each subpopulation, $\underline{x} \in \mathcal{X}^T$, is characterized by its distribution of unobservables, i.e. $F_{\mathcal{A}_i,\mathcal{U}_{it}|X_i}(.|\underline{x})$ for $t = 1, 2, ..., T$. Together with the structural function, this yields a distribution of the outcome variable for each subpopulation. Hence, we can think of the subpopulation as infinitely many realizations from the same distribution. This is not to be confused with individuals in the same subpopulation having the same 'fixed effect'.[15] Back to our job training example, individuals in a subpopulation have the same distribution of ability, this does not mean that they all have the same ability or that they all respond in the same way to macroeconomic shocks. But rather that they have the same likelihood of being high-achieving or fast-learners.

[14] For the latter subpopulations, all columns of X_i are equal, i.e. $X_{i1} = ... = X_{iT}$ as defined in Chamberlain (1982).

[15] The case here is really one where correlated random effects and fixed effects are equivalent. Correlated random effects are approaches that integrate over the distribution of individual effects, whereas fixed effects treat them as fixed in repeated sampling. In the linear model, fixed effects and correlated random effects are equivalent under certain assumptions, as shown in Mundlak (1978).

3.3.2 Object of Interest and the Quasi-differencing Approach

Now we would like to learn about the average effect of a discrete regressor X on Y for a particular subpopulation $\underline{x}$. The problem is that there is both individual and time heterogeneity that may confound our effect if we simply look at the change in Y as X changes over time. Formally, our object of interest is the APE of changing X_{it} from x to x', $x \neq x'$, for subpopulation $X_i = \underline{x}$, given as follows

$$\begin{aligned} & \beta_t(x \to x' | X_i = \underline{x}) \\ = \; & E[Y_{it}^{x'} | X_i = \underline{x}] - E[Y_{it}^{x} | X_i = \underline{x}], \end{aligned} \tag{3.3}$$

where $Y_{it}^x = \xi_t(x, a, u)$ is the counterfactual notation, i.e. Y_{it}^x is the outcome variable if the regressor vector was fixed at x. We distinguish between x and x' and the columns of $\underline{x}$, since they may not be equal. For instance, if we are interested in the effect of obtaining a master degree (x') on top of a bachelor degree (x), then for individuals that only have a high school degree, i.e. $\underline{x} = (12, ..., 12)$, $\underline{x}_t \neq x$ and $\underline{x}_t \neq x'$ for all $t = 1, 2, ..., T$.[16]

Note that the APE is indexed by the time period t and the subpopulation $\underline{x}$. This reflects the presence of time heterogeneity (*the effect is different across time periods*) as well as individual heterogeneity (*the effect is different across subpopulations*).

[16]Note that in this situation, identification of the APE for these subpopulations requires assumptions on individual heterogeneity in the spirit of Angrist (2004).

Given our DGP in Assumption 3.3.1, we can write the APE as follows

$$\begin{aligned} & E[Y_{it}^{x'}|X_i = \underline{x}] - E[Y_{it}^{x}|X_i = \underline{x}] \\ = & \int \xi_t(x', a, u) dF_{\mathcal{A}_i, \mathcal{U}_{it}|X_i}(a, u|\underline{x}) - \int \xi_t(x, a, u) dF_{\mathcal{A}_i, \mathcal{U}_{it}|X_i}(a, u|\underline{x}) \end{aligned} \quad (3.4)$$

It is important to note that in (3.4) the only difference between the two terms is that one is evaluated at x and the other at x'. Otherwise, the structural function ξ_t and the unobservable distribution $F_{\mathcal{A}_i, \mathcal{U}_{it}|X_i}(.,.|\underline{x})$ are the same. This is the key difficulty in identifying a *ceteris paribus* effect in panel data. In an experiment, one could randomize individuals from the same subpopulation in the same period, thereby holding individual and time heterogeneity fixed. In observational settings, assumptions on $\{\xi_t,\, F_{\mathcal{A}_i, \mathcal{U}_{it}|X_i}\}_{t=1}^{T}$ are required to ensure that average within-group or within-period changes are taken with respect to the same structural function and unobservable distribution. This is the key intuition behind quasi-differencing in general nonseparable panel data models that will be examined in the following.

3.3.3 Characterization of the Trade-off between Identifying Assumptions

To simplify notation, we assume $T = 2$. More specifically, we would like to identify the APE of a subpopulation (x, x')

$$\begin{aligned} & \beta_t(x \to x'|X_i = (x, x')) \\ = & \int (\xi_t(x', a, u) - \xi_t(x, a, u)) dF_{\mathcal{A}_i, \mathcal{U}_{it}|X_i}(a, u|(x, x')) \end{aligned} \quad (3.5)$$

In observational settings, an obvious candidate for the identification of $\beta_t(x \to x'|X_i = (x, x'))$ is $E[Y_{i2} - Y_{i1}|X_i = (x, x')]$, which we can decompose as follows

$$\begin{aligned} & E[Y_{i2} - Y_{i1}|X_i = (x, x')] \\ = \; & E[Y_{i2} - Y_{i2}^x|X_i = (x, x')] + E[Y_{i2}^x - Y_{i1}|X_i = (x, x')] \\ = \; & \beta_2(x \to x'|X_i = (x, x')) + E[Y_{i2}^x - Y_{i1}|X_i = (x, x')]. \end{aligned}$$

The identification of $E[Y_{i2}^x - Y_{i1}|X_i = (x, x')]$ is necessary and sufficient for the identification of $\beta_2(x \to x'|X_i = (x, x'))$. We will refer to the former as the counterfactual trend, i.e. it is the change that would have occurred to the subpopulation (x, x') had the change from x to x' not occurred.

In the following, we will give a sufficient condition under which we can identify the counterfactual trend, $E[Y_{i2}^x - Y_{i1}|X_i = (x, x')]$, from a stayer subpopulation, (x, x), i.e.

$$\begin{aligned} & E[Y_{i2}^x - Y_{i1}|X_i = (x, x')] \\ = \; & E[Y_{i2} - Y_{i1}|X_i = (x, x)]. \end{aligned} \tag{3.6}$$

This condition will help us characterize the trade-off between various identifying assumptions. We first introduce some standard regularity conditions on the distribution of unobservables.

Assumption 3.3.2 *(Distribution of Unobservables)*

(i) $F_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(.,.,.|.)$ *admits a density,* $f_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(.,.,.|.)$,

(ii) $f_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(.,.,.|.) > 0$, $f_{\mathcal{A}_i,\mathcal{U}_{it}|X_i}(.,.|.) > 0$ *for* $t = 1, 2$, $f_{\mathcal{A}_i|X_i}(.|.) > 0$.

The following theorem gives a sufficient condition for (3.6).

Theorem 3.3.1 *Let Assumptions 3.3.1 and 3.3.2 hold,*

$$E[Y_{i2}^{x} - Y_{i1}|X_i = (x, x')] \quad = \quad E[Y_{i2} - Y_{i1}|X_i = (x, x)]$$

if

$$\begin{aligned} & (\xi_2(x, a, u_2) - \xi_1(x, a, u_1)) \\ \times \quad & (f_{\mathcal{A}_i, \mathcal{U}_{i1}, \mathcal{U}_{i2}|X_i}(a, u_1, u_2|(x, x')) - f_{\mathcal{A}_i, \mathcal{U}_{i1}, \mathcal{U}_{i2}|X_i}(a, u_1, u_2|(x, x))) = 0. \\ & \forall (a, u_1, u_2) \in \mathbb{R}^3 \end{aligned}$$

All proofs of Section 3.3 are given in Appendix 3.6.1. The above condition depends on the change in the structural function and the difference between unobservable distributions, both across time and subpopulations.

We illustrate how the condition characterizes the trade-off between different identifying assumptions in the following variations on the above theorem. The first two variations reflect a situation where the researcher would like to leave individual heterogeneity unrestricted, which reflects the standard fixed-effects approach.

Variation 3.3.1 *(Stationary Structural Function up to Generalized Time Effect) Let Assumptions 3.3.1 and 3.3.2 hold.*
Under arbitrary individual heterogeneity and $\mathcal{U}_{i1}|X_i, \mathcal{A}_i \stackrel{d}{=} \mathcal{U}_{i2}|X_i, \mathcal{A}_i$*,*

$$E[Y_{i2}^{x} - Y_{i1}|X_i = (x, x')] \quad = \quad E[Y_{i2} - Y_{i1}|X_i = (x, x)]$$

if

$$\xi_t(x, a, u) \quad = \quad \xi(x, a, u) + \lambda_t(x) \, \forall (a, u) \in \mathbb{R}^2$$

In the above variation on Theorem 3.3.1, time homogeneity, i.e. $\mathcal{U}_{i1}|X_i, \mathcal{A}_i \overset{d}{=} \mathcal{U}_{i2}|X_i, \mathcal{A}_i$, is maintained. If individual heterogeneity is left unrestricted, then the structural function has to be decomposed into a stationary and nonstationary component, $\xi(x, a, u)$ and $\lambda_t(x)$, respectively. The nonstationary component may only depend on observable regressors. We will refer to this as stationary in unobservables. Hence, in the quasi-differencing approach, time homogeneity alone is not sufficient. Hoderlein and White (2012) and CFHN2013 assume both time homogeneity and a stationary structural function without time effects to achieve identification of the APE for discrete and continuous regressors, respectively.[17] Note that for continuous regressors, conditional independence restrictions are required to ensure that the marginal change in the regressor vector does not change the distribution of the unobservables, as in Altonji and Matzkin (2005) and Hoderlein and White (2012).

Variation 3.3.2 *(Time Homogeneity) Let Assumptions 3.3.1 and 3.3.2 hold. Under arbitrary individual heterogeneity and* $\xi_t(x, a, u) = \xi(x, a, u) + \lambda_t(x)$,

$$E[Y_{i2}^x - Y_{i1}|X_i = (x, x')] \quad = \quad E[Y_{i2} - Y_{i1}|X_i = (x, x)]$$

if $\forall \underline{x} \in \{(x, x), (x, x')\}$

$$f_{\mathcal{U}_{i2}|X_i,\mathcal{A}_i}(u_2|\underline{x}, a) \quad = \quad f_{\mathcal{U}_{i1}|X_i,\mathcal{A}_i}(u_1|\underline{x}, a) \;\; \forall (a, u_1, u_2) \in \mathbb{R}^2$$

In Variation 3.3.2, the structural function is assumed to be stationary up to a generalized time effect. In this case, time homogeneity has to be assumed. Hence, the last two variations show that the stationarity of the structural function

[17] With an additional restriction, CFHN2013 can allow for nonstochastic location and scale time effects, i.e. $Y_{it} = \xi(X_{it}, \mathcal{A}_i, \mathcal{U}_{it})\sigma_t + \lambda_t$.

and time homogeneity are both required to identify the APE using within-group changes across time. This is not at all surprising for the quasi-differencing approach, since we simply take average within-group changes and remain agnostic about the functional form of the structural function and the distribution of unobservables. Thus, without information about these objects, we cannot hope to use time homogeneity without the stationarity of the structural function in unobservables and vice versa.

Variations 3.3.1 and 3.3.2 characterize a situation where identification would be achieved within-group, since individual heterogeneity is unrestricted. Section 3.3.4 generalizes the setup here where subpopulations other than (x, x) can be used to identify the generalized time effect.

So far we have seen that restrictions on time heterogeneity and the stationarity of the structural function are required if we would like individual heterogeneity to be unrestricted. In this setup, we use average within-group changes across time to identify the APE for a subpopulation. Now we move to the case where we would like to have unrestricted time heterogeneity. Here, we will see that we have to impose restrictions on individual heterogeneity.

Variation 3.3.3 *(Individual Homogeneity) Given Assumptions 3.3.1, 3.3.2. Under arbitrary time heterogeneity*

$$E[Y_{i2}^{x} - Y_{i1} | X_i = (x, x')] \quad = \quad E[Y_{i2} - Y_{i1} | X_i = (x, x)]$$

if

$$f_{\mathcal{A}_i, \mathcal{U}_{i1}, \mathcal{U}_{i2}|X_i}(a, u_1, u_2|(x, x')) \quad = \quad f_{\mathcal{A}_i, \mathcal{U}_{i1}, \mathcal{U}_{i2}|X_i}(a, u_1, u_2|(x, x)) \; \forall (a, u_1, u_2) \in \mathbb{R}^3$$

The above condition implies that the subpopulations (x, x') and (x, x) are homogeneous, i.e. they have the same distribution of unobservables.[18] This allows us to identify the APE using average within-period changes across subpopulations. Specifically,

$$\begin{aligned}
& \beta_2(x \to x' | X_i = (x, x')) \\
= & \int (\xi_2(x', a, u) - \xi_2(x, a, u)) dF_{\mathcal{A}_i, \mathcal{U}_{i2}|X_i}(a, u|(x, x')) \\
= & \int \xi_2(x', a, u) dF_{\mathcal{A}_i, \mathcal{U}_{i2}|X_i}(a, u|(x, x')) - \int \xi_2(x, a, u) dF_{\mathcal{A}_i, \mathcal{U}_{i2}|X_i}(a, u|(x, x)) \\
= & E[Y_{i2}|X_i = (x, x')] - E[Y_{i2}|X_i = (x, x)],
\end{aligned}$$

where the penultimate equality follows from the individual homogeneity assumption. Note that in this setup, we are using the information obtained from the first time period, i.e. X_{i1}, to identify the APE in the second time period, which is a cross-section of all individuals.[19] Section 3.3.4 generalizes this type of homogeneity assumptions to allow subpopulations other than (x, x) as a control group for (x, x'). Angrist (2004) discusses the use of this type of homogeneity assumptions to relate the local average treatment effect (LATE) to the average treatment effect (ATE) for instrumental variables methods in the cross-sectional setup.

So far we have characterized the obvious trade-off between individual and time heterogeneity. Since we only have two dimensions that we are differencing over, we cannot leave both unrestricted. In the next variation on Theorem 3.3.1, the setup assumes that we would like to leave both individual and time heterogeneity unrestricted. The result shows that separability and conditional independence

[18] It is important to note here that in the above case, the result holds because $E[Y_{i2}^x|X_i = (x, x')] = E[Y_{i2}|X_i = (x, x)]$ and $E[Y_{i1}|X_i = (x, x')] = E[Y_{i1}|X_i = (x, x)]$.

[19] Recall that microeconomic panel data models are also called cross-sectional time series. For each time period, we observe a cross-section.

assumptions are required to aid identification. We first introduce some more regularity conditions.

Assumption 3.3.3 *(Smoothness and Dominating Function)*

(i) $\partial\xi(x,a,u)/\partial a$, $\partial^2\xi(x,a,u)/\partial a\partial u_t$, $f_{\mathcal{U}_{it}|X_i,\mathcal{A}_i}(u|\underline{x},a)$, $\partial f_{\mathcal{U}_{it}|X_i,\mathcal{A}_i}(u|\underline{x},a)/\partial a$ *exist and are continuous for all* a,u, $x\in\mathcal{X}$, $\underline{x}\in\mathcal{X}^T$ *and* $t=1,2,...,T$,

(ii) $|\xi(x,a,u_t)f_{\mathcal{U}_{it}|X_i,\mathcal{A}_i}(u|\underline{x},a)|\leq g(x,u,\underline{x})$, *where* $E[|g(x,\mathcal{U}_{it},X_i)||X_i=\underline{x}]<\infty$ $\forall\underline{x}\in\mathcal{X}^T$.

The separability of a function in two variables is characterized by the cross-partial derivative being zero.[20] Hence, Assumption 3.3.3 (i) ensures that sufficient smoothness conditions hold. Assumption 3.3.3 (ii) ensures that we can apply the dominating convergence theorem.

Variation 3.3.4 *(Separability and Independence) Let Assumptions 3.3.1, 3.3.2 and 3.3.3 hold.*
Under unrestricted individual and time heterogeneity, and $\xi_t(x,a,u)=\xi(x,a,u)$, $t=1,2$,

$$E[Y_{i2}^x-Y_{i1}|X_i=(x,x')] \;=\; E[Y_{i2}-Y_{i1}|X_i=(x,x)]$$

if

$$\partial^2\xi(x,a,u)/\partial a\partial u \;=\; 0\ \forall(a,u)\in\mathbb{R}^2, \qquad \textit{(Separability)}$$

$$(\mathcal{U}_{i1},\mathcal{U}_{i2}) \;\perp\; (X_i,\mathcal{A}_i). \qquad \textit{(Independence)}$$

[20] For instance, $m(a,u)=a^2+u^2$, then $\partial m(a,u)/\partial a=2a$ and $\partial m(a,u)/\partial u=2u$. Thus, $\partial^2 m(a,u)/\partial a\partial u=0$.

There are two interesting findings in the above result. The first is that separability aids identification in this setup if it is coupled with independence. The separability of $\mathcal{A}_i$ and $\mathcal{U}_{it}$ alone is not sufficient if they are dependent. We can think of independence as "stochastic separability".[21] Hence, both functional-form and "stochastic separability" are required. The other interesting issue to note here is that even though we have arbitrary individual and time heterogeneity, i.e. $F_{\mathcal{A}_i|X_i}$ is unrestricted, and $F_{\mathcal{U}_{it}|X_i,\mathcal{A}_i}$ is allowed to change over time, the idiosyncratic shocks, $\mathcal{U}_{it}$ are independent of individual heterogeneity and observables, since $F_{\mathcal{U}_{it}|X_i,\mathcal{A}_i} = F_{\mathcal{U}_{it}}$.

Relation to the Classical Identification Literature. In the classical identification literature, stronger assumptions are imposed to identify all structural objects, whereas the quasi-differencing is much more parsimonious. In the following, we will discuss the relative merits of both approaches with special attention given to the model in Evdokimov (2010). It is given by the following

$$Y_{it} \quad = \quad m(X_{it}, \mathcal{A}_i) + \mathcal{U}_{it}. \tag{3.7}$$

Under the separability of $\mathcal{U}_{it}$ in the structural function, the conditional independence $F_{\mathcal{U}_{it}|X_i,\mathcal{A}_i,\mathcal{U}_{i(t)}} = F_{\mathcal{U}_{it}|X_{it}}$,[22] monotonicity of $m(x,a)$ in a and regularity conditions, Evdokimov (2010) shows the identification of $F_{\mathcal{U}_{it}|X_{it}}$ for $t = 1, 2$, $F_{\mathcal{A}_i|X_i}$ and $m(x,a)$, $\forall x, a$. Under the above model, we can decompose the average within-

[21] This term is used here to convey the intuition and is not related to the definition of separability of stochastic processes.

[22] $\mathcal{U}_{i(t)}$ is the idiosyncratic shock for the other period $\tau \neq t$.

group change for subpopulation (x, x') as follows,

$$\begin{aligned} E[Y_{i2} - Y_{i1}|X_i = (x, x')] &= E[m(x', \mathcal{A}_i) - m(x, \mathcal{A}_i)|X_i = (x, x')] \\ &+ E[\mathcal{U}_{i2}|X_{i2} = x'] - E[\mathcal{U}_{i1}|X_{i1} = x] \\ &= \beta(x \to x'|X_i = (x, x')) + \Delta. \end{aligned} \tag{3.8}$$

Since $F_{\mathcal{U}_{it}|X_{it}}$ is identified for $t = 1, 2$, we can identify Δ. Hence, the above identification strategy allows us to identify our object of interest in the presence of time heterogeneity, while allowing individual heterogeneity to be unrestricted.

The quasi-differencing approach gives two strategies to identify the same objects. First, Variation 3.3.1 and 3.3.2 show that neither time homogeneity nor stationarity of the structural function are sufficient on their own. Hence, the following model

$$\begin{aligned} Y_{it} &= \xi(X_{it}, \mathcal{A}_i, \mathcal{U}_{it}) + \lambda_t(X_{it}), \ \ t = 1, 2 \\ \mathcal{U}_{i1}|X_i, \mathcal{A}_i &\overset{d}{=} \mathcal{U}_{i2}|X_i, \mathcal{A}_i, \end{aligned} \tag{3.9}$$

where $\lambda_1(x) = 0$ for all $x \in \mathcal{X}$. Hence, $\xi_1(x, a, u) = \xi(x, a, u)$ and $\xi_2(x, a, u) = \xi(x, a, u) + \lambda_2(x)$. Now the APE of moving from x to x' for subpopulation (x, x') in the period t is given by

$$\begin{aligned} & \beta_t(x \to x'|X_i = (x, x')) \\ = & \int (\xi_t(x', a, u) - \xi_t(x, a, u)) dF_{\mathcal{A}_i, \mathcal{U}_{it}|X_i}(a, u|(x, x')) \\ \overset{(3.9)}{=} & \int (\xi_t(x', a, u) - \xi_t(x, a, u)) dF_{\mathcal{A}_i, \mathcal{U}_{i1}|X_i}(a, u|(x, x')) \end{aligned} \tag{3.10}$$

Note that using average within-group changes

$$\begin{aligned} & E[Y_{i2} - Y_{i1}|X_i = (x, x')] \\ = \; & E[Y_{i2} - Y^x_{i2}|X_i = (x, x')] + E[Y^x_{i2} - Y_{i1}|X_i = (x, x')] \\ = \; & \int (\xi_2(x', a, u) - \xi_2(x, a, u)) dF_{\mathcal{A}_i, \mathcal{U}_{i1}|X_i}(a, u|(x, x')) \\ + \; & \int (\xi_2(x, a, u) - \xi_1(x, a, u)) dF_{\mathcal{A}_i, \mathcal{U}_{i1}|X_i}(a, u|(x, x')) \\ = \; & \beta_2(x \to x'|X_i = (x, x')) + \lambda_2(x) \end{aligned} \tag{3.11}$$

The APE is indexed by the time period, since the structural function is not stationary in the observables, $\xi_2(x, a, u) = \xi_1(x, a, u) + \lambda_2(x)$.[23] Now to identify the APE in (3.11), we have to identify $\lambda_2(x)$, the counterfactual trend. Now noting that

$$\begin{aligned} & E[Y_{i2} - Y_{i1}|X_i = (x, x)] \\ = \; & \int (\xi_2(x, a, u) - \xi_1(x, a, u)) dF_{\mathcal{A}_i, \mathcal{U}_{i1}|X_i}(a, u|(x, x)) \\ = \; & \int \lambda_2(x) dF_{\mathcal{A}_i, \mathcal{U}_{i1}|X_i}(a, u|(x, x)) = \lambda_2(x). \end{aligned} \tag{3.12}$$

The last equality illustrates why identification of the counterfactual trend would not be achieved if the nonstationary component of the structural function depended on the unobservables as pointed out in Variation 3.3.1. Plugging (3.12) into (3.11) identifies the APE.

Note that the setup of Evdokimov (2010) may also accommodate generalized time effects as in $\lambda_t(X_{it})$. The key difference between the two approaches is

[23] This implies that $\xi_2(x', a, u) - \xi_1(x, a, u) = \xi_1(x', a, u) + \lambda(x') - \xi_1(x, a, u) - \lambda(x)$, which is equal to $\xi_1(x', a, u) - \xi_1(x, a, u)$ iff $\lambda_t(x') = \lambda_t(x)$. For instance, if $\lambda_t(x) = \lambda$, then this condition would be fulfilled and the APE would be constant across time.

that with the identification of the entire distribution of $\mathcal{U}_{it}|X_{it}$, Evdokimov (2010) can allow for time heterogeneity, whereas the quasi-differencing approach used here does not allow for time heterogeneity.

Since the structural functions in (3.7) and (3.9) are different, it may be hard to compare the two models. On the other hand, (3.7) is very similar to the setup in Variation 3.3.4, which is given by

$$\begin{aligned} Y_{it} &= \mu(X_{it}, \mathcal{A}_i) + \mathcal{U}_{it} \\ \mathcal{U}_{it}|X_i, \mathcal{A}_i &\overset{d}{=} \mathcal{U}_{it}, \quad t = 1, 2 \end{aligned} \tag{3.13}$$

$$\begin{aligned} & E[Y_{i2} - Y_{i1}|X_i = (x, x')] \\ = & \int (\mu(x', a) + u_2 - (\mu(x, a) + u_1)) dF_{\mathcal{A}_i, \mathcal{U}_{i1}, \mathcal{U}_{i2}|X_i}(a, u_1, u_2|(x, x')) \\ = & \int (\mu(x', a) - \mu(x, a)) dF_{\mathcal{A}_i|X_i}(a|(x, x')) \\ + & E[\mathcal{U}_{i2}|X_i = (x, x')] - E[\mathcal{U}_{i1}|X_i = (x, x')] \\ \overset{(3.13)}{=} & \beta(x \to x'|X_i = (x, x')) + E[\mathcal{U}_{i2}] - E[\mathcal{U}_{i1}] \end{aligned} \tag{3.14}$$

Here, the counterfactual trend is given by $\{E[\mathcal{U}_{i2}] - E[\mathcal{U}_{i1}]\}$. Since the distribution of $\mathcal{U}_{it}$ is independent of X_i and $\mathcal{A}_i$, we can use any stayer subpopulation to identify the counterfactual trend.

$$\begin{aligned} & E[Y_{i2} - Y_{i1}|X_i = (x, x)] \\ = & E[\mathcal{U}_{i2} - \mathcal{U}_{i1}|X_i = (x, x)] \overset{(3.13)}{=} E[\mathcal{U}_{i2}] - E[\mathcal{U}_{i1}]. \end{aligned} \tag{3.15}$$

Plugging (3.15) into (3.14) identifies the APE. In Evdokimov (2010), $E[\mathcal{U}_{it}|X_{it}] \neq$

$E[\mathcal{U}_{it}]$. If this were true, then we could not identify the APE using the quasi-differencing approach.

It is important to point out that the classical identification results in Evdokimov (2010) require stronger assumptions than the quasi-differencing approach. Hence, it is hard to compare the two approaches. The above examples are simply meant to illustrate the advantages of imposing stronger assumptions and identifying all structural objects.

Relation to the Classical Difference-in-Difference Model. Let $X_{it} \in \{0,1\}$ and $T = 2$. Now consider the following two linear models. Model 1 follows the setup in Variations 3.3.1 and 3.3.2,

$$\begin{aligned} Y_{it} &= \beta X_{it} + \mathcal{A}_i + \lambda_t + \mathcal{U}_{it}, \quad t = 1, 2 \\ \mathcal{U}_{i2}|X_i, \mathcal{A}_i &\overset{d}{=} \mathcal{U}_{i1}|X_i, \mathcal{A}_i. \end{aligned} \tag{3.16}$$

Model 2 follows the setup in Variation 3.3.4,

$$\begin{aligned} Y_{it} &= \beta X_{it} + \mathcal{A}_i + \mathcal{U}_{it} \\ \mathcal{U}_{it}|X_i, \mathcal{A}_i &\overset{d}{=} \mathcal{U}_{it}, \quad t = 1, 2. \end{aligned} \tag{3.17}$$

Under both models, β is identified by the probability limit of the difference-in-difference estimator $\hat{\beta} \equiv \sum_{i=1}^{n} \Delta Y_i 1\{X_i = (0,1)\} / \sum_{i=1}^{n} 1\{X_i = (0,1)\} - \sum_{i=1}^{n} \Delta Y_i \times 1\{X_i = (0,0)\} / \sum_{i=1}^{n} 1\{X_i = (0,0)\}$.

For Model 1,

$$\begin{aligned}
\hat{\beta} &\xrightarrow{p} E[Y_{i2} - Y_{i1}|X_i = (0,1)] - E[Y_{i2} - Y_{i1}|X_i = (0,0)] \\
&= \beta + E[\mathcal{A}_i + \lambda_2 + \mathcal{U}_{i2} - (\mathcal{A}_i + \lambda_1 + \mathcal{U}_{i1})|X_i = (0,1)] \\
&- E[\mathcal{A}_i + \lambda_2 + \mathcal{U}_{i2} - (\mathcal{A}_i + \lambda_1 + \mathcal{U}_{i1})|X_i = (0,0)] \\
&\overset{(3.16)}{=} \beta + E[\mathcal{U}_{i1} - \mathcal{U}_{i1}|X_i = (0,1)] - E[\mathcal{U}_{i1} - \mathcal{U}_{i1}|X_i = (0,0)] \\
&= \beta \qquad (3.18)
\end{aligned}$$

Note that in the above, idiosyncratic shocks may be correlated with the regressors as well as unobservable individual heterogeneity in arbitrary ways. However, time homogeneity has to hold, i.e. the conditional distribution of idiosyncratic shocks has to be the same across time. Furthermore, the structural relationship has to be stationary up to a nonstochastic time effect. Thus, the generalization of Model 1 to a nonseparable model yields Model 1', given by

$$\begin{aligned}
Y_{it} &= \xi(X_{it}, \mathcal{A}_i, \mathcal{U}_{it}) + \lambda_t, \;\; t = 1, 2 \\
\mathcal{U}_{i2}|X_i, \mathcal{A}_i &\overset{d}{=} \mathcal{U}_{i1}|X_i, \mathcal{A}_i. \qquad (3.19)
\end{aligned}$$

The above is a generalization of Athey and Imbens (2006) for the panel context, where a nonstochastic time effect is allowed for.

For Model 2, on the other hand,

$$\begin{aligned}
\hat{\beta} &\xrightarrow{p} E[Y_{i2} - Y_{i1}|X_i = (0,1)] - E[Y_{i2} - Y_{i1}|X_i = (0,0)] \\
&= \beta + E[\mathcal{U}_{i2} - \mathcal{U}_{i1}|X_i = (0,1)] - E[\mathcal{U}_{i2} - \mathcal{U}_{i1}|X_i = (0,0)] \\
&\overset{(3.17)}{=} \beta + E[\mathcal{U}_{i2} - \mathcal{U}_{i1}] - E[\mathcal{U}_{i2} - \mathcal{U}_{i1}] \\
&= \beta. \qquad (3.20)
\end{aligned}$$

The intuition here is that, even though idiosyncratic shocks are heterogeneously distributed across time, they are independent of regressors and individual heterogeneity. Hence, counterfactual trends for mover subpopulations, (x, x') can be identified from stayer subpopulations, (x, x). Generalizing this model to the nonseparable setup yields Model 2', given by

$$\begin{aligned} Y_{it} &= \xi(X_{it}, \mathcal{A}_i, \mathcal{U}_{it}), \\ \mathcal{U}_{it}|X_i, \mathcal{A}_i &\overset{d}{=} \mathcal{U}_{it}.\ t = 1, 2 \end{aligned} \tag{3.21}$$

Recall that the parallel-trends assumption in the linear model implies that $Cov(\mathcal{U}_{it}, \mathcal{A}_i) = 0$ and $Cov(\mathcal{U}_{it}, X_{it}) = 0$ for $t = 1, 2$. Hence, the above may be viewed as the nonseparable version of the parallel-trends assumption.[24]

3.3.4 Menu of Identifying Assumptions

The previous section gives directions to some new identification strategies. For within-group identification, we use time homogeneity and the stationarity of the structural function in unobservables. For instance, the structural function may be given as follows

$$Y_{it} = \xi(X_{it}, \mathcal{A}_i, \mathcal{U}_{it}) + \lambda_t(X_{it}). \tag{3.22}$$

For within-period identification, restrictions on individual heterogeneity, such as $F_{\mathcal{A}_i, \mathcal{U}_{it}|X_i}(.,.|(x,x)) = F_{\mathcal{A}_i, \mathcal{U}_{it}|X_i}(.,.|(x,x'))$, allow us to identify the effect of interest

[24] In the empirical literature, the parallel-trends assumption is tested by look at pre-trial years and ensuring that both control and treatment groups follow the same trends in average within-group changes across time. It is important to note that if only changes are considered, both Model 1' and Model 2' would yield the same 'parallel-trends' prediction. However, they have different testable implications if we look at the distribution of unobservables as a whole. This issue will be left for future work.

in period t in the presence of time heterogeneity.

In the following section, we will give generalizations of (3.22) as well as other options for within-period identification.

Within-Group Identification: Restrictions on Structural Function and Time Heterogeneity

Time homogeneity was proposed in CFHN2013, where the structural function is assumed to be stationary in both observables and unobservables. In this paper, we allow the structural function to be nonstationary in observables, however we maintain its stationarnity in unobservables. The following theorem gives conditions under which one can identify the APE of a subpopulation using average within-group changes across time, when we allow for generalized time effects in the structural function.

$$Y_{it} = \xi(X_{it}, \mathcal{A}_i, \mathcal{U}_{it}) + \lambda_t(h(X_i)) \tag{3.23}$$

where $h : \mathcal{X}^T \mapsto \mathbb{R}$.[25] The following theorem gives conditions under which one can identify the APE given the above structural relationship. Let $\underline{x}_t$ denote the t^{th} column of $\underline{x}$.

Theorem 3.3.2 *(Within-Group Identification)*
Given Assumption 3.3.1, if the following conditions are satisfied for some $t, \tau \in \{1, 2, ...T\}$, $\tau \neq t$,

(i) $F_{\mathcal{U}_{it}|X_i,\mathcal{A}_i}(.|.,.) = F_{\mathcal{U}_{i\tau}|X_i,\mathcal{A}_i}(.|.,.)$,

(ii) $\xi_\tau(x, a, u) = \xi_t(x, a, u) + \lambda_\tau(h(\underline{x}))$, $\forall (x, a, u) \in \mathcal{X} \times \mathbb{R}^2$

[25]Note that this mapping is not onto. $\mathcal{X}^T$ is a finite set, whereas $\mathbb{R}$ is infinite.

(iii) $h(\underline{x}) = h(\underline{x}^c)$ *and* $\underline{x}_t^c = \underline{x}_\tau^c$,

then

$$\beta_t(\underline{x}_t \to \underline{x}_\tau | X_i = \underline{x}) \quad = \quad E[Y_{i\tau} - Y_{it} | X_i = \underline{x}] - E[Y_{i\tau} - Y_{it} | X_i = \underline{x}^c]$$

The above theorem shows that in order to allow for a time effect that depends on X_i, there has to be a restriction $h(X_i)$ that allows one to identify $\lambda_\tau(h(\underline{x}))$ from another subpopulation, a control group, $\underline{x}^c$, this is implied by condition (iii) in the above theorem. To illustrate this, given (i) and (ii),

$$\begin{aligned} E[Y_{i\tau} - Y_{it} | X_i = \underline{x}] &= \int (\xi_\tau(\underline{x}_\tau, a, u) - \xi_t(\underline{x}_t, a, u)) dF_{\mathcal{A}_i, \mathcal{U}_{i1} | X_i}(a, u | \underline{x}) \\ &= \beta_\tau(\underline{x}_t \to \underline{x}_\tau | X_i = \underline{x}) + \lambda_\tau(h(\underline{x})). \end{aligned}$$

Hence, the counterfactual trend here is $\lambda_\tau(h(\underline{x}))$. Now we would like to identify this counterfactual trend from a subpopulation $\underline{x}^c$, given by

$$\begin{aligned} E[Y_{i\tau} - Y_{it} | X_i = \underline{x}^c] &= \int (\xi_\tau(\underline{x}_\tau^c, a, u) - \xi_t(\underline{x}_t^c, a, u)) dF_{\mathcal{A}_i, \mathcal{U}_{i1} | X_i}(a, u | \underline{x}^c) \\ &= \int (\xi_\tau(\underline{x}_\tau^c, a, u) - \xi_t(\underline{x}_t^c, a, u)) dF_{\mathcal{A}_i, \mathcal{U}_{i1} | X_i}(a, u | \underline{x}^c) \\ &\overset{(\underline{x}_\tau^c = \underline{x}_t^c)}{=} \lambda_\tau(h(\underline{x}^c)) \\ &\overset{(h(\underline{x}) = h(\underline{x}^c))}{=} \lambda_\tau(h(\underline{x})). \end{aligned}$$

The last two equalities follow from the conditions given in (iii).

Within-group identification strategies are very popular in the empirical literature, the above result can be thought of as a generalization of the fixed-effects methods, such as the difference-in-difference model discussed above. Section 3.4

proposes tests for the models implied by the above theorem, which may be viewed as nonparametric tests of the fixed-effects assumption.

Within-Period Identification: Restrictions on Individual Heterogeneity

Within-period identification solves a cross-sectional identification problem using the panel information. Traditional cross-sectional identification strategies may be applied, such as nonparametric instrumental variables approaches which require completeness conditions in the nonparametric setup, which are known to be very strong. What we propose here are alternative strategies to the classical cross-sectional identification strategies. For continuous regressors, Bester and Hansen (2009) show how average effects in a specific time period can be identified through index restrictions on the distribution of individual heterogeneity, while relying on special properties of continuous variables.[26] In many ways, within-period identification strategies presented here may be thought of as the counterpart of Bester and Hansen (2009) for discrete regressors.

Theorem 3.3.3 *(Within-Period Identification)*
Given Assumption 3.3.1, if

(i) $F_{\mathcal{A}_i,\mathcal{U}_{it}|X_i}(.,.|.) = F_{\mathcal{A}_i,\mathcal{U}_{it}|h(X_i)}(.,.|.)$,

(ii) $h(\underline{x}) = h(\underline{x}^c)$, *where* $\underline{x} \neq \underline{x}^c$,

then

$$\begin{aligned}\beta_t(\underline{x}_t^c \to \underline{x}_t | X_i = \underline{x}) &= \beta_t(\underline{x}_t^c \to \underline{x}_t | X_i = \underline{x}^c) \\ &= E[Y_{it}|X_i = \underline{x}] - E[Y_{it}|X_i = \underline{x}^c].\end{aligned}$$

[26] Bester and Hansen (2009) impose the restriction that $\mathcal{U}_{it}|X_i, \mathcal{A}_i \overset{d}{=} \mathcal{U}_{it}|X_{it}, \mathcal{A}_i$.

Note in the above that $\beta_t(\underline{x}_t^c \to \underline{x}_t | X_i = \underline{x}) = \beta_t(\underline{x}_t^c \to \underline{x}_t | X_i = \underline{x}^c)$ by (i) and (ii). Hence, we identify the effect for the control group as well. The particular effect that we identify is determined by the realizations of X_{it} for both subpopulations in period t, $\underline{x}_t$ and $\underline{x}_t^c$, respectively.

There are two types of restrictions that may be of particular interest here, (1) exchangeability restrictions and (2) exclusion restrictions. Exchangeability restrictions, which are similar in spirit to the restrictions in Bester and Hansen (2009), were proposed in Altonji and Matzkin (2005) to relax the conditional independence assumption in the identification of average derivatives.[27] In that case, exchangeability is not sufficient for identification. However, for discrete regressors, exchangeability is sufficient. We give an empirical example where we use the exchangeability restriction to identify the effect of interest.

Example 3.3.1 *(Female Labor Participation: Exchangeability Restriction)*
Let Y_{it} be a binary variable for employment status, X_{it} a binary variable for having a child under 6 months of age.

$$Y_{it} = \xi_t(X_{it}, \mathcal{A}_i, \mathcal{U}_{it}) \tag{3.24}$$

The effect of child-bearing on female labor participation is a classical example of panel data models using fixed effects approach, i.e. within-group identification, whether in linear or nonlinear models, as in Heckman and Heckman and MaCurdy (1980), Heckman and MaCurdy (1982), Hyslop (1999), and Fernandez-Val (2009). In the presence of time heterogeneity, within-group identification could not identify the APE of child-bearing on female labor participation. However, if we assume

[27] Note that in Hoderlein and White (2012) time homogeneity alone is not sufficient for the identification of the local average structural derivative, but a conditional independence restriction is also imposed.

$F_{\mathcal{A}_i,\mathcal{U}_{it}|X_i} = F_{\mathcal{A}_i,\mathcal{U}_{it}|\sum_t X_{it}}$, then within-period identification is possible. The justification behind such an assumption is that, generally speaking, individuals are more likely to select how many children to have in a given time period but they cannot fully control when their children are born. This would imply that the number of times that an individual has children under 6 months of age, i.e. $\sum_{t=1}^{T} X_{it}$, characterizes unobservable characteristics, as opposed to X_i. For $T = 2$, this assumption implies that subpopulations $(0,1)$ and $(1,0)$ have the same distribution of unobservables, i.e. $F_{\mathcal{A}_i,\mathcal{U}_{it}|X_i}(.,.|(0,1)) = F_{\mathcal{A}_i,\mathcal{U}_{it}|X_i}(.,.|(1,0))$ for $t = 1, 2$. Hence, they can be used to identify the counterfactual for one another. This yield the following APEs for the first and second time period.

$$\begin{aligned} & \beta_1(0 \to 1|X_i = (0,1)) \\ = \; & \beta_1(0 \to 1|X_i = (1,0)) = E[Y_{i1}|X_i = (1,0)] - E[Y_{i1}|X_i = (0,1)] \\ & \beta_2(0 \to 1|X_i = (0,1)) \\ = \; & \beta_2(0 \to 1|X_i = (1,0)) = E[Y_{i2}|X_i = (0,1)] - E[Y_{i2}|X_i = (1,0)]. \end{aligned}$$

Hence, within-period changes across subpopulations identify the effect of interest for individuals that switch their X_{it} across time. Note that in the presence of time heterogeneity, $\beta_1(0 \to 1|X_i = (0,1)) \neq \beta_2(0 \to 1|X_i = (0,1))$. The intuition here is that in the presence of changing macroeconomic conditions, the APE in one year is generally different from another, even for the same subpopulation.

Another class of restrictions is exclusion restrictions such as $F_{\mathcal{A}_i,\mathcal{U}_{it}|X_i}(.) = F_{\mathcal{A}_i,\mathcal{U}_{it}|X_{i(\tau)}}(.)$, where $X_{i(t)} = \{X_{i1}, .., X_{i,\tau-1}, X_{i,\tau+1}, ..., X_{iT}\}$. The distribution of unobservables is determined by observables, $X_{i(\tau)}$, hence $X_{i\tau}$ does not provide additional information about these unobservables. This assumption implies selection on observables, $X_{i(\tau)}$, as in Heckman and Robb (1985).

Example 3.3.2 *(Hypothetical Example: Family Panel Model with Exclusion Restriction)*

Now let us think of a panel of families and their children. Our outcome variable is income of the child, and the variable of interest is a binary variable for whether the parents helped their child with college tuition or not.

$$Y_{ij} \quad = \quad \xi_j(X_{ij}, \mathcal{A}_i, \mathcal{U}_{ij}), \quad i = 1, 2, ..., n, j = 1, 2, ..., J. \tag{3.25}$$

For simplicity, assume that all families in the panel have exactly 2 children, i.e. J=2. In this setup, it is very important to allow for unobservable heterogeneity at the child level, since every child has his/her own abilities. Thus, a within-period approach is more suitable. In this situation, once we condition on whether parents help their first child with their college tuition, whether they help their second child contains no further information about the unobservable characteristics of the family. This justifies an exclusion restriction on the distribution of unobservables, $F_{\mathcal{A}_i,\mathcal{U}_{ij}|X_i} = F_{\mathcal{A}_i,\mathcal{U}_{ij}|X_{i1}}$, $j = 1, 2$. *As a result, we can identify the effect of interest for the second child. If we observe all subpopulations,* $(0,0)$, $(0,1)$, $(1,0)$ *and* $(1,1)$, *then we can identify the following objects:*

$$\begin{aligned}
& \beta_2(0 \to 1 | X_i = (0,1)) \\
= \quad & \beta_2(0 \to 1 | X_i = (0,1)) = E[Y_{i2} | X_i = (0,1)] - E[Y_{i2} | X_i = (0,0)] \\
& \beta_2(0 \to 1 | X_i = (1,0)) \\
= \quad & \beta_2(0 \to 0 | X_i = (1,1)) = E[Y_{i2} | X_i = (1,1)] - E[Y_{i2} | X_i = (1,0)],
\end{aligned}$$

which are the APE of parents' help with college tuition on the second child's income for subpopulations that did not help their first child with college tuition and the

subpopulation that helped their first child with college tuition, respectively.

3.4 Testing Identifying Assumptions

The purpose of the previous section is to better understand how identification of a specific object is achieved and to characterize the trade-off between individual and time heterogeneity as well as assumptions on the the structural function. It also provides different assumptions that can achieve identification of the same object and may not be justified *a priori*. Fortunately, the identification strategies proposed here have testable implications. In this section, we propose tests for these implications.

3.4.1 Basic Testing Problem

The identifying assumptions proposed above impose restrictions on $\{\xi_t, F_{\mathcal{A}_i,\mathcal{U}_{it}|X_i}\}_{t=1}^{T}$, which imply equality restrictions on the conditional distribution of the outcome variable.[28] The equality of two distributions is a well-known problem in statistics, the so-called two-sample problem, where the two samples are random and independent of each other. When testing within-group identification in the panel setting, the two samples are not independent, since both samples are realizations of the same individuals across time. For within-period identification, our statistic is a linear combination of KS and CM statistics. For both cases, bootstrap methods are proposed to obtain the critical values for the statistics.

It is important to note that the tests proposed here are not over-identification

[28] As a result, these restrictions imply restrictions on the set of distributions, $\mathcal{D} \equiv \{F_{Y_{it}^x|X_i}(.|\underline{x}) : t = 1, 2, ..., T, x \in \mathcal{X}, \underline{x} \in \mathcal{X}^T\}$, where X_{it} is fixed. Note that this set is not observable. Now let us define the set of observable distributions $\mathcal{D}^o \equiv \{F_{Y_{it}|X_i}(.|\underline{x}) : t = 1, 2, ..., T, \underline{x} \in \mathcal{X}^T\}$. If the restrictions imply that the map $\mathcal{D}^o \to \mathcal{D}$ is non-injective, then the identifying assumptions imply equality restrictions on the conditional distribution of the outcome variable.

tests. They test implications of the identifying assumptions. Hence, rejecting them is clear evidence against the identifying assumptions. Two examples are given below.

Example 3.4.1 *(Time Homogeneity for the Scalar Binary Example)*
Let $T = 2$ and $X_{it} \in \{0, 1\}$. Assume a stationary structural function and time homogeneity, $\xi_t = \xi$ and $F_{\mathcal{A}_i,\mathcal{U}_{i1}|X_i} = F_{\mathcal{A}_i,\mathcal{U}_{i2}|X_i}$ for $t = 1, 2$. This implies the following restriction

$$F_{Y_{i1}^{\underline{x}}|X_i}(.|\underline{x}) = F_{Y_{i2}^{\underline{x}}|X_i}(.|\underline{x}), \quad \forall \underline{x}. \tag{3.26}$$

Back to our job training program, if a subpopulation does not change its job training status across time, then the distribution of its outcome variable should also not change over time. Note that this is exactly what allows us to identify the effect of job training on earnings for subpopulations that switch their job training status across time.
Thus, we have the following restrictions

$$\begin{aligned} F_{Y_{i1}|X_i}(.|(0,0)) &= F_{Y_{i2}|X_i}(.|(0,0)) \\ F_{Y_{i1}|X_i}(.|(1,1)) &= F_{Y_{i2}|X_i}(.|(1,1)) \end{aligned} \tag{3.27}$$

However, note that the APE for subpopulations $(0, 1)$ and $(1, 0)$ is just-identified as follows

$$\begin{aligned} \beta_1(0 \to 1|X_i = (0,1)) &= E[Y_{i2}|X_i = (0,1)] - E[Y_{i1}|X_i = (0,1)] \\ \beta_1(0 \to 1|X_i = (1,0)) &= E[Y_{i1}|X_i = (1,0)] - E[Y_{i2}|X_i = (1,0)]. \end{aligned}$$

Thus, the tests proposed here may be implemented even if there are no over-

identifying restrictions.

Example 3.4.2 *(Exclusion Restriction Example for the Scalar Binary Example) Again, let $T = 2$ and $X_{it} \in \{0, 1\}$. We now impose the following exclusion restriction $F_{\mathcal{A}_i, \mathcal{U}_{it}|X_i}(., .|\underline{x}) = F_{\mathcal{A}_i, \mathcal{U}_{it}|X_{i1}}(., .|\underline{x}_1)$. This assumption implies that*

$$F_{Y_{i1}^x|X_i}(., .|(x, x')) = F_{Y_{i1}^x|X_i}(., .|(x, x'')) \quad \forall x, x', x'' \in \mathcal{X}, x' \neq x'' \tag{3.28}$$

Hence, we have testable implications for the first period, where $F_{Y_{i1}|X_i}(.|(0, 0)) = F_{Y_{i1}|X_i}(.|(0, 1))$ and $F_{Y_{i1}|X_i}(.|(1, 0)) = F_{Y_{i1}|X_i}(.|(1, 1))$. Here our object of interest is also just identified.

$$\begin{aligned} \beta_2(0 \to 1|X_i = (0, 1)) &= E[Y_{i2}|X_i = (0, 1)] - E[Y_{i2}|X_i = (0, 0)] \\ \beta_2(0 \to 1|X_i = (1, 0)) &= E[Y_{i2}|X_i = (1, 1)] - E[Y_{i2}|X_i = (1, 0)] \end{aligned}$$

In the following, the classical two-sample problem is reviewed. Then, the validity of the bootstrap procedures for within-group and within-period identification is shown.

3.4.2 Review of the Classical Two-Sample Problem

Given two independent random samples of continuous variables, $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^m$, i.e. all the random variables are mutually independent, we would like to test the equality of the two distributions from which the samples are drawn, i.e.

$$H_0 : F_X(.) = F_Y(.) \text{ versus } H_1 : F_X(.) \neq F_Y(.) \tag{3.29}$$

where $F_W(.)$ denotes the distribution of W. Now let $F_{W,n}(.) = \sum_{i=1}^{n} 1\{W_i \leq .\}/n$, the empirical cumulative distribution function (cdf). Now we can define the Kolmogo-rov-Smirnov (KS) and Cramer-von-Mises (CM) statistics as follows,

$$KS_{n\mathcal{Z}} = \sup_{z\in\mathcal{Z}} |F_{X,n}(z) - F_{Y,n}(z)| \equiv \|F_{X,n}(.) - F_{Y,m}(.)\|_{\infty,\mathcal{Z}} \tag{3.30}$$

and

$$CM_{n,F_Z} = \int (F_{X,n}(z) - F_{Y,n}(z))^2 dF_Z(z) \equiv \|F_{X,n}(.) - F_{Y,m}(.)\|_{2,F_Z}, \tag{3.31}$$

respectively. F_Z is the pooled distribution. Note that $F_Z(.) = F_X(.) = F_W(.)$ under the null. Since $F_Z(.)$ is unknown, CM_{n,F_Z} is infeasible. There is a feasible statistic equivalent to CM_{n,F_Z} that will be discussed in the following.

Under the independence of the two random samples and the continuity of the underlying pooled distribution, both statistics are pivotal, i.e. their distribution does not depend on the distributions of X and Y. The result follows from the probability integral transform theorem and is given for the one-sample KS statistic in Gibbons (1985).[29] For the two-sample CM statistic, Anderson (1962) gives the asymptotic distribution of the two-sample CM statistic. For a detailed outline of the development of the KS statistic, see Andrews (1997).

Let $N = n + m$, and $\{Z_i\}_{i=1}^{N}$ be the pooled sample of $\{X_i\}_{i=1}^{n}$ and $\{Y_i\}_{i=1}^{m}$. Under the above conditions, the aforementioned statistics are equivalent to the

[29]The extension to the two-sample case is straightforward and is given in Section 3.6.3.

following

$$KS_{n,N} = \max_{1\leq i\leq N} |F_{X,n}(Z_i) - F_{Y,n}(Z_i)| \equiv \|F_{X,n}(.) - F_{Y,m}(.)\|_{\infty,N} \quad (3.32)$$

and

$$CM_{n,N} = \frac{nm}{N^2}\sum_{i=1}^{N}(F_{X,n}(Z_i) - F_{Y,n}(Z_i))^2 \equiv \|F_{X,n}(.) - F_{Y,m}(.)\|_{2,N}, \quad (3.33)$$

which are based on the empirical measure.

There are two related testing problems that we are interested in here. In the following, the cross-sectional independence assumption is maintained. First, the testing of within-group identifying assumptions, which are generalizations of time homogeneity, is a paired-sample problem, which deviates from the classical two-sample problem due to the dependence between the samples. The dependence is a result of the nature of panel data, where we track the same individuals across time. Hence, the sources of dependence between the observations across time are (1) time-invariant unobservables and (2) possible dependence between idiosyncratic shocks.

Testing the equality of two distributions with dependence has recently attracted some interest in the statistics literature. Quessy and Ethier (2012) propose the use of the multiplier method to adjust CM and characteristic function tests for the k-sample problem, where the samples are dependent. In the presence of a time effect, the data must be appropriately demeaned before testing the equality of distributions. In section 3.4.3, a bootstrap method is proposed for both the KS and CM tests that is shown to be asymptotically valid to test time homogeneity

in the presence of generalized time effects.

Testing within-period identification under cross-sectional independence is a k-sample problem, where the samples are independent. Hence, it is a more straightforward extension of the two-sample problem and is discussed in section 3.4.4.

3.4.3 Testing Within-Group Identification

For within-group identification, there are two specific cases, time homogeneity with and without a generalized time effect. In the following, bootstrap-adjusted KS and CM statistics to test both variants of time homogeneity are proposed. To simplify illustration, the case where $T = 2$ is examined. Extensions to $T > 2$ are discussed in section 3.4.3.

Time Homogeneity

Under time homogeneity, we assume that for $i = 1, 2, ..., n$

$$\begin{aligned} Y_{it} &= \xi(X_{it}, \mathcal{A}_i, \mathcal{U}_{it}), \quad t = 1, 2 \\ F_{\mathcal{U}_{i2}|X_i,\mathcal{A}_i}(.|.) &= F_{\mathcal{U}_{i1}|X_i,\mathcal{A}_i}(.|.). \end{aligned} \tag{3.34}$$

It is important to note that the two time periods need not be adjacent to each other. We refer to them as period 1 and 2 for simplicity, but they could be any two periods in a time series.

Let $X_i = (X_{i1}, X_{i2})$, time homogeneity implies the following

$$F_{Y_{i1}|X_i}(.|(x,x)) = F_{Y_{i2}|X_i}(.|(x,x)),\ \forall x \in \mathcal{X}. \tag{3.35}$$

Recall that $|\mathcal{X}| = K$ by Assumption 3.3.1. Integrating the above with respect to $F_{X_i|\Delta X_i}(.|0)$, where $\Delta X_i \equiv X_{i2} - X_{i1}$, we obtain the following hypothesis

$$\begin{aligned} & H_0^1 : F_{Y_{i1}|\Delta X_i}(.|0) = F_{Y_{i2}|\Delta X_i}(.|0) \\ \text{vs.} \quad & H_A^1 : F_{Y_{i1}|\Delta X_i}(.|0) \neq F_{Y_{i2}|\Delta X_i}(.|0) \end{aligned} \tag{3.36}$$

Since Y_{i1} and Y_{i2} are observations from the same individual, the above testing problem is a paired-sample problem. To simplify notation, let $F_t(.|\Delta X_i = 0) \equiv F_{Y_{it}|X_i}(.|0)$ and $F_{t,n}(.|\Delta X_i = 0)$ denote the empirical counterpart of $F_t(.|\Delta X_i = 0)$ for $t = 1, 2$, given as follows

$$F_{t,n}(.|\Delta X_i = 0) = \frac{\sum_{i=1}^n 1\{Y_{it} \leq .\}1\{\Delta X_i = 0\}}{\sum_{i=1}^n 1\{\Delta X_i = 0\}}.$$

Now the paired KS and CM tests are defined by the following

$$\begin{aligned} KS_{n,\mathcal{Y}} &= \sup_{y \in \mathcal{Y}} |\sqrt{n}(F_{1,n}(.|\Delta X_i = 0) - F_{2,n}(.|\Delta X_i = 0))| \\ &\equiv \|F_{1,n}(.|\Delta X_i = 0) - F_{2,n}(.|\Delta X_i = 0)\|_{\infty,\mathcal{Y}} \end{aligned} \tag{3.37}$$

$$\begin{aligned} CM_{n,\phi} &= \int \{\sqrt{n}(F_{1,n}(y|\Delta X_i = 0) - F_{2,n}(y|\Delta X_i = 0))\}^2 \phi(y) dy \\ &\equiv \|F_{1,n}(.|\Delta X_i = 0) - F_{2,n}(.|\Delta X_i = 0)\|_{2,\phi} \end{aligned} \tag{3.38}$$

where ϕ is some user-specified density.

Given the dependence between Y_{i1} and Y_{i2}, the probability-integral transform theorem no longer applies and hence the above statistics are not pivotal. The following bootstrap procedure is proposed to adjust the p-value of the above statistic. Let $\{Y_i, X_i\} = \{(Y_{i1}, Y_{i2}), (X_{i1}, X_{i2})\}$.[30] Let $\hat{F}^b_{t,n}(.|\Delta X_i = 0)$ be the empirical cdf of the b^{th} bootstrap sample, $(\hat{Y}^b_i, \hat{X}^b_i)$ given as follows

$$\hat{F}^b_{t,n}(.|\Delta X_i = 0) \quad = \quad \frac{\sum_{i=1}^n 1\{\hat{Y}^b_{it} \leq .\}1\{\Delta \hat{X}^b_i = 0\}}{\sum_{i=1}^n 1\{\Delta \hat{X}^b_i = 0\}}.$$

The bootstrap procedure proposed here is given by the following:

1. Compute the statistic $KS_{n,\mathcal{Y}}$ and $CM_{n,\phi}$ for $\{\{Y_1, X_1\}, ..., \{Y_n, X_n\}\}$, hereinafter the original sample.

2. Resample n observations $\{\{\hat{Y}_1, \hat{X}_1\}, ..., \{\hat{Y}_n, \hat{X}_n\}\}$ with replacement from the original sample. Compute the centered statistics

$$\begin{aligned} & KS^b_{n,\mathcal{Y}} \\ = \; & \|\sqrt{n}\{\hat{F}^b_{1,n}(.|\Delta X_i = 0) - \hat{F}^b_{2,n}(.|\Delta X_i = 0) \\ & -(F_{1,n}(.|\Delta X_i = 0) - F_{2,n}(.|\Delta X_i = 0))\}\|_{\infty,\mathcal{Y}} \\ & CM^b_{n,\phi} \\ = \; & \|\sqrt{n}\{\hat{F}^b_{1,n}(.|\Delta X_i = 0) - \hat{F}^b_{2,n}(.|\Delta X_i = 0) \\ & -(F_{1,n}(.|\Delta X_i = 0) - F_{2,n}(.|\Delta X_i = 0))\}\|_{2,\phi} \end{aligned}$$

3. Repeat 1-2 B times.

[30] Note this is not a matrix.

4. Calculate the p-values of the tests with

$$\begin{aligned} p_{KS,n} &= \sum_{b=1}^{B} 1\{KS^b_{n,\mathcal{Y}} > KS_{n,\mathcal{Y}}\} \\ p_{CM,n} &= \sum_{b=1}^{B} 1\{CM^b_{n,\phi} > CM_{n,\phi}\}. \end{aligned}$$

Reject if p-value is smaller than some significance level α.

Note that as a paired sample problem, the resampling procedure is very similar to a one-sample problem, where we just resample across i. The above procedure approximates the distribution of the KS and CM statistics, under the null $F_1(.|\Delta X_i = 0) = F_2(.|\Delta X_i = 0)$, which allows us to then estimate the p-value. The following theorem shows that the bootstrap-adjusted tests have asymptotic level α and are consistent against fixed alternatives.

Theorem 3.4.1 *Given that $\{Y_i, X_i\}_{i=1}^n$ is an iid sequence, $|\mathcal{X}| = K$, $P(\Delta X_i = 0) > 0$, and $F_t(.|\Delta X_i)$ are non-degenerate for $t = 1, 2$, the procedure described in 1-4 for $KS_{n,\mathcal{Y}}$ and $CM_{n,\phi}$ to test H_0^1 (i) provides correct asymptotic size α and (ii) is consistent against any fixed alternative.*

The proof is in Appendix 3.6.2. The convergence of the bootstrap empirical process follows by straightforward application of results in der Vaart and Wellner (2000). Since the limit process is a tight Brownian bridge and both test statistics are norms thereof, the statistics have positive densities. Hence, the correct asymptotic size and consistency of the test follows.

Time Homogeneity with Time Effects

For time homogeneity with generalized time effects, we assume that

$$\begin{aligned} Y_{it} &= \xi(X_{it}, \mathcal{A}_i, \mathcal{U}_{it}) + \lambda_t(X_{it}), \quad t = 1, 2 \\ F_{\mathcal{U}_{i2}|X_i,\mathcal{A}_i}(.|.) &= F_{\mathcal{U}_{i1}|X_i,\mathcal{A}_i}(.|.), \end{aligned} \tag{3.39}$$

where we normalize $\lambda_1(.)$ to be zero and hence drop the subscript for $\lambda_2(.) = \lambda(.)$. Hence,

$$\begin{aligned} F_{Y^x_{i1}|X_i}(.|\underline{x}) &= F_{Y^x_{i2}-\lambda_2(x)|X_i}(.|\underline{x}) \\ &= F_{Y^x_{i2}|X_i}(. + \lambda_2(x)|\underline{x}), \quad \forall x, \underline{x} \end{aligned} \tag{3.40}$$

which implies that

$$F_{Y_{i1}|X_i}(.|(x,x)) = F_{Y_{i2}|X_i}(. + \lambda_2(x)|(x,x)), \quad \forall x \tag{3.41}$$

Now let $\Lambda = (\lambda(x^1), ..., \lambda(x^K))'$ and recall that $|\mathcal{X}| = K$, we introduce the following notation

$$F_{Y_{i2}|\Delta X_i}(., \Lambda|0) = \sum_{k=1}^{K} P(X_i = (x^k, x^k)) F_{Y_{i2}|X_i}(. + \lambda_2(x^k)|(x^k, x^k)).$$

Hence, we can write our hypothesis as follows

$$\begin{aligned} & H_0^2 : F_{Y_{i1}|\Delta X_i}(.|0) = F_{Y_{i2}|\Delta X_i}(., \Lambda|0) \\ \text{vs.} \quad & H_A^2 : F_{Y_{i1}|\Delta X_i}(.|0) \neq F_{Y_{i2}|\Delta X_i}(., \Lambda|0) \end{aligned} \tag{3.42}$$

We will adapt the simplified notation from above $F_2(.,\Lambda|\Delta X_i = 0) \equiv F_{Y_{i2}|\Delta X_i}(.,\Lambda|0)$.

$$\begin{aligned} & F_2(.,\Lambda|\Delta X_i = 0) \\ = & \sum_{k=1}^{K} P(X_{i1} = X_{i2} = x^k|\Delta X_i = 0) F_2(. + \lambda(x^k)|X_{i1} = X_{i2} = x^k). \end{aligned}$$

$F_{t,n}(.|.)$ is the empirical cdf of $F_t(.|.)$. Λ_n is the sample analogue of Λ, given by

$$\begin{aligned} \Lambda_n &= \begin{pmatrix} \lambda_n(x^1) \\ ... \\ \lambda_n(x^K) \end{pmatrix} \\ &= \begin{pmatrix} \frac{\sum_{i=1}^n \Delta Y_i 1\{X_i=(x^1,x^1)\}}{\sum_{i=1}^n 1\{X_i=(x^1,x^1)} \\ ... \\ \frac{\sum_{i=1}^n \Delta Y_i 1\{X_i=(x^K,x^K)\}}{\sum_{i=1}^n 1\{X_i=(x^K,x^K)} \end{pmatrix} \end{aligned} \tag{3.43}$$

$$\begin{aligned} KS_{n,\mathcal{Y}}(\Lambda_n) &= \|\sqrt{n}(F_{1,n}(.|\Delta X_i = 0) - F_{2,n}(.,\Lambda_n|\Delta X_i = 0))\|_{\infty,\mathcal{Y}} \\ CM_{n,\phi}(\Lambda_n) &= \|\sqrt{n}(F_{1,n}(.,\Lambda_n|\Delta X_i = 0) - F_{2,n}(.|\Delta X_i = 0))\|_{2,\phi}. \end{aligned} \tag{3.44}$$

The following bootstrap procedure can be used to adjust the p-values of the above statistics. Let $\hat{\Lambda}_n^b$ be the estimator of Λ in the b^{th} bootstrap sample.

1. Compute the statistics $KS_{n,\mathcal{Y}}(\Lambda_n)$ and $CM_{n,\phi}(\Lambda_n)$ for $\{\{Y_1, X_1\}, ..., \{Y_n, X_n\}\}$, hereinafter the original sample.

2. Resample n observations $\{\{\hat{Y}_1, \hat{X}_1\}, ..., \{\hat{Y}_n, \hat{X}_n\}\}$ with replacement from the

original sample. Compute the centered statistics

$$\begin{aligned} & KS^b_{n,\mathcal{Y}}(\hat{\Lambda}_n) \\ = \; & \|\sqrt{n}\{\hat{F}^b_{1,n}(.|\Delta X_i = 0) - \hat{F}^b_{2,n}(., \hat{\Lambda}^b_n|\Delta X_i = 0) \\ & -(F_{1,n}(.|\Delta X_i = 0) - F_{2,n}(., \Lambda_n|\Delta X_i = 0))\}\|_{\infty,\mathcal{Y}} \\ & CM^b_{n,\phi}(\hat{\Lambda}_n) \\ = \; & \|\sqrt{n}\{\hat{F}^b_{1,n}(.|\Delta X_i = 0) - \hat{F}^b_{2,n}(., \hat{\Lambda}^b_n|\Delta X_i = 0) \\ & -(F_{1,n}(.|\Delta X_i = 0) - F_{2,n}(., \Lambda_n|\Delta X_i = 0))\}\|_{2,\phi} \end{aligned}$$

3. Repeat 1-2 B times.

4. Calculate the p-values of the tests with

$$\begin{aligned} p_{KS,n} &= \sum_{b=1}^{B} 1\{KS^b_{n,\mathcal{Y}}(\hat{\Lambda}_n) > KS_{n,\mathcal{Y}}(\Lambda_n)\} \\ p_{CM,n} &= \sum_{b=1}^{B} 1\{CM^b_{n,\phi}(\hat{\Lambda}_n) > CM_{n,\phi}(\Lambda_n)\}. \end{aligned}$$

Reject if p-value is smaller than some significance level α.

The presence of the generalized time effects adds noise to the empirical cdf. In order to ensure the convergence of the empirical process that are used to compute the KS and CM statistics, we impose the following condition, which ensures that the underlying distribution is uniformly continuous.

Assumption 3.4.1 *(Bounded Density)*

$F_t(.)$ has a density $f_t(.)$ that is bounded, i.e. $\sup_{y\in\mathcal{Y}} |f_t(y)| < \infty$, $t = 1, 2$.[31]

[31] Since we only demean Y_{i2}, it is sufficient to impose the above condition for that time period. It is imposed on both time periods, since the choice of demeaning the variables in the second time period is arbitrary.

Theorem 3.4.2 *Given that* $\{Y_i, X_i\}_{i=1}^n$ *is an iid sequence,* $|\mathcal{X}| = K$*,* $P(\Delta X_i = 0) > 0$*,* $F_t(.|\Delta X_i = 0)$ *is non-degenerate for* $t = 1, 2$*, and Assumption 3.4.1 holds, the procedure described in 1-4 for* $KS_{n,\mathcal{Y}}(\Lambda_n)$ *and* $CM_{n,\phi}(\Lambda_n)$ *to test* H_0^2 *(i) provides correct asymptotic size* α *and (ii) is consistent against any fixed alternative.*

The proof is given in Appendix 3.6.2. The key difference between the above result and Theorem 3.4.1 is that we have to ensure that the functional Delta method applies to ensure that the empirical process converges, hence we impose Assumption 3.4.1. The intuition behind imposing this regularity condition is that by demeaning the variables, we are introducing asymptotically normal noise to the empirical process. Assumption 3.4.1 ensures that the empirical process converges nonetheless to a Brownian bridge, by allowing us to apply the functional delta method. From here, it is straightforward to show that the bootstrap empirical process converges to the same tight limit process as the empirical process. Then, we show that the bootstrap-adjusted tests have correct asymptotic size and are consistent against fixed alternatives.

Extensions to $T > 2$

For the case where $T > 2$, there are two possible approaches. First, one could approach this problem as a multiple testing problem, where the hypothesis would be which two periods in the time series exhibit time homogeneity. In this situation, for the different pairs of time periods, the p-values of the statistics can be computed by the bootstrap procedure given above. A multiple-testing correction, as in Romano and Shaikh (2006), can then be applied to control the family-wise error rate. An alternative approach would be to test that all time periods have the same distribution. In this situation, one can apply the bootstrap procedure above

to a convex combination of the statistics for the different pairs of time periods.

3.4.4 Testing Within-Period Identification

For within-period identification, we have the following setup

$$\begin{aligned} Y_{it} &= \xi_t(X_{it}, \mathcal{A}_i, \mathcal{U}_{it}) \\ F_{\mathcal{A}_i,\mathcal{U}_{it}|X_i}(.,.|\underline{x}) &= F_{\mathcal{A}_i,\mathcal{U}_{it}|h(X_i)}(.,.|h(\underline{x})) \quad \text{for } t = 1, 2, ..., T. \end{aligned} \tag{3.45}$$

By the finiteness of $\mathcal{X}^T$, $h(\underline{x})$ also has finite-support.[32] Let $\mathcal{H}$ denote the support of $h(\underline{x})$ and $|\mathcal{H}| = L$. Let $h_l \in \mathcal{H}$. For $l = 1, ..., L$, define $\mathcal{X}_l \equiv \{\underline{x} \in \mathcal{X}^T : h(\underline{x}) = h_l\}$. Let $k_l = |\mathcal{X}_l|$.

$$\begin{aligned} & H_0^3 : F_{Y_{it}|X_i}(.|\underline{x}_p) = F_{Y_{it}|X_i}(.|\underline{x}_q), \\ & \forall p, q = 1, 2, ..., k_l, l = 1, ..., L, t = 1, 2, ..., T \\ vs. \quad & H_A^3 : F_{Y_{it}|X_i}(.|\underline{x}_p) \neq F_{Y_{it}|X_i}(.|\underline{x}_q), \\ & for some p, q = 1, 2, ..., k_l, l = 1, 2, ..., L, t = 1, 2, ..., T \end{aligned} \tag{3.46}$$

Similar to Quessy and Ethier (2012), we define

$$\bar{F}_{t,l} = \frac{1}{k_l} \sum_{\underline{x} \in \mathcal{X}_l} F_{Y_{it}|X_i}(.|\underline{x}). \tag{3.47}$$

[32] For identification purposes, its support has to be smaller than $\mathcal{X}$, otherwise there would be no identification gain from the restriction as shown in Theorem 3.3.3.

The KS and CM test statistics for the restrictions in each time period are given by

$$\begin{aligned} & KS_{n,t} \\ = & \sum_{l=1}^{L} P(h(X_i) = h_l) \sum_{\underline{x} \in \mathcal{X}_l} P(X_i = \underline{x} | h(X_i) = h_l) \| \sqrt{n} \{ F_{t,n}(.|\underline{x}) - \bar{F}_{t,l,n}(.) \} \|_{\infty, \mathcal{Y}} \end{aligned}$$

$$\begin{aligned} & CM_{n,t} \\ = & \sum_{l=1}^{L} P(h(X_i) = h_l) \sum_{\underline{x} \in \mathcal{X}_l} P(X_i = \underline{x} | h(X_i) = h_l) \| \sqrt{n} \{ F_{t,n}(.|\underline{x}) - \bar{F}_{t,l,n}(.) \} \|_{2, \phi} \end{aligned}$$

Averaging the above statistics over t, we obtain the following,

$$\begin{aligned} KS_{n,\mathcal{Y}} &= \frac{1}{T} \sum_{t=1}^{T} KS_{n,t} \\ CM_{n,\phi} &= \frac{1}{T} \sum_{i=1}^{T} CM_{n,t}. \end{aligned}$$

Now the following bootstrap procedure is used to approximate the distribution of the above statistics.

1. Compute the statistics $KS_{n,\mathcal{Y}}$ and $CM_{n,\phi}$ for $\{\{Y_1, X_1\}, ..., \{Y_n, X_n\}\}$, hereinafter the original sample.

2. Resample n observations $\{\{\hat{Y}_1, \hat{X}_1\}, ..., \{\hat{Y}_n, \hat{X}_n\}\}$ with replacement from the original sample. Compute the centered statistics

$$\begin{aligned} KS^b_{n,\mathcal{Y}} &= \frac{1}{T} \sum_{t=1}^{T} KS^b_{n,t} \\ CM^b_{n,\phi} &= \frac{1}{T} \sum_{t=1}^{T} CM^b_{n,t}, \end{aligned}$$

   where $KS^b_{n,t}$ and $CM^b_{n,t}$ are given below.

3. Repeat 1-2 B times.

4. Calculate the p-values of the tests with

$$\begin{aligned} p_{KS,n} &= \sum_{b=1}^{B} 1\{KS^b_{n,\mathcal{Y}}(\hat{\Lambda}_n) > KS_{n,\mathcal{Y}}(\Lambda_n)\} \\ p_{CM,n} &= \sum_{b=1}^{B} 1\{CM^b_{n,\phi}(\hat{\Lambda}_n) > CM_{n,\phi}(\Lambda_n)\}. \end{aligned}$$

Reject if p-value is smaller than some significance level α.

For event A_i, let $\hat{P}^b_n(A_i)$ be the empirical probability of A_i in the b^{th} bootstrap sample. $KS^b_{n,t}$ and $CM^b_{n,t}$ are given as follows:

$$\begin{aligned} & KS^b_{n,t} \\ = & \sum_{l=1}^{L} \hat{P}^b_n(h(X_i) = h_l) \sum_{\underline{x} \in \mathcal{X}_l} \hat{P}^b_n(X_i = \underline{x} | h(X_i) = h_l) \\ & \times \|\sqrt{n}\{\hat{F}_{1,n}(.|\underline{x}) - \hat{\bar{F}}_{1,l,n}(.|\underline{x}) - (F_{1,n}(.|\underline{x}) - \bar{F}_{l,n}(.|\underline{x}))\}\|_{\infty,\mathcal{Y}} \\ & CM^b_{n,t} \\ = & \sum_{l=1}^{L} \hat{P}^b_n(h(X_i) = h_l) \sum_{\underline{x} \in \mathcal{X}_l} \hat{P}^b_n(X_i = \underline{x} | h(X_i) = h_l) \\ & \times \|\sqrt{n}\{\hat{F}_{t,n}(.|\underline{x}) - \hat{\bar{F}}_{t,l,n}(.|\underline{x}) - (F_{1,n}(.|\underline{x}) - \bar{F}_{l,n}(.|\underline{x}))\}\|_{2,\phi}. \end{aligned}$$

The following theorem shows that the bootstrap-adjusted tests are justified asymptotically.

Theorem 3.4.3 *Given $\{Y_i, X_i\}_{i=1}^n$ is an iid sequence, $|\mathcal{X}| = K$, $P(X_i = \underline{x}) > 0$ for all $\underline{x} \in \mathcal{X}^T$, and $F_{Y_{it}|X_i}(.)$ is nondegenerate for all t, the procedure described in 1-4 for $KS_{n,\mathcal{Y}}$ and $CM_{n,\phi}$ to test H_0^3 (i) provides correct asymptotic size α and (ii) is consistent against fixed alternatives.*

The proof is in Appendix 3.6.2. The convergence of the empirical and bootstrap empirical processes to a tight Brownian bridge follows from results in der Vaart and Wellner (2000). The remainder of the proof follows by similar arguments to Theorem 3.4.2.

3.4.5 Monte Carlo Study

The baseline model is from Evdokimov (2010), but is adapted to have a binary regressor.

$$\begin{aligned} Y_{it} &= m(X_{it}, \mathcal{A}_i) + \mathcal{U}_{it} \\ m(x,a) &= 2a + (2+a)(2x-1)^3 \\ X_{it} &\sim \text{i.i.d.Bernoulli(0.5)}, \\ \mathcal{A}_i &= \frac{\rho}{\sqrt{T}} \sum_{t=1}^{T} \sqrt{12}(X_{it} - 0.5) + \sqrt{1-\rho^2}\psi_i \\ \psi_i &\sim \text{i.i.d.N(0,1)} \\ \epsilon_{it} &\sim N(0,1). \end{aligned}$$

where $\rho = 0.5$.

Note that the above model exhibits time homogeneity without a time effect. Now we also include the following three variants of the above model for our simulations:

(A) Time Homogeneity with no Trend

$\mathcal{U}_{it} = (1 + X_{it})\epsilon_{it}$

$Y_{it} = m(X_{it}, \mathcal{A}_i) + \mathcal{U}_{it}$, $t = 1, 2$

(B) Time Homogeneity up to a Time Effect

$Y_{it} = m(X_{it}, \mathcal{A}_i) + \mathcal{U}_{it} + \lambda_t$, where $\lambda_1 = 0$, $\lambda_2 = 0.5$

(C) Time Homogeneity up to a Generalized Time Effect

$Y_{it} = m(X_{it}, \mathcal{A}_i) + \mathcal{U}_{it} + \lambda_t(X_{it})$, where $\lambda_1(0) = \lambda_2(0) = 0$, $\lambda_2(0) = -0.5$, $\lambda_2(1) = 0.5$

(D) Time Heterogeneity with Exclusion Restriction $F_{\mathcal{A}_i,\mathcal{U}_{it}|X_i}(.) = F_{\mathcal{A}_i,\mathcal{U}_{it}|X_{i1}}(.)$

$\mathcal{A}_i = \rho\sqrt{\frac{12}{T}}(X_{i1} - 0.5) + \sqrt{1-\rho^2}\psi_i$

$\mathcal{U}_{i1} = (1 + X_{i1})\epsilon_{i1}$

$\mathcal{U}_{i2} = (1 + X_{i1})(\epsilon_{i2} + \lambda_2)\sigma_2$, where $\lambda_2 = 0.5$, $\sigma_2 = 1.5$ $Y_{it} = m(X_{it}, \mathcal{A}_i) + \mathcal{U}_{it}$, $t = 1, 2$

Under each model, the behavior of the bootstrap procedure for the following statistics is examined, where we use the notation introduced in sections 3.4.3 and 3.4.4.

$$KS^{nt}_{n,\mathcal{Y}} = \|F_{1,n}(.|\Delta X_i = 0) - F_{2,n}(.|\Delta X_i = 0)\|_{\infty,\mathcal{Y}} \tag{3.48}$$

$$KS^{pt}_{n,\mathcal{Y}} = \|F_{1,n}(.|\Delta X_i = 0) - F_{2,n}(. + \lambda_n|\Delta X_i = 0)\|_{\infty,\mathcal{Y}} \tag{3.49}$$

$$KS^{gt}_{n,\mathcal{Y}} = \|F_{1,n}(.|\Delta X_i = 0) - F_{2,n}(., \Lambda_n|\Delta X_i = 0)\|_{\infty,\mathcal{Y}} \tag{3.50}$$

$$\begin{aligned} KS^{excl}_{n,\mathcal{Y}} &= P_n(X_{i1} = 0)\|F_{1,n}(.|X_i = (0,0)) - F_{1,n}(.|X_i = (0,1))\|_{\infty,\mathcal{Y}} \\ &+ P_n(X_{i1} = 1)\|F_{1,n}(.|X_i = (1,0)) - F_{1,n}(.|X_i = (1,1))\|_{\infty,\mathcal{Y}} \end{aligned} \tag{3.51}$$

$$CM^{nt}_{n,\phi} = \|F_{1,n}(.|\Delta X_i = 0) - F_{2,n}(.|\Delta X_i = 0)\|_{2,\phi} \tag{3.52}$$

$$CM^{pt}_{n,\phi} = \|F_{1,n}(.|\Delta X_i = 0) - F_{2,n}(. + \lambda_n|\Delta X_i = 0)\|_{2,\phi} \tag{3.53}$$

$$CM^{gt}_{n,\phi} = \|F_{1,n}(.|\Delta X_i = 0) - F_{2,n}(., \Lambda|\Delta X_i = 0)\|_{2,\phi} \tag{3.54}$$

$$\begin{aligned} CM^{excl}_{n,\phi} &= P_n(X_{i1} = 0)\|F_{1,n}(.|X_i = (0,0)) - F_{1,n}(.|X_i = (0,1))\|_{2,\phi} \\ &+ P_n(X_{i1} = 1)\|F_{1,n}(.|X_i = (1,0)) - F_{1,n}(.|X_i = (1,1))\|_{2,\phi}. \end{aligned} \tag{3.55}$$

where ϕ is the standard normal density.

Note that the statistics with nt super-script test time homogeneity with no trend, pt time homogeneity with a 'pure' time effect, gt time homogeneity with a generalized time effect, $\lambda_t(X_{it})$, and $excl$ the exclusion restriction $F_{\mathcal{A}_i,\mathcal{U}_{it}|X_i} = F_{\mathcal{A}_i,\mathcal{U}_{it}|X_{i1}}$.

Under the baseline model (A), all variants of time homogeneity are not violated, whereas the exclusion restriction is. Under Model (B), both the exclusion restriction and time homogeneity with no trend are violated. Under Models (C) and (D), only time homogeneity up to a generalized time effect and the exclusion restriction, respectively, are not violated.

Table A.7 reports the coverage probabilities of all of the above statistics under the four different models, when $n = 1000, 1500, 2000$, and $T = 2$, where we perform 1000 simulation replications (S).

The CM statistics follow our asymptotic results in finite samples, since they control size when their respective null is true, and reject with high probability when their respective null does not hold. The KS statistic for the time homogeneity with no trend also performs well in finite samples. However, for time homogeneity with both pure and generalized time effect, the KS statistic is significantly under-sized. However, as n increases, the size properties improve. As for the exclusion restriction, the KS statistic tends to over-reject.[33]

[33]For $n = 3000$, under Model (D), its finite-sample performance matches the asymptotic results, and it exhibits good size control.

3.5 Empirical Illustration: Returns to Schooling

The standard function for estimating returns to schooling is Mincer's human capital earnings function, given by

$$Y = \alpha + \beta S + \gamma E + \delta E^2 + \mathcal{U} \tag{3.56}$$

where Y is log earnings, S is years of completed education, and E is the number of years an individual has worked. Since potential experience is used instead, where $E = Age - S - 6$, many researchers just control for S and Age. Angrist and Newey (1991) examine fixed effects estimation of Mincer's equation using a subsample of the national longitudinal survey of youth (NLSY), since they observe changes in schooling status for 20% of their sample.[34] They also propose a test of the over-identifying restrictions of the linear fixed effects model and reject it for Mincer's equation. In the following, we will test time homogeneity up to a location time effect, which may be viewed as a nonparametric test of the fixed-effects assumption in the presence of time effects. The model is given by the following equation,

$$\begin{aligned} Y_{it} &= \xi(S_{it}, \mathcal{A}_i, \mathcal{U}_{it}) + \lambda_t \\ \mathcal{U}_{it}|S_i, \mathcal{A}_i &\stackrel{d}{=} \mathcal{U}_{i1}|S_i, \mathcal{A}_i. \end{aligned}$$

Our results indicate that we cannot reject the time homogeneity assumption. Hence, we cannot reject the fixed effects assumption nonparametrically. Since our nonparametric estimates of the APE indicate violations of the linear model, we conjecture that the rejection of the over-identification test in Angrist and Newey

[34] Panel data methods are not often used for the identification of returns to schooling, since it is unlikely to observe changes in schooling over the short span of a micro-panel. For instance, Chamberlain (1984) includes schooling as a time-invariant control variable in the union-wage example.

(1991) is due to linearity.

3.5.1 Data and Methodology

Angrist and Newey (1991) use a NLSY random subsample of young men from 1983-1987. The young men are aged 18-26 in 1983. 20% of the subsample exhibits changes in schooling. We use a revised version of this sample with 1087 young men. The descriptive statistics are reported in Table A.8.[35]

The linear specification is the most widely used specification of Mincer's equation. Card (1999) however points out that there is no economic justification for the linear specification and cites empirical findings of possible nonlinearities in the relationship between schooling and earnings. In this spirit, we propose the following model exhibiting time homogeneity up to a location time effect

$$\begin{aligned} Y_{it} &= \xi(S_{it}, \mathcal{A}_i, \mathcal{U}_{it}) + \lambda_t \\ \mathcal{U}_{it}|S_i, \mathcal{A}_i &\stackrel{d}{=} \mathcal{U}_{i1}|S_i, \mathcal{A}_i. \end{aligned}$$

where Y_{it} is log earnings, and we normalize λ_1 to zero. As noted above, this model allows us to identify the effect of schooling on earnings using within-group variation.

The model implies that individuals that do not change their schooling status, i.e. stayers, should have the same distribution for log earnings across time, once

[35] Angrist and Newey (1991) had 1045 in their sample. Despite the difference between their original sample and the revised one used here, the descriptive statistics are quite similar. Hence, this difference cannot be driving the difference in our results. We also replicate their estimation results in Table A.11.

appropriately demeaned. Formally, the testable implication is given as follows

$$F_{Y_{i1}|\Delta S_i}(.|0) \quad = \quad F_{Y_{i2}|\Delta S_i}(. + \lambda_2|0). \tag{3.57}$$

Now we test the above implication for the following year pairs, 1983-84, 1984-85, 1985-86, and 1986-87, using the following KS and CM statistics

$$\begin{aligned} KS_{n,\mathcal{Y}} &= \sup_{y \in \mathcal{Y}} \left| \frac{\sum_{i=1}^n (1\{Y_{i1} \leq y\} - 1\{Y_{i2} - \lambda_n \leq y\}) 1\{\Delta S_i = 0\}}{\sum_{i=1}^n 1\{\Delta S_i = 0\}} \right| \\ CM_{n,\phi} &= \int \left(\frac{\sum_{i=1}^n (1\{Y_{i1} \leq y\} - 1\{Y_{i2} - \lambda_n \leq y\}) 1\{\Delta S_i = 0\}}{\sum_{i=1}^n 1\{\Delta S_i = 0\}} \right)^2 \phi(y) dy, \end{aligned}$$

where $\mathcal{Y}$[36] is the support of Y_{it} and ϕ is the standard normal density, but any other density may also be used. $\lambda_n = \sum_{i=1}^n (Y_{i2} - Y_{i1}) 1\{\Delta S_i = 0\} / \sum_{i=1}^n 1\{\Delta S_i = 0\}$. The p-values for both statistics were obtained using the bootstrap procedure outlined above.

The above model has another implication. All stayer subpopulations regardless of the years of schooling completed must exhibit the same mean shift across time. Formally,

$$E[Y_{i2} - Y_{i1}|S_i = (s,s)] \quad = \quad E[Y_{i2} - Y_{i1}|S_i = (s',s')] \ \ \forall s, s' \in \mathcal{S}, s \neq s', \tag{3.58}$$

where $\mathcal{S}$ denotes the support of S_{it}. This implication can be tested by an F-test, which we also report.

[36] To implement the KS statistic in practice, I search for the maximum over a fine grid between the minimum and maximum value of Y_{it} in the sample.

For every year pair, the APEs for movers are then estimated as follows

$$\begin{aligned} & \hat{\beta}(s \to s+1 | S_i = (s, s+1)) \\ = \; & \frac{\sum_{i=1}^{n} (Y_{i2} - Y_{i1}) 1\{S_i = (s, s+1)\}}{\sum_{i=1}^{n} 1\{S_i = (s, s+1)\}} - \hat{\lambda}_{2,n}, \end{aligned} \tag{3.59}$$

where $\hat{\lambda}_{2,n} = \sum_{i=1}^{n} (Y_{i2} - Y_{i1}) 1\{\Delta S_i = 0\} / \sum_{i=1}^{n} 1\{\Delta S_i = 0\}$. The APE for all movers[37] is given by

$$\begin{aligned} & \hat{\beta}(\Delta S_i = 1) \\ = \; & \sum_{s \in \mathcal{S}} P_n(S_i = (s, s+1) | \Delta S_i = 1) \hat{\beta}(s \to s+1 | S_i = (s, s+1)), \end{aligned} \tag{3.60}$$

The standard errors are computed under the assumption of cross-sectional independence.

3.5.2 Results and Discussion

Table A.9 shows the p-values for the bootstrap-adjusted KS and CM tests as well as the F test for time homogeneity up to a time effect. We use 500 bootstrap replications to adjust the p-value of the KS and CM statistics. The p-values indicate that we cannot reject the fixed-effects assumption for any of the year-pairs. Comparing our findings to Angrist and Newey (1991), we conjecture that the rejection in Angrist and Newey (1991) is due to misspecification of the linear model rather than a violation of the fixed-effects assumption, especially given our estimates of the APE, which we give below.

The estimates of the APE for mover subpopulations are reported in Table

[37] Since schooling only increases by unit increments only, $\Delta S_i = 1$ characterizes all mover subpopulations.

A.10. Similar to Angrist and Newey (1991), our APE estimates are not significant, except in 1985-1986. Our results also provides evidence against the implication of the linear model that the APE for identified subpopulations is constant across time. Furthermore, it is important to note here that the object we are estimating is not a *ceteris paribus* effect. Recall that $E = Age - S - 6$. Since everyone's age increases across time, individuals that increase their schooling are forgoing a year of potential experience, while individuals that do not increase their schooling gain a year of potential experience. Hence, the estimate of the APE here identifies the effect of schooling corrected for the opportunity cost of forgoing one year of potential experience.

Now the significance of the APE in 1985-86 is driven solely by individuals that completed their bachelor degree, i.e. 16 years of schooling. This is in line with the empirical evidence quoted in Card (1999) that the changes in schooling status have particularly significant effects on earnings around the completion of terminal degrees.

3.6 Mathematical Proofs

3.6.1 Proofs of Section 3.3 Results

Proof (Theorem 3.3.1)

Note that

$$\begin{aligned}
& E[Y_{i2}^x - Y_{i1}|X_i = (x, x')] - E[Y_{i2} - Y_{i1}|X_i = (x, x)] \\
= & \int (\xi_2(x, a, u_2) - \xi_1(x, a, u_1)) \\
& \times (dF_{\mathcal{A}_i, \mathcal{U}_{i1}, \mathcal{U}_{i2}|X_i}(a, u_1, u_2|(x, x')) - dF_{\mathcal{A}_i, \mathcal{U}_{i1}, \mathcal{U}_{i2}|X_i}(a, u_1, u_2|(x, x))) \\
= & \int (\xi_2(x, a, u_2) - \xi_1(x, a, u_1)) \\
& \times (f_{\mathcal{A}_i, \mathcal{U}_{i1}, \mathcal{U}_{i2}|X_i}(a, u_1, u_2|(x, x')) - f_{\mathcal{A}_i, \mathcal{U}_{i1}, \mathcal{U}_{i2}|X_i}(a, u_1, u_2|(x, x)))d(a, u_1, u_2)
\end{aligned}$$

Let $\lambda(x) = \xi_2(x, a, u) - \xi_1(x, a, u)$, where $\lambda(x)$ is the component of the difference between the structural function that only depends on the regressors. Now adding and subtracting $\lambda(x)$ to the integrand, we obtain the following,

$$\begin{aligned}
& E[Y_{i2}^x - Y_{i1}|X_i = (x, x')] - E[Y_{i2} - Y_{i1}|X_i = (x, x)] \\
= & \int (\xi_2(x, a, u_2) - \xi_1(x, a, u_1) - \lambda(x)) \\
& \times (f_{\mathcal{A}_i, \mathcal{U}_{i1}, \mathcal{U}_{i2}|X_i}(a, u_1, u_2|(x, x')) - f_{\mathcal{A}_i, \mathcal{U}_{i1}, \mathcal{U}_{i2}|X_i}(a, u_1, u_2|(x, x)))d(a, u_1, u_2) \\
= & \int (\xi_2(x, a, u_2) - \xi_1(x, a, u_1) - \lambda(x)) \\
& \times \left(\frac{f_{\mathcal{A}_i, \mathcal{U}_{i1}, \mathcal{U}_{i2}|X_i}(a, u_1, u_2|(x, x'))}{f_{\mathcal{A}_i, \mathcal{U}_{i1}, \mathcal{U}_{i2}|X_i}(a, u_1, u_2|(x, x))} - 1 \right) f_{\mathcal{A}_i, \mathcal{U}_{i1}, \mathcal{U}_{i2}|X_i}(a, u_1, u_2|(x, x))d(a, u_1, u_2).
\end{aligned}$$

The above is equal to zero if

$$(\xi_2(x, a, u_2) - \xi_1(x, a, u_1) - \lambda(x)) \left(\frac{f_{\mathcal{A}_i, \mathcal{U}_{i1}, \mathcal{U}_{i2}|X_i}(a, u_1, u_2|(x, x'))}{f_{\mathcal{A}_i, \mathcal{U}_{i1}, \mathcal{U}_{i2}|X_i}(a, u_1, u_2|(x, x))} - 1 \right) = 0.$$

$\forall a, u_1, u_2$. □

Proof (Variation 3.3.1)

Under time homogeneity, the condition in Theorem 3.3.1 simplifies to

$$\begin{aligned} & (\xi_2(x,a,u) - \xi_1(x,a,u) - \lambda(x)) \\ \times \quad & \left(f_{\mathcal{A}_i,\mathcal{U}_{i1}|X_i}(a,u|(x,x')) - f_{\mathcal{A}_i,\mathcal{U}_{i1}|X_i}(a,u|(x,x))\right) = 0, \end{aligned}$$

$\forall a, u$. Under unrestricted individual heterogeneity, the second term is not equal to zero. Hence, the above is zero if

$$\xi_2(x,a,u) - \xi_1(x,a,u) \quad = \quad \lambda(x)\, \forall a, u$$

Without loss of generality, we can write the above as $\xi_t(x,a,u) = \xi(x,a,u) + \lambda_t(x)$. □

Proof (Variation 3.3.2)

Recall that $E[Y_{i2}^x - Y_{i1}|X_i = (x,x')] = E[Y_{i2} - Y_{i1}|X_i = (x,x)]$ can be written as

$$\begin{aligned} & \int (\xi_2(x,a,u_2) - \xi_1(x,a,u_1) - \lambda(x)) \\ & \times \left(f_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a,u_1,u_2|(x,x')) - f_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a,u_1,u_2|(x,x))\right) d(a,u_1,u_2) = 0. \end{aligned}$$

By the conditions of this variation, $\xi_2(x,a,u_2) = \xi_1(x,a,u_2) + \lambda(x)$. Plugging this into the integrand of the previous equation, it follows that

$$\begin{aligned} & \int (\xi_1(x,a,u_2) - \xi_1(x,a,u_1)) \\ & \times (f_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a,u_1,u_2|(x,x')) - f_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a,u_1,u_2|(x,x))) = 0 \end{aligned}$$

Under unrestricted individual heterogeneity and $f_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(.) > 0$ by Assumption

3.3.2, the above equals to zero if each term is zero. Hence, for $\underline{x} \in \{(x,x),(x,x')\}$,

$$\int (\xi_1(x,a,u_2) f_{\mathcal{A}_i,\mathcal{U}_{i2}|X_i}(a,u_2|\underline{x}) - \xi_1(x,a,u_1) f_{\mathcal{A}_i,\mathcal{U}_{i1}|X_i}(a,u_1|\underline{x}))d(a,u_1,u_2) \quad = \quad 0$$

The above holds if

$$\xi_1(x,a,u) f_{\mathcal{A}_i,\mathcal{U}_{i2}|X_i}(a,u|\underline{x}) \quad = \quad \xi_1(x,a,u) f_{\mathcal{A}_i,\mathcal{U}_{i1}|X_i}(a,u|\underline{x}) \ \ \forall a,u$$

which holds if $f_{\mathcal{A}_i,\mathcal{U}_{i1}|X_i}(a,u|\underline{x}) = f_{\mathcal{A}_i,\mathcal{U}_{i2}|X_i}(a,u|\underline{x})\ \forall a,u$, which implies time homogeneity, i.e. $\mathcal{U}_{i1}|X_i,\mathcal{A}_i \stackrel{d}{=} \mathcal{U}_{i2}|X_i,\mathcal{A}_i$. □

Proof (Variation 3.3.3)

Recall the condition in Theorem 3.3.1

$$\begin{aligned}&(\xi_2(x,a,u_2) - \xi_1(x,a,u_1) - \lambda(x)) \\ &\times \left(f_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a,u_1,u_2|(x,x')) - f_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a,u_1,u_2|(x,x))\right) = 0.\end{aligned}$$

Clearly, without restrictions on time heterogeneity and the structural function, the above is true if

$$f_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a,u_1,u_2|(x,x')) \quad = \quad f_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a,u_1,u_2|(x,x)) \ \ \forall a,u_1,u_2.$$

□

Proof (Variation 3.3.4) For notational convenience, we set $\xi_1(x,a,u) = \xi(x,a,u)$.

By Assumption 3.3.1(iii), Fubini's theorem applies as follows,

$$\begin{aligned} & \int (\xi(x,a,u_2) - \xi(x,a,u_1) - \lambda(x)) \\ & \times d(F_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a,u_1,u_2|(x,x')) - F_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a,u_1,u_2|(x,x))) \\ = & \int (m_2(x,a,X_i) - m_1(x,a,X_i) - \lambda(x)) \\ & \times d(F_{\mathcal{A}_i|X_i}(a|(x,x')) - F_{\mathcal{A}_i|X_i}(a|(x,x))), \end{aligned}$$

where $m_t(x,a,\underline{x}) = \int \xi(x,a,u) dF_{\mathcal{U}_{it}|X_i,\mathcal{A}_i}(u|\underline{x},a)$.

Under arbitrary individual heterogeneity and $f_{\mathcal{A}_i|X_i}(.|.) > 0$, the above is zero if

$$(m_2(x,a,\underline{x}) - m_1(x,a,\underline{x}) - \lambda(x)) = 0 \; \forall a, \underline{x}, \tag{3.61}$$

By Assumption 3.3.3 (i), (ii) and the dominated convergence theorem, $\partial(m_2(x,a,\underline{x}) - m_1(x,a,\underline{x}))/\partial a$ exists and is continuous. Now note that

$$\begin{aligned} & m_2(x',a',\underline{x}') - m_1(x',a',\underline{x}') - (m_2(x,a,\underline{x}) - m_1(x,a,\underline{x})) \\ = \; & m_2(x',a',\underline{x}') - m_1(x',a',\underline{x}') - (m_2(x',a',\underline{x}) - m_1(x',a',\underline{x})) \\ + \; & m_2(x',a',\underline{x}) - m_1(x',a',\underline{x}) - (m_2(x',a,\underline{x}) - m_1(x',a,\underline{x})) \\ + \; & m_2(x',a,\underline{x}) - m_1(x',a,\underline{x}) - (m_2(x,a,\underline{x}) - m_1(x,a,\underline{x})) \end{aligned}$$

$$\begin{aligned} = \; & m_2(x',a',\underline{x}') - m_1(x',a',\underline{x}') - (m_2(x',a',\underline{x}) - m_1(x',a',\underline{x})) \\ + \; & \int_a^{a'} \frac{\partial(m_2(x',a,\underline{x}) - m_1(x',a,\underline{x}))}{\partial a} da \\ + \; & m_2(x',a,\underline{x}) - m_1(x',a,\underline{x}) - (m_2(x,a,\underline{x}) - m_1(x,a,\underline{x})) \\ & \forall x, x', a, a', \underline{x}, \underline{x}' \end{aligned} \tag{3.62}$$

where the last equality follows by the first fundamental theorem of calculus and the continuity of $\partial(m_2(x', a, \underline{x}) - m_1(x', a, \underline{x}))/\partial a$. Now if the following conditions hold

$$\frac{\partial(m_2(x, a, \underline{x}) - m_1(x, a, \underline{x}))}{\partial a} = 0 \quad \forall a, \underline{x} \tag{3.63}$$

$$\frac{\bar{\partial}(m_2(x, a, \underline{x}) - m_1(x, a, \underline{x}))}{\bar{\partial}(\underline{x}, \underline{x}')} = 0 \quad \forall a, \underline{x}, \underline{x}' \tag{3.64}$$

where $\bar{\partial} g(x)/\bar{\partial}(x, x') = g(x) - g(x')$, (3.62) yields the following

$$\begin{aligned} m_2(x', a', \underline{x}') - m_1(x', a', \underline{x}') &= m_2(x', a, \underline{x}) - m_1(x', a, \underline{x}) \\ &\quad \forall x, x', a, a', \underline{x}, \underline{x}' \end{aligned} \tag{3.65}$$

since $a \neq a'$ and $\underline{x} \neq \underline{x}'$ for some a' and $\underline{x}'$, (3.65) implies (3.61) by the arbitrariness of $\lambda(x)$.

Note that we can re-write (3.63) as follows

$$\begin{aligned} &\frac{\partial(m_2(x, a, \underline{x}) - m_1(x, a, \underline{x}))}{\partial a} \\ = &\int (\xi(x, a, u_2) - \xi(x, a, u_1)) dF_{\mathcal{U}_{i1}, \mathcal{U}_{i2}|X_i, \mathcal{A}_i}(u_1, u_2|\underline{x}, a) \end{aligned}$$

By Assumption 3.3.3(ii) and the dominated convergence theorem,

$$\begin{aligned}
&\frac{\partial(m_2(x,a,\underline{x})-m_1(x,a,\underline{x}))}{\partial a}\\
=&\int\frac{\partial(\xi(x,a,u_2)-\xi(x,a,u_1))f_{\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i,\mathcal{A}_i}(u_1,u_2|\underline{x},a))}{\partial a}d(u_1,u_2),\\
=&\int\frac{\partial(\xi(x,a,u_2)-\xi(x,a,u_1))}{\partial a}f_{\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i,\mathcal{A}_i}(u_1,u_2|\underline{x},a)d(u_1,u_2)\\
+&\int(\xi(x,a,u_2)-\xi(x,a,u_1))f^a_{\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i,\mathcal{A}_i}(u_1,u_2|\underline{x},a)d(u_1,u_2)\\
=&\int\frac{\partial(\xi(x,a,u_2)-\xi(x,a,u_1))}{\partial a}f_{\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i,\mathcal{A}_i}(u_1,u_2|\underline{x},a)d(u_1,u_2)\\
+&\int(\xi(x,a,u_2)-\xi(x,a,u_1))\frac{f^a_{\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i,\mathcal{A}_i}(u_1,u_2|\underline{x},a)}{f_{\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i,\mathcal{A}_i}(u_1,u_2|\underline{x},a)}\\
&\times f_{\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i,\mathcal{A}_i}(u_1,u_2|\underline{x},a)d(u_1,u_2) \qquad (3.66)
\end{aligned}$$

Now the above equals to zero if

$$\begin{aligned}
&\frac{\partial(\xi(x,a,u_2)-\xi(x,a,u_1))}{\partial a}f_{\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i,\mathcal{A}_i}(u_1,u_2|\underline{x},a)\\
+&(\xi(x,a,u_2)-\xi(x,a,u_1))f^a_{\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i,\mathcal{A}_i}(u_1,u_2|\underline{x},a)=0 \quad \forall a,u_1,u_2,\underline{x}
\end{aligned}$$

The above is true if the following conditions hold

$$\frac{\partial(\xi(x,a,u_2)-\xi(x,a,u_1))}{\partial a} = 0\ \forall a,u_1,u_2 \qquad (3.67)$$

$$f^a_{\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i,\mathcal{A}_i}(u_1,u_2|\underline{x},a) = 0\ \forall a,u_1,u_2,\underline{x}. \qquad (3.68)$$

As for (3.64), let $\bar{f}^z(z)=\bar{\partial}f(z)/\bar{\partial}(z,z')$.

$$\begin{aligned}
&\frac{\bar{\partial}(m_2(x,a,\underline{x})-m_1(x,a,\underline{x}))}{\bar{\partial}\underline{x}}\\
=&\int(\xi(x,a,u_2)-\xi(x,a,u_1))\bar{f}^{\underline{x}}_{\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i,\mathcal{A}_i}(u_1,u_2|\underline{x},a)d(u_1,u_2)
\end{aligned}$$

The above equals to zero if

$$(\xi(x,a,u_2)-\xi(x,a,u_1))\bar{f}^{\underline{x}}_{\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i,\mathcal{A}_i}(u_1,u_2|\underline{x},a) \;=\; 0,\;\; \forall a,u_1,u_2,\underline{x} \quad (3.69)$$

since $(\xi(x,a,u_2)-\xi(x,a,u_1)) \neq 0$ for some a,u_1,u_2, it follows that the above is true iff

$$\bar{f}^{\underline{x}}_{\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i,\mathcal{A}_i}(u_2|\underline{x},a) \;=\; 0 \quad (3.70)$$

We obtain (i) by noting that (3.67) is true if

$$\partial\xi^a(x,a,u)/\partial u \;=\; 0.\;\; \forall a,u$$

Now (3.68) and (3.70) are true if $f_{\mathcal{U}_{it}|X_i,\mathcal{A}_i}(.) = f_{\mathcal{U}_{it}}(.)$, which is equivalent to (ii) by Assumption 3.3.3(i). □

Proof (Theorem 3.3.2)

By (iii) $\underline{x}^c_t = \underline{x}^c_\tau = x$,

$$\begin{aligned} E[Y_{i\tau}-Y_{it}|X_i=\underline{x}^c] &\overset{(i)}{=} \int (\xi_\tau(x,a,u)-\xi_t(x,a,u))dF_{\mathcal{A}_i,\mathcal{U}_{it}|X_i}(a,u|\underline{x}^c) \\ &\overset{(ii)}{=} \lambda_\tau(h(\underline{x}^c)) \end{aligned} \quad (3.71)$$

Note that $h(\underline{x}) = h(\underline{x}^c)$.

$$\begin{aligned} E[Y_{i\tau}-Y_{it}|X_i=\underline{x}] &\overset{(i)}{=} \int (\xi_\tau(\underline{x}_\tau,a,u)-\xi_t(\underline{x}_t,a,u))dF_{\mathcal{A}_i,\mathcal{U}_{it}|X_i}(a,u|\underline{x}) \\ &= \int (\xi_t(\underline{x}_\tau,a,u)-\xi_t(\underline{x}_t,a,u))dF_{\mathcal{A}_i,\mathcal{U}_{it}|X_i}(a,u|\underline{x}) + \lambda_\tau(h(\underline{x})) \\ &= \beta_t(\underline{x}_t \to \underline{x}_\tau|X_i=\underline{x}) + \lambda_\tau(h(\underline{x})) \end{aligned}$$

Thus, by (iii)

$$\beta_t(\underline{x}_t \to \underline{x}_\tau | X_i = \underline{x}) \;\; = \;\; E[Y_{i\tau} - Y_{it} | X_i = \underline{x}] - E[Y_{i\tau} - Y_{it} | X_i = \underline{x}^c]$$

□

Proof (Theorem 3.3.3)

$$\begin{aligned}
& E[Y_{it}|X_i = \underline{x}] - E[Y_{it}|X_i = \underline{x}^c] \\
= \;& \int \xi_t(\underline{x}_t, a, u) dF_{\mathcal{A}_i, \mathcal{U}_{it}|X_i}(a, u|\underline{x}) - \int \xi_t(\underline{x}_t^c, a, u) dF_{\mathcal{A}_i, \mathcal{U}_{it}|X_i}(a, u|\underline{x}^c) \\
\overset{\text{(i) \& (ii)}}{=} \;& \int (\xi_t(\underline{x}_t, a, u) - \xi_t(\underline{x}_t^c, a, u)) dF_{\mathcal{A}_i, \mathcal{U}_{it}|h(X_i)}(a, u|h(\underline{x})) \\
= \;& \beta_t(\underline{x}_t^c \to \underline{x}_t | X_i = \underline{x}) \\
= \;& \beta_t(\underline{x}_t^c \to \underline{x}_t | X_i = \underline{x}^c)
\end{aligned}$$

□

3.6.2 Proofs of Section 3.4 Results

Proof (Theorem 3.4.1)

In order to show (i) and (ii), we have to show that the following empirical and bootstrap empirical processes converge to the same tight Brownian bridge. Let $\mathbb{G}_{n|\Delta X_i=0}$ and $\hat{\mathbb{G}}_{n|\Delta X_i=0}$ denote the empirical and bootstrap empirical processes,

respectively.

$$\begin{aligned}
& \mathbb{G}_{n|\Delta X_i=0} \\
= \; & \sqrt{n}(F_{1,n}(.|\Delta X_i = 0) - F_1(.|\Delta X_i = 0)) \\
- \; & \sqrt{n}(F_{2,n}(.|\Delta X_i = 0) - F_2(.|\Delta X_i = 0)) \\
& \hat{\mathbb{G}}_{n|\Delta X_i=0} \\
= \; & \sqrt{n}(\hat{F}_{1,n}(.|\Delta X_i = 0) - F_{1,n}(.|\Delta X_i = 0)) \\
& -\sqrt{n}(\hat{F}_{2,n}(.|\Delta X_i = 0) - F_{2,n}(.|\Delta X_i = 0))),
\end{aligned}$$

where $\hat{F}_{t,n} = \sum_{i=1}^{n} M_{ni} 1\{Y_{it} \leq y\} 1\{\Delta X_i = 0\} / \sum_{i=1}^{n} M_{ni} 1\{\Delta X_i = 0\}$.

Since the unconditional cdfs trivially fulfill the Donsker property, it follows that they converge to a tight Brownian bridge

$$\sqrt{n} \begin{pmatrix} F_{1,n}(.) - F_1(.) \\ F_{2,n}(.) - F_2(.) \end{pmatrix} \rightsquigarrow \begin{pmatrix} \mathbb{G}_1 \\ \mathbb{G}_2 \end{pmatrix}, \tag{3.72}$$

where $\mathbb{G}_1$ and $\mathbb{G}_2$ are each a tight Brownian bridge on $L^\infty(\mathcal{Y})$.

Since $P(\Delta X_i = 0) > 0$ and $\{Y_i, X_i\}_{i=1}^n$ is an i.i.d. sequence, Lemma 3.6.3 applies,

which implies that

$$\begin{aligned}
& \sqrt{n}\begin{pmatrix} F_{1,n}(.|\Delta X_i) - F_1(.|\Delta X_i) \\ F_{2,n}(.|\Delta X_i = 0) - F_2(.|\Delta X_i = 0) \end{pmatrix} \\
\rightsquigarrow & \begin{pmatrix} \frac{\mathbb{G}_1}{P(\Delta X_i = 0)} + \mathcal{Z}F_1(., \Delta X_i = 0) \\ \frac{\mathbb{G}_2}{P(\Delta X_i = 0)} + \mathcal{Z}F_2(., \Delta X_i = 0) \end{pmatrix} \\
\equiv & \begin{pmatrix} \mathbb{G}_{1|\Delta X=0} \\ \mathbb{G}_{2|\Delta X=0} \end{pmatrix} \\
& \sqrt{n}\begin{pmatrix} \hat{F}_{1,n}(.|\Delta X_i) - F_{1,n}(.|\Delta X_i) \\ \hat{F}_{2,n}(.|\Delta X_i = 0) - F_{2,n}(.|\Delta X_i = 0) \end{pmatrix} \\
\rightsquigarrow & \begin{pmatrix} \mathbb{G}_{1|\Delta X=0} \\ \mathbb{G}_{2|\Delta X=0} \end{pmatrix}
\end{aligned}$$

where $\sqrt{n}(P_n(\Delta X_i = 0) - P(\Delta X_i = 0)) \rightsquigarrow \mathcal{Z}$, where $P_n(\Delta X_i = 0) = \sum_{i=1}^{n} 1\{\Delta X_i = 0\}/n$. By the tightness of $\mathbb{G}_t$ for $t = 1, 2$ and $\mathcal{Z}$, it follows that $\mathbb{G}_{t|\Delta X=0}$ is a tight Brownian bridge for $t = 1, 2$. By continuous mapping theorem, it follows that

$$\begin{aligned}
\mathbb{G}_{n|\Delta X_i=0} & \rightsquigarrow \frac{\mathbb{G}_1 - \mathbb{G}_2}{P(\Delta X_i = 0)} + \mathcal{Z}(F_1(., \Delta X_i = 0) - F_2(., \Delta X_i = 0)) \equiv \mathbb{H} \\
\hat{\mathbb{G}}_{n|\Delta X_i=0} & \rightsquigarrow \mathbb{H}.
\end{aligned}$$

where $(\mathbb{G}_1 - \mathbb{G}_2)$ is a tight Brownian bridge in $\mathcal{L}^\infty(\mathcal{F})$, where $\mathcal{F} = \{1\{Y_{i1} \leq y\} - 1\{Y_{i2} \leq y\} : y \in \mathcal{Y}\}$. Thus, $\mathbb{H}$ is a tight Brownian bridge.

Since $\|.\|_{\infty,\mathcal{Y}}$ and $\|.\|_{2,\phi}$ are continuous, convex functionals, it follows that $\|\mathbb{H}\|_{\infty,\mathcal{Y}}$ and $\|\mathbb{H}\|_{2,\phi}$ has absolutely continuous and strictly increasing distribution on its support $[0, \infty)$, except possibly at zero, by Theorem 11.1 in Davydov, Lifshits, and Smorodina (1998). Since $F_1(.|\Delta X_i = 0)$ and $F_2(.|\Delta X_i = 0)$ are non-degenerate,

then $P(\|\mathbb{H}\|_{\infty,\mathcal{Y}} = 0) = 0$ and $P(\|\mathbb{H}\|_{2,\phi} = 0) = 0$. Thus, both norms of $\mathbb{H}$ have absolutely continuous distributions on $[0, \infty)$. Now the critical values of the bootstrap-adjusted KS and CM tests are given by

$$\begin{aligned} \hat{c}_n^{KS} &= \inf\{t : \hat{P}_n(\|\hat{\mathbb{G}}_{n|\Delta X_i=0}\|_{\infty,\mathcal{Y}} > t) \leq \alpha\}, \\ \hat{c}_n^{CM} &= \inf\{t : \hat{P}_n(\|\hat{\mathbb{G}}_{n|\Delta X_i=0}\|_{2,\phi} > t) \leq \alpha\}, \end{aligned}$$

where $\hat{P}_n$ is the bootstrap probability measure for the sample. Given the above, it follows that

$$\begin{aligned} \hat{c}_n^{KS} \xrightarrow{p} c^{KS} &= \inf\{t : P(\|\mathbb{H}\|_{\infty,\mathcal{Y}} > t) \leq \alpha\}, \\ \hat{c}_n^{CM} \xrightarrow{p} \tilde{c}^{CM} &= \inf\{t : P(\|\mathbb{H}\|_{2,\phi} > t) \leq \alpha\}. \end{aligned}$$

Thus, under the null, the bootstrap-adjusted statistics have correct asymptotic size. Hence, we have shown (i). By the tightness of the limit process, it follows that $\hat{c}_n^{KS}$ and $\hat{c}_n^{CM}$ are bounded in probability. Thus, the tests are consistent against any fixed alternative, which proves (ii). □

Proof (Theorem 3.4.2)

In order to show (i) and (ii), we first have to show that the underlying empirical and bootstrap empirical processes, given below, converge to the same tight Brownian bridge. Once this is established, the proofs of (i) and (ii) follow by similar arguments to Theorem 3.4.1.

We define the empirical and bootstrap empirical processes, $\mathbb{G}_{n|\Delta X_i=0}(\Lambda_n)$ and

$\hat{\mathbb{G}}_{n|\Delta X_i=0}(\hat{\Lambda}_n)$

$$\begin{aligned}
\mathbb{G}_{n|\Delta X_i=0}(\Lambda_n) &= \sqrt{n}(F_{1,n}(.|\Delta X_i=0) - F_1(.|\Delta X_i=0)) \\
&\quad -\sqrt{n}(F_{2,n}(.,\Lambda_n|\Delta X_i=0) - F_2(.,\Lambda|\Delta X_i=0)) \\
\hat{\mathbb{G}}_{n|\Delta X_i=0}(\hat{\Lambda}_n) &= \sqrt{n}(\hat{F}_{1,n}(.|\Delta X_i=0) - F_{1,n}(.|\Delta X_i=0)) \\
&\quad -\sqrt{n}(\hat{F}_{2,n}(.,\hat{\Lambda}_n|\Delta X_i=0) - F_{2,n}(.,\Lambda_n|\Delta X_i=0))
\end{aligned}$$

Now note that

$$\begin{aligned}
F_{2,n}(.|\Delta X_i=0) &= \sum_{k=1}^{K} P_n(X_{i1}=X_{i2}=x^k) F_{2,n}(.+\lambda(x)|X_{i1}=X_{i2}=x^k) \\
F_2(.|\Delta X_i=0) &= \sum_{k=1}^{K} P(X_{i1}=X_{i2}=x^k) F_2(.+\lambda(x)|X_{i1}=X_{i2}=x^k) \quad (3.73)
\end{aligned}$$

Since Assumption 3.4.1 holds, it follows that, for $x \in \mathcal{X}$, $F_2(.+\lambda(x)|X_{i1}=X_{i2}=x)$ is Hadamard differentiable tangentially to $\mathbb{D}\times\mathcal{Y}$ by Lemma 3.7.1, where $\mathbb{D} = \{g \in \mathcal{L}^\infty(\mathcal{F}) : \text{g is } \rho_2\text{-uniformly continuous}\}$, where ρ_2 is the variance metric. The Hadamard derivative is given by $\phi'_{F_2(.|x),\lambda(x)}(g,\epsilon) = g(.+\lambda(x)) + \epsilon f_2(.+\lambda(x)|X_{i1}=X_{i2}=x)$, where the subscript $F_2(.|x)$ denotes $F_2(.|X_{i1}=X_{i2}=x)$. Now $F_1(.|X_{i1}=X_{i2}=x) - F_2(.+\lambda(x)|X_{i1}=X_{i2}=x)$ is trivially Hadamard differentiable tangentially to $\mathbb{D}\times\mathcal{L}^\infty(\mathcal{Y})\times\mathcal{Y}$.

Let $F_{t|x}(.) = F_t(.|X_{i1}=X_{i2}=x)$ and $F_{t|x,n}(.)$ be the sample analogue thereof. Now noting that

$$\sqrt{n}\begin{pmatrix} F_{1|.,n}(.) - F_{1|.}(.) \\ F_{2|.,n}(.) - F_{2|.}(.) \\ \Lambda_n - \Lambda \end{pmatrix} \rightsquigarrow \begin{pmatrix} \mathbb{G}_{1|.} \\ \mathbb{G}_{2|.} \\ \mathcal{E} \end{pmatrix} \quad (3.74)$$

where $\mathbb{G}_{1|.}$ and $\mathbb{G}_{2|.}$ are each $K \times 1$ tight Brownian bridges on $\{\mathcal{L}^\infty(\mathcal{Y})\}^K$ and $\mathcal{E}$ is a K-dimensional normal random vector. For the following, we define $\mathcal{E}(x)$ as follows, $\sqrt{n}(\lambda_n(x) - \lambda(x)) \xrightarrow{p} \mathcal{E}(x)$.

By Theorem 3.9.4 in der Vaart and Wellner (2000), it follows that

$$\sqrt{n}\left(F_{1|.,n}(.) - F_{2|.,n}(. + \lambda_n(.)) - (F_{1|.}(.) - F_{2|.}(. + \lambda(.)))\right) \mapsto \mathbb{G}_{1,2|.}, \qquad (3.75)$$

where $\mathbb{G}_{1,2|x} = \mathbb{G}_{1|x} - \phi'_{F_{2|x},\lambda(x)}(\mathbb{G}_{2|x}, \mathcal{E}(x))$, which is a tight Brownian bridge.

To show the weak convergence of the bootstrap empirical process, given below, we have to check that the conditions in Theorem 3.9.11 in der Vaart and Wellner (2000),

$$\sqrt{n}\left(\hat{F}_{1|.,n}(.) - \hat{F}_{2|.,n}(. + \hat{\lambda}_n(.)) - (F_{1|.,n}(.) - F_{2|.,n}(. + \lambda_n(.)))\right) \mapsto \mathbb{G}_{1,2|.}. \quad (3.76)$$

The conditions in Theorem 3.9.11 include (a) Hadamard-differentiability tangentially to a subspace $\mathbb{D} \times \mathcal{L}^\infty(\mathcal{Y}) \times \mathcal{Y}$, (b) the underlying empirical processes converge to a separable limit, and (c) Condition (3.9.9), p. 378, in Van der Vaart and Wellner holds in outer probability. (a) follows from the above. Now (b) follows by (3.74) and tightness, since the latter implies separability. Finally, (c) is fulfilled if the conditions of Theorem 3.6.2 hold. We can consider $(\mathbb{G}_{1|.}, \mathbb{G}_{2|.}, \mathcal{E})$ as a tight Brownian bridge on $\mathcal{L}^\infty(\mathcal{F}) \times \mathcal{L}^\infty(\mathcal{F}) \times \mathcal{Y}$, where $\mathcal{F} = \{1\{y \leq t\} : t \in \mathcal{Y}\}$. Note that $\mathcal{E}$ is finite-dimensional, it suffices to show that Theorem 3.6.2 applies to $\mathcal{F}$. Since $\mathcal{F}$ is clearly Donsker and $\sup_{f \in \mathcal{F}} | \int (f - \int f dF_t(y|X_{i1} = X_{i2} = x))^2 dF_t(y|X_{i1} = X_{i2} = x)| < \infty$, the conditions in Theorem 3.6.2. hold. Thus, (c) holds, which implies (3.76).

Now we relate (3.75) and (3.76) to $\mathbb{G}_{n|\Delta X_i=0}$ and $\hat{\mathbb{G}}_{n|\Delta X_i=0}$, respectively. Let

$F_{1|x,n}(.) \equiv F_{1,n}(.|X_{i1} = X_{i2} = x)$.

$$
\begin{aligned}
& \mathbb{G}_{n|\Delta X_i=0} \\
= \quad & \sqrt{n}(F_{1,n}(.|\Delta X_i) - F_{2,n}(., \Lambda_n|\Delta X_i = 0)) \\
- \quad & \sqrt{n}(F_1(.|\Delta X_i = 0) - F_2(., \Lambda|\Delta X_i = 0)) \\
= \quad & \sqrt{n}\sum_{k=1}^{K} P_n(X_{i1} = X_{i2} = x^k|\Delta X_i = 0)\left(F_{1|x^k,n}(.) - F_{2|x^k,n}(. + \lambda_n(x^k))\right) \\
& -\sqrt{n}\sum_{k=1}^{K} P_n(X_{i1} = X_{i2} = x^k|\Delta X_i = 0)\left(F_{1|x^k}(.) - F_{2|x^k}(. + \lambda(x^k))\right) \\
= \quad & \sqrt{n}\sum_{k=1}^{K} P(X_{i1} = X_{i2} = x^k|\Delta X_i = 0)\left(F_{1|x^k,n}(.) - F_{2|x^k,n}(. + \lambda_n(x^k))\right) \\
& -\sqrt{n}\sum_{k=1}^{K} P(X_{i1} = X_{i2} = x^k|\Delta X_i = 0)\left(F_{1|x^k}(.) - F_{2|x^k}(. + \lambda(x^k))\right) \\
+ \quad & \sum_{k=1}^{K}(P_n(X_{i1} = X_{i2} = x^k|\Delta X_i = 0) - P(X_{i1} = X_{i2} = x^k|\Delta X_i = 0)) \\
& \times\sqrt{n}\left(F_{1|x^k,n}(.) - F_{2|x^k,n}(. + \lambda_n(x^k)) - (F_{1|x^k}(.) - F_{2|x^k}(. + \lambda(x^k)))\right) \quad (3.77) \\
\rightsquigarrow \quad & \sum_{k=1}^{K} P(X_{i1} = X_{i2} = x^k|\Delta X_i = 0)(\mathbb{G}_{1|x^k} - \phi'_{F_2,\lambda(x^k)}(\mathbb{G}_{2|x^k}, \mathcal{E}(x^k))) \\
\equiv \quad & \mathbb{H}(\Lambda) \quad (3.78)
\end{aligned}
$$

since the last term of the last equality converges to zero in probability by the weak convergence of (3.77) and $P_n(X_{i1} = X_{i2} = x^k|\Delta X_i = 0) \xrightarrow{p} P(X_{i1} = X_{i2} = x^k|\Delta X_i = 0)$ for all $k = 1, 2, ..., K$. Hence, we have shown that the empirical process, $\mathbb{G}_{n|\Delta X_i=0}$ converges to a tight Brownian bridge.

Now the bootstrap empirical process can be decomposed as follows

$$
\begin{aligned}
& \hat{\mathbb{G}}_{n|\Delta X_i=0} \\
= \;& \sqrt{n}\left(\hat{F}_{1,n}(.|\Delta X_i=0) - \hat{F}_{2,n}(.,\hat{\Lambda}_n|\Delta X_i=0)\right) \\
- \;& \sqrt{n}\left(F_{1,n}(.|\Delta X_i=0) - F_{2,n}(.,\Lambda_n|\Delta X_i=0)\right) \\
= \;& \sqrt{n}\sum_{k=1}^{K}\hat{P}_n(X_{i1}=X_{i2}=x^k|\Delta X_i=0) \\
& \times\left(\hat{F}_{1|x^k,n}(.+\hat{\lambda}_n(x^k)) - \hat{F}_{2|x^k,n}(.) - (F_{1|x^k,n}(.+\lambda_n(x^k)) - F_{2|x^k,n}(.))\right) \\
= \;& \sqrt{n}\sum_{k=1}^{K}P(X_{i1}=X_{i2}=x^k|\Delta X_i=0) \\
& \times\left(\hat{F}_{1|x^k,n}(.) - \hat{F}_{2|x^k,n}(.+\hat{\lambda}_n(x^k)) - (F_{1|x^k,n}(.) - F_{2|x^k,n}(.+\lambda_n(x^k)))\right) \\
+ \;& \sum_{k=1}^{K}(\hat{P}_n(X_{i1}=X_{i2}=x^k|\Delta X_i=0) - P(X_{i1}=X_{i2}=x^k|\Delta X_i=0)) \\
& \times\sqrt{n}\left(\hat{F}_{1|x^k,n}(.) - \hat{F}_{2|x^k,n}(.+\hat{\lambda}_n(x^k)) - (F_{1|x^k,n}(.) - F_{2|x^k,n}(.+\lambda_n(x^k)))\right) \\
\rightsquigarrow \;& \mathbb{H}(\Lambda)
\end{aligned}
\tag{3.79}
$$

where the first term of the last equality follows by the continuous mapping theorem and (3.76). The second term converges to zero by (3.76) and $(\hat{P}_n(X_{i1}=X_{i2}=x^k|\Delta X_i=0) - P(X_{i1}=X_{i2}=x^k|\Delta X_i=0)) \xrightarrow{p} 0$ by Lemma 3.6.1. Thus, the bootstrap empirical process converges to the same tight Brownian bridge as the empirical process. (i) and (ii) follow by the same arguments as Theorem 3.4.1. □

Proof (Theorem 3.4.3)

We first have to show that the underlying empirical and bootstrap empirical processes converge to the same tight Brownian bridge. Let $m_l = |\{\underline{x} \in \mathcal{X}_l : \underline{x} \in \mathcal{X}\}|$, where $|S|$ denotes the cardinality of a set S. Our statistics can be written as

follows:

$$\begin{aligned} & KS_{n,\mathcal{Y}} \\ = \; & \frac{1}{T}\sum_{t=1}^{T}\sum_{l=1}^{L} P_n(h(X_i)=h_l)\sum_{j=1}^{m_l} P_n(X_i=\underline{x}_j|h(X_i)=h_l) \\ & \times\|\sqrt{n}\{F_{t,n}(.|\underline{x}_j)-\bar{F}_{t,n,l}(.)\}\|_{\infty,\mathcal{Y}} \end{aligned}$$

$$\begin{aligned} & CM_{n,\phi} \\ = \; & \frac{1}{T}\sum_{t=1}^{T}\sum_{l=1}^{L} P_n(h(X_i)=h_l)\sum_{j=1}^{m_l} P_n(X_i=\underline{x}_j|h(X_i)=h_l) \\ & \times\|\sqrt{n}\{F_{t,n}(.|\underline{x}_j)-\bar{F}_{t,n,l}(.)\}\|_{2,\phi}. \end{aligned}$$

Let $(\zeta(.,\underline{x}))_{\underline{x}\in\mathcal{X}^T}$ be the vector that contains the elements $\{\zeta(.,\underline{x}):\underline{x}\in\mathcal{X}^T\}$.

$$\begin{pmatrix} \sqrt{n}\left(F_{1,n}(.|X_i=\underline{x})-F_1(.|X_i=\underline{x})\right)_{\underline{x}\in\mathcal{X}^T} \\ \sqrt{n}\left(F_{2,n}(.|X_i=\underline{x})-F_2(.|X_i=\underline{x})\right)_{\underline{x}\in\mathcal{X}^T} \\ \cdots \\ \sqrt{n}\left(F_{T,n}(.|X_i=\underline{x})-F_T(.|X_i=\underline{x})\right)_{\underline{x}\in\mathcal{X}^T} \end{pmatrix} \rightsquigarrow \mathbb{G} \tag{3.80}$$

Since $T<\infty$ and $|\mathcal{X}^T|<\infty$, the joint distribution of the centered empirical conditional cdfs converges to a tight Brownian bridge. Now note that a linear combination of the above yields the empirical process that we use to construct our statistics.

$$\left(\left(\begin{array}{c} \sqrt{n}(F_{1,n}(.|\underline{x}_j)-\bar{F}_{1,n,l}(.)-(F_1(.|\underline{x}_j)-\bar{F}_{1,l}(.))) \\ \cdots \\ \sqrt{n}(F_{T,n}(.|\underline{x}_j)-\bar{F}_{T,n,l}(.)-(F_T(.|\underline{x}_j)-\bar{F}_{T,l}(.))) \end{array}\right)_{j=1,2,\ldots,m_l}\right)_{l=1,2,\ldots,L} \rightsquigarrow \mathbb{H},$$

where $\mathbb{H}\equiv((\mathbb{H}_{j,l})_{j=1,..,m_l})_{l=1,..,L}$. Note that the above process is a $(T\sum_{l=1}^{L}m_l)\times 1$

vector of functionals. Since all of the above processes are defined on a Donsker class, the bootstrap empirical process also converges to the same limit process by Theorem 3.6.1 in der Vaart and Wellner (2000).

$$\left(\left(\begin{array}{c} \sqrt{n}(\hat{F}_{1,n}(.|\underline{x}_j) - \hat{\bar{F}}_{1,n,l}(.) - (F_{1,n}(.|\underline{x}_j) - \bar{F}_{1,n,l}(.))) \\ \cdots \\ \sqrt{n}(\hat{F}_{T,n}(.|\underline{x}_j) - \hat{\bar{F}}_{T,n,l}(.) - (F_{T,n}(.|\underline{x}_j) - \bar{F}_{T,n,l}(.))) \end{array} \right)_{j=1,2,...,m_l} \right)_{l=1,2,...,L}$$
$$\rightsquigarrow \quad \mathbb{H}.$$

Now we give the limiting statistics as follows,

$$\begin{aligned} KS_{\mathcal{Y}} &= \frac{1}{T}\sum_{t=1}^{T}\sum_{l=1}^{L} P(h(X_i) = h_l) \sum_{j=1}^{m_l} P(X_i = \underline{x}|h(X_i) = h_l)\|\mathbb{H}_{j,l}\|_{\infty,\mathcal{Y}} \\ CM_{\phi} &= \frac{1}{T}\sum_{t=1}^{T}\sum_{l=1}^{L} P(h(X_i) = h_l) \sum_{j=1}^{m_l} P_n(X_i = \underline{x}|h(X_i) = h_l)\|\mathbb{H}_{j,l}\|_{2,\phi}. \end{aligned} \quad (3.81)$$

Since the above is a linear combination of convex continuous functionals, it follows that Theorem 11.1 in Davydov, Lifshits, and Smorodina (1998) applies. Thus, the distributions of $KS_{\mathcal{Y}}$ and CM_{ϕ} are absolutely continuous and strictly increasing on $(0, \infty)$. Since $F_t(.|X_i = \underline{x})$ is non-degenerate for $\underline{x} \in \mathcal{X}^T$ and $t = 1, 2..., T$, it follows that the $P(KS_{\mathcal{Y}} = 0) = 0$ and $P(CM_{\phi} = 0) = 0$. Hence, it follows that their distribution is absolutely continuous on $[0, \infty)$. Now it remains to show that $KS_{n,\mathcal{Y}}$ and $CM_{n,\phi}$ converge to $KS_{\mathcal{Y}}$ and CM_{ϕ}, respectively.

Let T_n with norm $\|.\|$ denote either the KS or CM with their respective norms,

and let $\mathbb{H}_{n,j,l}$ denote the relevant empirical process

$$\begin{aligned}
T_n &= \frac{1}{T}\sum_{t=1}^{T}\sum_{l=1}^{L} P_n(h(X_i)=h_l)\sum_{j=1}^{m_l} P_n(X_i=\underline{x}_j|h(X_i)=h_l)\|\mathbb{H}_{n,j,l}\| \\
&= \frac{1}{T}\sum_{t=1}^{T}\sum_{l=1}^{L} P(h(X_i)=h_l) \\
&\quad \times \sum_{j=1}^{m_l} P(X_i=\underline{x}|h(X_i)=h_l)\|\mathbb{H}_{n,j,l}\| \\
&+ \frac{1}{T}\sum_{t=1}^{T}\sum_{l=1}^{L} P_n(h(X_i)=h_l) \\
&\quad \times \sum_{j=1}^{m_l} (P_n(X_i=\underline{x}|h(X_i)=h_l) - P(X_i=\underline{x}|h(X_i)=h_l))\|\mathbb{H}_{n,j,l}\| \\
&+ \frac{1}{T}\sum_{t=1}^{T}\sum_{l=1}^{L} (P_n(h(X_i)=h_l) - P(h(X_i)=h_l)) \\
&\quad \times \sum_{j=1}^{m_l} P_n(X_i=\underline{x}|h(X_i)=h_l)\|\mathbb{H}_{n,j,l}\| \\
&\rightsquigarrow \mathcal{T}
\end{aligned}$$

where $\mathcal{T}$ equals $KS_{\mathcal{Y}}$ and CM_{ϕ} for the KS and CM statistics, respectively. The convergence follows since the latter two terms converge in probability to zero, since $(P_n(X_i=\underline{x}|h(X_i)=h_l) - P(X_i=\underline{x}|h(X_i)=h_l)) \xrightarrow{p} 0$ and $(P_n(h(X_i)=h_l) - P(h(X_i)=h_l)) \xrightarrow{p} 0$, and both terms are multiplied by $O_p(1)$ terms.

Now it follows that the tests based on the bootstrap critical values yield correct asymptotic size, which gives (i). Since the empirical processes on which the tests are based are all tight, the critical values of the tests are bounded in probability. Hence, the tests are consistent against any fixed alternatives, which yields (ii). □

3.6.3 Supplementary Results: Testing Identifying Restrictions

Theorem 3.6.1 *Given two random samples of size n and m from $F_X(.)$ and $F_Y(.)$, respectively, where $F_X(.)$ and $F_Y(.)$ are atomless distribution functions, $D_{n_1,m}$ is distribution-free, where*

$$D_{n,m} = \sup_{z\in\mathcal{Z}} |\hat{F}_X(z) - \hat{F}_Y(z)|, \tag{3.82}$$

Proof Note that $D_{n,m} = \max\{D^+_{n,m}, D^-_{n,m}\}$, where $D^+_{n,m} = \sup_{z\in\mathcal{Z}}[\hat{F}_X(z) - \hat{F}_Y(z)]$ and $D_{n,m} = \sup_{z\in\mathcal{Z}}[\hat{F}_Y(z) - \hat{F}_X(z)]$. Let $\{Z_i\}_{i=1}^{n+m}$ be the combined observations from both samples, and let $Z_{(i)}$ be the i^{th} order statistic and add $Z_{(0)} = -\infty$ and $Z_{(n+m+1)} = \infty$. Also, $r_i \equiv \sum_{j=1}^{n}\{X_j \le Z_{(i)}\}$ and $s_i \equiv \sum_{j=1}^{m}\{Y_j \le Z_{(i)}\}$.

$$\begin{aligned}
D^+_{n,m} &= \max_{0\le i\le n+m} \sup_{Z_{(i)}\le y<Z_{(i+1)}} \left[\hat{F}_1(y) - F_X(y) - (\hat{F}_2(y) - F_2(y)) + F_X(y) - F_2(y)\right] \\
&= \max_{0\le i\le n+m} \left[\frac{r_i}{n} - \inf_{Z_{(i)}\le y<Z_{(i+1)}} F_X(y) - \left(\frac{s_{i-1}}{m} - \sup_{Z_{(i)}\le y<Z_{(i+1)}} F_2(y)\right)\right] \\
&\quad + \max_{0\le i\le n+m} \left[\sup_{Z_{(i)}\le y<Z_{(i+1)}} (F_X(y) - F_2(y))\right] \\
&= \max_{0\le i\le n+m} \left[\frac{r_i}{n} - F_X(Z_{(i)}) - \left(\frac{s_{i-1}}{m} - F_2(Z_{(i)})\right)\right] \\
&\quad + \max_{0\le i\le n+m} \left[F_X(Z_{(i)}) - F_2(Z_{(i)})\right] \\
&= \max[\max_{1\le i\le n+m} \left[\frac{r_i}{n} - F_X(Z_{(i)}) - \left(\frac{s_{i-1}}{m} - F_Y(Z_{(i)})\right)\right] \\
&\quad + \max_{1\le i\le n+m} \left[F_X(Z_{(i)}) - F_Y(Z_{(i)})\right], 0]
\end{aligned} \tag{3.83}$$

By the probability-integral transform theorem and continuity, $F_X(Z_{(j)})$ and $F_Y(Z_{(j)})$ are both j^{th} order statistics from $U(0,1)$ variables, for $j = 1, 2$. hence $D^+_{n,m}$ is

distribution-free. Similarly,

$$\begin{aligned} D^{-}_{n,m} &= \max[\max_{1\le i\le n}\left[\frac{s_i}{m} - F_Y(Z_{(i)}) - (\frac{r_{i-1}}{n} - F_X(Z_{(i)}))\right] \\ &\quad - \min_{1\le i\le n}\left[(F_X(Z_{(i)}) - F_Y(Z_{(i)}))\right], 0] \end{aligned}$$

$$\begin{aligned} D_{n,m} &= \max[\max_{1\le i\le n}\left[\frac{s_i}{m} - F_Y(Z_{(i)}) - \left(\frac{r_{i-1}}{n} - F_X(Z_{(i)})\right)\right] \\ &\quad + \max[\max_{1\le i\le n}\left[F_X(Z_{(i)}) - F_Y(Z_{(i)})\right], \\ &\quad \max_{1\le i\le n+m}\left[\frac{r_i}{n} - F_X(Z_{(i)}) - \left(\frac{s_i}{m} - F_Y(Z_{(i)})\right)\right]] \\ &\quad + \max_{1\le i\le n+m}\left[(F_X(Z_{(i)}) - F_Y(Z_{(i)}))\right], 0] \end{aligned}$$

□

Lemma 3.6.1 *For an iid sequence of events,* $\{A_i\}_{i=1}^n$, $P(A_i) > 0$, $\hat{P}_n(A_i) \equiv \sum_{i=1}^n M_{ni} 1\{A\}/n \xrightarrow{p} P(A)$ *as* $n \to \infty$.

Proof Note that $M_{ni} \sim Bin(n, n^{-1})$ independent of $\{A_i\}_{i=1}^n$. Now,

$$E\left[\frac{\sum_{i=1}^n M_{ni} 1\{A_i\}}{n}\right] = \frac{\sum_{i=1}^n 1\{A_i\}}{n} \tag{3.84}$$

Now since M_{ni} is not iid across i, we cannot apply a law of large numbers directly. We use the Poissonized process, where $\{M_{N_n,i}\}_{i=1}^n$ are iid Poisson variables with mean 1.

$$\hat{P}_n(A_i) = \frac{\sum_{i=1}^n (M_{ni} - M_{N_n,i}) 1\{A_i\}}{n} + \frac{\sum_{i=1}^n M_{N_n,i} 1\{A_i\}}{n} \tag{3.85}$$

So we can apply a weak law of large numbers to the latter term to show convergence

to $P(A_i)$, since given $\{Y_i, X_i\}_{i=1}^n$ and N_n

$$E\left[\frac{\sum_{i=1}^n M_{N_n,i} 1\{A_i\}}{n}\right] = \frac{\sum_{i=1}^n 1\{A_i\}}{n} \tag{3.86}$$

Since $E[N_n] = n$ and it is independent of $\{A_i\}_{i=1}^n$, by the law of large numbers

$$\frac{\sum_{i=1}^n M_{N_n,i} 1\{A_i\}}{N_n} \xrightarrow{p} P(A_i) \tag{3.87}$$

It remains to show that the residual term $\frac{\sum_{i=1}^n (M_{ni} - M_{N_n,i}) 1\{\Delta X_i = 0\}}{n}$ converges to zero in probability. From the proof of Theorem 3.6.2. in der Vaart and Wellner (2000), we have that

$$P\left(\max_{1\le i\le n} |M_{N_n,i} - M_{n,i}| > 2\right) \to \epsilon \tag{3.88}$$

So we can write

$$\begin{aligned}
&\frac{\sum_{i=1}^n |M_{N_n,i} - M_{ni}|}{n} 1\{\max_{1\le i\le n} |M_{N_n,i} - M_{n,i}| > 2\} \\
&+\frac{\sum_{i=1}^n |M_{N_n,i} - M_{ni}|}{n} 1\{\max_{1\le i\le n} |M_{N_n,i} - M_{n,i}| \le 2\} \\
= \;& o_p(1) + \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^\infty 1\{|M_{N_n,i} - M_{ni}| \le j\} 1\{\max_{1\le i\le n} |M_{N_n,i} - M_{n,i}| \le 2\}
\end{aligned}$$

Now we can write the difference between the bootstrap empirical process and the

Poissonized process as follows

$$\begin{aligned}
& \frac{\sum_{i=1}^{n}(M_{n,i}-M_{ni})1\{\Delta X_i=0\}}{n} \\
\leq\ & \frac{\sum_{i=1}^{n}|M_{n,i}-M_{ni}|1\{\Delta X_i=0\}}{n} \\
=\ & \frac{\sum_{i=1}^{n}|M_{N_n,i}-M_{ni}|}{n}1\{\max_{1\leq i\leq n}|M_{N_n,i}-M_{n,i}|>2\}1\{\Delta X_i=0\} \\
& +\frac{\sum_{i=1}^{n}|M_{N_n,i}-M_{ni}|}{n}1\{\max_{1\leq i\leq n}|M_{N_n,i}-M_{n,i}|\leq 2\}1\{\Delta X_i=0\}
\end{aligned}$$

$$\begin{aligned}
\leq\ & o_p(1)+\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{\infty}1\{|M_{N_n,i}-M_{ni}|\leq j\}1\{\max_{1\leq i\leq n}|M_{N_n,i}-M_{n,i}|\leq 2\} \\
=\ & o_p(1)+\frac{1}{n}\sum_{j=1}^{\infty}\frac{\#I_n^j}{n}\left(\frac{1}{\#I_n^j}\sum_{i\in I_n^j}1\{\Delta X_i=0\}\right) \\
\leq\ & o_p(1)+\frac{1}{n}\sum_{j=1}^{2}\frac{\#I_n^j}{n}\left(\frac{n}{\sum_{i=1}^{n}1\{|M_{N_n,i}-M_{n,i}|\geq j\}}\frac{\sum_{i=1}^{n}1\{\Delta X_i=0\}}{n}\right) \\
=\ & o_p(1)+\frac{O_p(n^{1/2})}{n}O_p(1)=o_p(1) \qquad (3.89)
\end{aligned}$$

where $I_n^j=\{i\in\{1,2,...,n\}:|M_{N_n,i}-M_{ni}|\geq j\}$. Now the penultimate equality follows from $\#I_n^j\leq|N_n-n|=O_p(n^{1/2})$ for all j. Now the second term follows by the law of large numbers and $P(|M_{N_n,i}-M_{n,i}|=0,\hat{\tau},\tilde{\tau})=1-2\epsilon$, such that $P(|M_{N_n,i}-M_{n,i}|=j)>0$ for $j=0,1,2$. □

Lemma 3.6.2 *Given Assumption 3.4.1,* $\rho_2(s+\epsilon,s)\leq K|\epsilon|$.

Proof (Lemma 3.6.2)

$$\begin{aligned}
& \rho_2(s+\epsilon,s) \\
=\ & E[(\mathbb{G}(s+\epsilon)-\mathbb{G}(s))^2] \\
=\ & E[\mathbb{G}(s+\epsilon)^2|\epsilon]+E[\mathbb{G}(s)^2]-2E[\mathbb{G}(s+\epsilon)\mathbb{G}(s)]
\end{aligned}$$

$$
\begin{aligned}
&= |F(s+\epsilon)(1-F(s+\epsilon)) + F(s)(1-F(s)) - 2\{F(\{s+\epsilon\}\wedge s) - F(s+\epsilon)F(s)\}| \\
&\leq |F(s+\epsilon)(1-F(s+\epsilon)) - \{F(\{s+\epsilon\}\wedge s) - F(s+\epsilon)F(s)\}| \qquad (3.90) \\
&+ |F(s)(1-F(s)) - \{F(\{s+\epsilon\}\wedge s) - F(s+\epsilon)F(s)\}| \qquad (3.91)
\end{aligned}
$$

Now for (3.90)

$$
\begin{aligned}
& |F(s+\epsilon)(1-F(s+\epsilon)) - \{F(\{s+\epsilon\}\wedge s) - F(s+\epsilon)F(s)\}| \\
= \; & |F(s+\epsilon)(1-F(s+\epsilon)) - F(\{s+\epsilon\}\wedge s)(1-F(\{s+\epsilon\}\vee s))\}| \\
\leq \; & |F(s+\epsilon)(F(\{s+\epsilon\}\vee s) - F(\{s+\epsilon\}\vee s))| \\
& +|(F(s+\epsilon) - F(\{s+\epsilon\}\wedge s))(1-F(\{s+\epsilon\}\vee s))| \\
\leq \; & |F(\{s+\epsilon\}\vee s) - F(s+\epsilon)| + |F(s+\epsilon) - F(\{s+\epsilon\}\wedge s)| \\
\leq \; & f(s)\{|\max(s+\epsilon, s) - (s+\epsilon)| + |s+\epsilon - \min(s+\epsilon, s)| \\
= \; & f(s)\{|\max(0,-\epsilon)| + |min(0,\epsilon)|\} \\
\leq \; & 2|f(s)||\epsilon|
\end{aligned}
$$

where $f(s)$ is the density of F. Similar manipulation of (3.91) yields

$$
|F(s)(1-F(s)) - \{F(\{s+\epsilon\}\wedge s) - F(s+\epsilon)F(s)\}| \;\leq\; 2|f(s)||\epsilon_n|
$$

Assumption 3.4.1 delivers the result with $K = 4\sup_{s\in\mathbb{R}}|f(s)|$. □

We need to introduce some notation for the following lemma. Let $\mathbb{G}_n \equiv \sqrt{n}(F_n(.) - F(.))$. $\mathbb{G}_{n_A} \equiv \sqrt{n}(F_n(.|A_i) - F(.|A_i))$ and $\hat{\mathbb{G}}_{n_A} \equiv \sqrt{n}(\hat{F}_n(.|A_i) - F_n(.|A_i))$. $\mathbb{G}_n \rightsquigarrow \mathbb{G}$, a $\mathbb{P}$-Brownian bridge in $\mathcal{L}^\infty(\mathcal{Y})$, where $\mathcal{Y}$ is the support of Y_i. Clearly,

Lemma 3.6.3 *Given $\{Y_i, A_i\}_{i=1}^n$ an iid sequence, where A_i is an event with $P(A_i) > 0$, then*

(i) $\mathbb{G}_{n|A} \rightsquigarrow \mathbb{G}_A = \frac{\mathbb{H}}{P(A_i)} + \mathcal{Z}F(., A_i)$

(ii) $\hat{\mathbb{G}}_{n|A} \rightsquigarrow \mathbb{G}_A$

where $\mathbb{H}$ *is* $\mathbb{P}$*-Brownian bridge in* $\mathcal{L}^\infty(\mathcal{F})$*, where* $\mathcal{F} = \{1\{y \leq t\}1\{A_i\} : t \in \mathcal{Y}\}$*, and* $\sqrt{n}((P_n(A_i))^{-1} - (P(A_i))^{-1}) \rightsquigarrow \mathcal{Z}$*.*

Proof For (i), let $P_n(A) \equiv \sum_{i=1}^n 1\{A_i\}$.

$$\begin{aligned}
\mathbb{G}_{n_A} &= \sqrt{n}(F_n(.|A) - F(.|A)) \\
&= \sqrt{n}\left(\frac{F_n(., A) - F(., A)}{P_n(A)} + F(., A)\left(\frac{1}{P_n(A)} - \frac{1}{P(A)}\right)\right) \\
&\rightsquigarrow \frac{\mathbb{H}}{P(A_i)} + \mathcal{Z}F(., A_i) \equiv \mathbb{G}_A
\end{aligned}$$

The result for the first term follows by the Donsker property of $\mathcal{F}$ and Slutzky's theorem. The latter trivially follows by $P(A_i) > 0$ and the continuous mapping theorem.

For (ii), let $\hat{P}_n(A) \equiv \sum_{i=1}^n M_{ni}1\{A_i\}/n$

$$\begin{aligned}
\hat{\mathbb{G}}_{n_A} &= \sqrt{n}(\hat{F}_n(.|A_i) - F_n(.|A_i)) \\
&= \sqrt{n}\left(\frac{\hat{F}_n(., A_i)}{\hat{P}_n(A_i)} - \frac{F_n(., A_i)}{P_n(A_i)}\right) \\
&= \sqrt{n}\left(\frac{\hat{F}_n(., A_i) - F(., A_i)}{\hat{P}_n(A_i)} + F_n(., A_i)\left(\frac{1}{\hat{P}_n(A_i)} - \frac{1}{P_n(A_i)}\right)\right) \\
&= \frac{\hat{\mathbb{G}}_n}{\hat{P}_n(A_i)} + F_n(., A_i)\sqrt{n}\left(\frac{1}{\hat{P}_n(A_i)} - \frac{1}{P_n(A_i)}\right) \\
&\rightsquigarrow \frac{\mathbb{H}}{P(A_i)} - F(., A_i)\mathcal{Z} \qquad (3.92)
\end{aligned}$$

where the weak convergence of the first term follows by the Donsker property of $\mathcal{F}$, which implies that Theorem 3.6.1 of der Vaart and Wellner (2000) applies.

The second term follows by the Glivenko-Cantelli property of $F(.,A_i)$ and the continuous mapping theorem. □

3.7 Paired-Sample Problem

We observe an iid sequence $\{Y_{i1}, Y_{i2}\}_{i=1}^n$. We are interested in testing the equality of the distribution of the demeaned variables $W_{i,n} = \{Y_{i1} - \lambda_n, Y_{i2}\}$, where $\lambda_n = \bar{Y}_{1,n} = \sum_{i=1}^n (Y_{i1} - Y_{i2})/n$. Now we will introduce some notation. Let $F_1(.)$ and $F_2(.)$ denote the distributions of Y_{i1} and Y_{i2}, respectively. The following are their empirical versions

$$F_{1,n}(t,\lambda_n) = \frac{1}{n}\sum_{i=1}^n 1\{Y_{i1} - \lambda_n \leq t\} \qquad F_{2,n} = \frac{1}{n}\sum_{i=1}^n 1\{Y_{i2} \leq t\} \tag{3.93}$$

and their bootstrap empirical versions

$$\hat{F}_{1,n}(t+\hat{\lambda}_n) = \frac{1}{n}\sum_{i=1}^n M_{ni} 1\{Y_{i1} - \hat{\lambda}_n \leq t\} \qquad \hat{F}_{2,n}(t) = \frac{1}{n}\sum_{i=1}^n M_{ni} 1\{Y_{i2} \leq t\}$$

Now we need to show weak convergence of the following empirical process

$$\mathbb{H}_n(\lambda_n) \;=\; \sqrt{n}(F_{1,n}(.+\lambda_n) - F_{2,n}(.) - (F_1(.+\lambda) - F_2(.))) \tag{3.94}$$

and its bootstrap empirical counterpart,

$$\hat{\mathbb{H}}_n(\hat{\lambda}_n) \;=\; \sqrt{n}(\hat{F}_{1,n}(.+\hat{\lambda}) - \hat{F}_{2,n}(.) - (F_{1,n}(.,\lambda) - F_{2,n}(.))). \tag{3.95}$$

Now to show weak convergence of the above process, we need to show that the delta method applies here to the above empirical process, such that we can use the

convergence of the following:

$$\sqrt{n}\begin{pmatrix} F_{1,n}(.) - F_1(.) \\ F_{2,n}(.) - F_2(.) \\ \lambda_n - \lambda \end{pmatrix} \rightsquigarrow \begin{pmatrix} \mathbb{G}_1 \\ \mathbb{G}_2 \\ \mathcal{E} \end{pmatrix} \tag{3.96}$$

The main complication is due to the evaluation of the distribution function at Y_{i1} after imposing a location shift. Let $\mathcal{Y}$ denote the support Y_{i1}. We first proceed to showing that the map $\phi : \mathcal{L}^\infty(\mathcal{Y}) \times \mathcal{Y} \mapsto \mathcal{L}^\infty(\mathcal{Y})$, which is given below, is Hadamard differentiable. Let G denote some distribution and γ a location shift.

$$\phi(G, \gamma) \quad = \quad G(. - \gamma) \tag{3.97}$$

We need to impose that the underlying distribution has a bounded density. For a Gaussian process, $\mathbb{G}$, define $\rho_2(s,t) = E|\mathbb{G}(s) - \mathbb{G}(t)|^2$ and $\mathbb{D} \equiv \{h \in \mathcal{L}^\infty(\mathcal{Y}) :$ h is ρ_2-uniformly continuous$\}$.

Lemma 3.7.1 *Assume that the distribution function F has a bounded density, $\phi(F, \lambda) : \mathcal{L}^\infty(\mathcal{Y}) \times \mathcal{Y} \mapsto \mathcal{L}^\infty(\mathcal{Y})$ is Hadamard differentiable at (F, λ) tangentially to $\mathbb{D} \times \mathcal{Y}$ with the following derivative*

$$\phi'_{F,\lambda}(g, \epsilon) \quad = \quad g(. - \lambda) - \epsilon f(. - \lambda),$$

for $(g, \epsilon) \in \mathbb{D} \times \mathbb{R}$.

Proof Note that $\phi'_{F,\lambda}(g, \epsilon)$ is clearly linear and continuous in g and ϵ. Now we need to show that

$$\frac{\phi(F + \tau_n g_n, \lambda + \tau_n \epsilon_n) - \phi(F, \lambda)}{\tau_n} \quad \rightarrow \quad \phi'_{F,\lambda}(g, \epsilon) \qquad n \rightarrow \infty \tag{3.98}$$

for all converging sequences $\tau_n \searrow 0$, $g_n \to g \in \mathbb{D}$, and $\epsilon_n \to \epsilon \in \mathbb{R}$.

$$\begin{aligned}
& \left\| \frac{\phi(F + \tau_n g_n, \lambda + \tau_n \epsilon_n) - \phi(F, \lambda)}{\tau_n} - \phi'_{F,\lambda}(g, \epsilon) \right\|_\infty \\
= & \left\| \frac{(F + \tau_n g_n)(. - \lambda - \tau_n \epsilon_n) - F(. - \lambda)}{\tau_n} - (g(. - \lambda) - \epsilon f(. - \lambda)) \right\|_\infty \\
\leq & \left\| \frac{F(. - \lambda - \tau_n \epsilon_n) - F(. - \lambda)}{\tau_n} \right. \\
& \left. + \epsilon f(. - \lambda) \right\|_\infty + \left\| g_n(. - \lambda - \tau_n \epsilon_n) - g(. - \lambda) \right\|_\infty
\end{aligned} \tag{3.99}$$

For the first term of (3.99),

$$\begin{aligned}
& \left\| \frac{F(. - \lambda - \tau_n \epsilon_n) - F(. - \lambda)}{\tau_n} + \epsilon f(. - \lambda) \right\|_\infty \\
\leq & \left\| \frac{F(. - \lambda - \tau_n \epsilon_n) - F(. - \lambda)}{\tau_n} - \frac{F(. - \lambda - \tau_n \epsilon) - F(. - \lambda)}{\tau_n} \right\|_\infty \\
+ & \left\| \frac{F(. - \lambda - \tau_n \epsilon) - F(. - \lambda)}{\tau_n} + \epsilon f(. - \lambda) \right\|_\infty \\
\leq & \sup_{t \in \mathbb{R}} |f(t)| |\epsilon_n - \epsilon| + |\epsilon| \sup_{t \in \mathbb{R}} \left| \frac{F(t - \lambda - \tau_n \epsilon) - F(t - \lambda)}{\epsilon \tau_n} - f(t - \lambda) \right| \\
= & \sup_{t \in \mathbb{R}} |f(t)| |\epsilon_n - \epsilon| + |\epsilon| \sup_{t \in \mathbb{R}} |f(w) - f(t - \lambda)| \;\; w \in (t - \lambda - \tau_n \epsilon, t - \lambda)
\end{aligned} \tag{3.100}$$

$$\leq \sup_{t \in \mathbb{R}} |f(t)| |\epsilon_n - \epsilon| + |\epsilon \delta| \tag{3.101}$$

The penultimate equality follows by the Mean Value Theorem.[38] The second term of the last equality follows by uniform continuity implied by Assumption 3.4.1, which implies that for $|w - t - \lambda| < |\tau_n \epsilon| < \nu$, there is a δ that bounds the second

[38] By the Mean Value Theorem, uniform continuity is equivalent to uniform differentiability.

term in (3.100) uniformly in t. Now as $\tau_n \to 0$ and $\epsilon_n \to \epsilon$, both terms converge to zero.

As for the second term of (3.99)

$$\begin{aligned} & \|g_n(.-\lambda-\tau_n\epsilon_n) - g(.-\lambda)\|_\infty \\ \leq\ & \|g_n(.-\lambda-\tau_n\epsilon_n) - g(.-\lambda-\tau_n\epsilon_n)\|_\infty \\ & + \sup_{t\in\mathbb{R}} |g(t-\lambda-\tau_n\epsilon_n) - g(t+\lambda)| \\ \leq\ & \|g_n(.-\lambda-\tau_n\epsilon_n) - g(.-\lambda-\tau_n\epsilon_n)\|_\infty \\ & + \sup_{\rho_2(t-\lambda-\tau_n\epsilon_n, t-\lambda)} |g(t-\lambda-\tau_n\epsilon_n) - g(t-\lambda)| \\ \to\ & 0 \qquad\qquad \tau_n \to 0, \epsilon_n \to \epsilon, g_n \to g \end{aligned} \tag{3.102}$$

where the first term follows by weak convergence for all $g \in \mathbb{D}$. As for the latter term, by Lemma 3.6.2 $\tau_n \to 0$ implies that $\rho_2(t-\lambda-\tau_n\epsilon_n, t-\lambda) \to 0$. Since $g \in \mathbb{D}$, which is uniformly ρ_2-continuous, the result follows. □

Theorem 3.7.1 *Assume that the distribution function F_1 has bounded density,*

(i) $\mathbb{H}_n(\lambda_n) \rightsquigarrow \mathbb{H}(\lambda)$,

(ii) $\hat{\mathbb{H}}_n(\hat{\lambda}_n) \rightsquigarrow \mathbb{H}(\lambda)$

where $\mathbb{H}(\lambda)$ is a tight Brownian bridge in $L^\infty(\mathcal{Y})$.

Proof (i) Note that

$$\sqrt{n}\begin{pmatrix} F_{1,n}(.) - F_1(.) \\ F_{2,n}(.) - F_2(.) \\ \lambda_n - \lambda \end{pmatrix} \rightsquigarrow \begin{pmatrix} \mathbb{G}_1 \\ \mathbb{G}_2 \\ \mathcal{E} \end{pmatrix} \tag{3.103}$$

where $\mathbb{G}_1$ and $\mathbb{G}_2$ are tight Brownian bridges in $L^\infty(\mathcal{Y})$ and $\mathcal{E}$ is a normal scalar random variable.

Now let $\phi(F, \lambda) = F_1(. - \lambda)$. By Lemma 3.7.1, $\phi : \mathcal{L}^\infty(\mathcal{Y}) \times \mathbb{R} \mapsto L^\infty(\mathbb{R})$ is Hadamard differentiable at F, λ tangentially to $\mathbb{D} \times \mathbb{R}$. Thus, by Theorem 3.9.4. in der Vaart and Wellner (2000) the delta method applies and

$$\sqrt{n} \begin{pmatrix} \phi(F_{1,n}, \lambda_n) - \phi(F_1, \lambda) \\ F_{2,n} - F_2 \end{pmatrix} \rightsquigarrow \begin{pmatrix} \phi'_{F,\lambda}(\mathbb{G}_1, \mathcal{E}) \\ \mathbb{G}_2 \end{pmatrix}$$

The result follows by the continuous mapping theorem.

(ii) It suffices to check the conditions of Theorem 3.9.11 in der Vaart and Wellner (2000), which are that the empirical process in (3.103), where $(\mathbb{G}_1, \mathbb{G}_2, \mathcal{E})$ are separable and in $\mathbb{D} \times L^\infty(\mathcal{Y}) \times \mathcal{Y}$. Since $\mathbb{G}_1$ and $\mathbb{G}_2$ are tight Brownian bridges and $\mathcal{E}$ is a scalar normal random variable, which implies that the limiting process is separable and by assumption it is in $\mathbb{D} \times \mathcal{L}^\infty(\mathcal{Y}) \times \mathcal{Y}$.

Furthermore, we need condition (3.9.9), p. 378, in der Vaart and Wellner (2000) to hold in probability. Theorem 3.6.2 implies that (3.9.9) holds almost surely, and hence in probability, if $\mathcal{F}$ is Donsker and $\|P(f - Pf)\|_{\mathcal{F}} < \infty$. Since $\mathcal{F} = \{1\{y_1 \leq t\} - 1\{y_2 \leq t : t \in \mathcal{Y}\}$, it is clearly Donsker. The second condition is also trivially fulfilled, since $\sup_{t \in \mathcal{Y}} |\int \{1\{y_1 \leq t\} - 1\{y_2 \leq t\} - (F_1(t) - F_2(t))\} dF(y)| \leq 2$. $\square$

Chapter 4

'Effortless' Perfection: Do Chinese cities manipulate "blue skies"?

4.1 Introduction

Air quality of major Chinese cities is among the worst in the world, a consequence of three decades of double-digit economic growth with lax environmental regulation (Zhang and Crooks, 2012). The central government of China started to address the issue of environmental regulation. Over a decade ago, major Chinese cities were required to report their daily air pollution index (API) and to make it available to the public. To incentivize air pollution abatement, performance evaluations of local officials included the number of "blue-sky days," which are days with API below 100.[1] These measures introduced both bottom-up and top-down pressures on the local governments to meet the target number of blue-sky days. In the absence of independent verification mechanisms, some cities are allegedly under-reporting their air pollution levels. The presence of a particular threshold for blue-sky days seems to be associated with dubious behavior of the API scores around the cut-off, which is illustrated in Figure A.1. We call this phenomenon "effortless perfection." Specifically, the API distributions of three municipalities, including Beijing, Tianjin and Chongqing, amass right below the cut-off. The irregularities have raised suspicion of systematic manipulation of air pollution data in China.

At the heart of this paper is a question about self-reported data that are used in performance evaluations of the reporting agencies. In such settings, there is good reason to suspect that the agencies manipulate the data to improve their performance evaluations, especially if such manipulation can go undetected. The relevance of this paper is hence not restricted to air pollution reporting. Using the unique data set that we have, our main goal is to understand if and how manipulation occurs in this setup. The importance of such a question goes beyond

[1] A blue-sky day is associated with low or moderate health concerns.

detecting cheaters. It can help us understand how manipulation can undermine the use of such self-reported data in economic studies.

Data manipulation is a concern for both policy makers and researchers. For policy makers, data manipulation defeats the purpose of policies that incentivize pollution abatement and other policy objectives. It might be effortless for a city to beautify its environmental record by just tweaking the numbers. However, society must then bear the costly consequence of misleading pollution data. First, misinformed urban residents cannot efficiently mitigate pollution-related health effects by changing their behavior, for example wearing masks or canceling outdoor activities. Second, promotion of officials whose cities have become involved in manipulation jeopardizes the public interest because their integrity is in question. In addition, environmental compliance through manipulation undermines the government's credibility.

Manipulation introduces measurement error into the data that may be used as regressors to study the effect of pollution on health and other outcomes. Such measurement error will bias studies that evaluate the impact of air pollution using such data on various outcomes.[2] The biased marginal effects of pollution might lead to incorrect policy recommendations. If the measurement error in the air pollution data violates the classical measurement-error assumptions, then standard econometric methods may not rectify the bias due to the measurement error in this data.

Our paper aims to identify irregularities in the air pollution data in order to provide insight on the nature of manipulation and the circumstances under which it occurs. We define manipulation as the behavior of not reporting the true pollution

[2]A large body of literature in environmental and health economics studies the impact of air pollution on productivity (Koop and Tole, 2004), property values (Chay and Greenstone, 2005), and health (Graff Zivin and Neidell, 2012). The reliability of the estimated marginal effects hinges on valid pollution data.

level, such as data falsification or hiding bad pollution data. It does not include strategic behavior such as temporary driving bans, closing factories, or requiring different fuels.[3] Although these command-and-control policies are inefficient, they can indeed reduce pollution in the short run so we cannot call them manipulation. The best way to detect manipulation is to use alternative sources of data to validate the official data. However, independent direct measures are not available for all the relevant pollutants, cities, and periods.[4] Therefore, this paper uses econometric methods to uncover suggestive evidence of manipulation from the self-reported data.

We pose two research questions. First, we investigate whether a city reports dubious pollution data around the cutoff for blue-sky days.[5] To answer this question, our empirical strategy is borrowed from the McCrary (2008) test, which is used to detect manipulation of a running variable in regression discontinuity design. The intuition here is that if the pollutant concentration on a particular day misses the cutoff for blue-sky day by a small amount, then there is an incentive for the city to under-report the pollutant concentration and score a blue-sky day. If such behavior occurs often, the distribution of the pollutant concentration exhibits a discontinuity at the cutoff for a blue-sky day. In the absence of manipulation, the distribution of air pollutant concentrations is expected to be continuous because polluters and regulators do not have complete control over the realized pollutant

[3]These policies were used during major events such as 2008 Olympic Games in Beijing and Expo 2010 in Shanghai. They are also used when local governments are desperate to meet their environmental targets at the end of a particular year.

[4]A notable independent measure is the U.S Embassy Beijing Air Quality Monitor. It has reported $PM_{2.5}$ particulates pollution since 2008. However, $PM_{2.5}$ was not regulated during 2001-2010. Some research institutions collected air quality data independently but the data are not widely available.

[5]Moral hazard arises where there is a cut-off in the performance evaluation and the information is asymmetric. Manipulation around the cut-off is not a unique phenomenon in economics. For example, lenders treat mortgage borrowers with credit scores just above certain thresholds differently from those with slightly lower scores. Borrowers may take advantage of this rule of screening by manipulating their FICO scores (Keys et al., 2010).

concentration (Brannlund and Lofgren, 1996). Thus, detection of this type of manipulation boils down to a test of discontinuity around the cutoff for blue-sky days in the distribution of pollutant concentrations. Irregularity around the cut-off is a red flag of potential manipulation.[6]

Second, we study the patterns of manipulation by proposing a panel matching approach. The ideal experiment to examine such patterns is to observe twin cities that are expected to have identical distributions of air quality, and hence of the API. The panel matching approach constructs pairs of cities that have the same geographic and provincial characteristics. Since true air quality is unobservable, we use visibility as a proxy together with other weather variables (Sloane and White, 1986). The key assumption that allows us to identify manipulation is that for a constructed city pair, we expect that the two cities have the same distribution of API conditional on visibility and other weather variables. This approach does not pin down the manipulator.[7] Instead, it identifies the conditions under which manipulation is most likely to occur.

This paper is not the first attempt to investigate dubious air pollution data in China. Andrews (2008a,b) first questioned the credibility of official data in Beijing and brought this issue to public attention. The author presented the evidence that the API has massive bunching below the cut-off in addition to other gimmicks to polish air quality reports. Chen et al. (2012) provide a formal econometric analysis on the accuracy of the air pollution data. They confirmed the anomaly around the cut-off based on the official data from 37 large cities during 2000-2009. We improve on this literature in three aspects.

[6]As noted above, command-and-control policies can also be used to ensure that the target number of blue-sky days is met. Such behavior can also lead to a bunching below the cutoff for blue-sky days.

[7]Inspection of the empirical CDFs can suggest which city is the manipulator, but the formal tests we perform do not answer this question.

First, we use a more comprehensive data set. In particular, we obtained the confidential daily air pollution data from the China National Environmental Monitoring Center (CNEMC). The data include non-disclosed variables underlying the calculation of the API. Our data set covers 113 cities during 2001-2010, which includes all cities that are required to disclose daily air pollution information. The detailed data have never been used in previous studies.

Second, our data set allows us to apply the discontinuity test to pollutant concentrations directly instead of the API. The distribution of pollutant concentrations satisfies the continuity assumption of the McCrary (2008) test, hereinafter the McCrary test, whereas that of the API does not. The API is a nonlinear transformation of pollutant concentrations. Its distribution is not continuous, which violates the assumption required for the McCrary test. Applying the McCrary test to the API scores directly may lead to biased results.

Third, our panel matching method is novel. The nonparametric specification is superior to a linear model because the former does not assume that manipulation and control variables such as visibility and weather conditions are separable, which allows for more realistic forms of manipulation which differ based on weather conditions. Furthermore, linear fixed effects models are known to be inconsistent when they are misspecified as shown in Chernozhukov et al. (2010) and Gibbons, Serrato, and Urbancic (2011). Since we do not expect the true relationship between API and other variables to be linear or to follow a particular functional form assumption, the nonparametric approach is preferable.

Our primary finding is that 55% of cities report dubious particulate matter (PM_{10}) pollution data. In comparison, the data for sulfur dioxide (SO_2) and nitrogen dioxide (NO_2) are much more trustworthy. This is not surprising, since PM_{10} is the primary culprit of non-blue-sky days in the majority of Chinese cities. In

terms of the circumstances under which manipulation occurs, we find that dubious data tend to occur on days with high visibility and low wind speed. When wind speed is low, nature is not doing its part and the pollutants are not simply "gone with the wind." On days with high visibility, manipulation is not easily detectable. These findings are very intuitive and illustrate the potential manipulation behavior of local governments. It is important to note that the two methods we use to detect manipulation may be applied by the monitoring agencies themselves to detect manipulators. Although we focus on air pollution, the methodology extends to other situations where the reported data are subject to manipulation due to the presence of a particular cutoff.

The rest of the paper is organized as follows. Section 2 gives an overview of the empirical setting. Section 3 describes the data and variables. Section 4 uses a discontinuity test to detect irregularities around the cut-off. Section 5 proposes the panel matching approach to identify the pattern of manipulation. Section 6 presents robustness checks and further discussions, and Section 7 concludes. Additional estimates and robustness checks are included in the online Appendix.

4.2 Empirical Background

This section introduces air pollution and its regulation in China. We focus on the institutional background on why some Chinese cities may engage in data manipulation.

4.2.1 Air Pollution and Regulation

Air pollution is a critical environmental concern in China. Asia Development Bank reports that less than 1% of the largest Chinese cities meet the air

quality standards recommended by the World Health Organization (Zhang and Crooks, 2012). With significant efforts to curb pollution, some pollutant emissions have slowed down and have not reached their peaks in recent years. For example, total sulfur dioxide emissions peaked in 2006 at 25.89 million metric tons, when the per-capita GDP was $2,064 USD in current prices.[8] Nonetheless, China is still the world's largest emitter of many air pollutants including sulfur dioxide and nitrogen oxide. Poor air quality is a result of rapid economic growth that heavily relies on fossil fuel consumption (Zhang and Wang, 2011). Coal accounts for about 70% of total energy use, which has led to severe SO_2, NO_2, and particulate matter pollution. In addition, motor vehicle usage has grown dramatically, since private car ownership has increased from 3.43 million in 2002 to 78.72 million in 2011.[9] Automotive consumption of gasoline has become a major source of air pollution in big cities.

Severe air pollution has caused tremendous health, economic, environmental, and social problems. Because of SO_2 and NO_2 emissions, acid rain occurred in 227 cities in 2011, or about half of all the monitored cities.[10] Although particulate-matter pollution has improved significantly since 2005, its concentrations are still five times higher than the safety level. A World Health Organization (WHO) report estimates that each year 656,000 premature deaths of Chinese citizens are attributed to the diseases triggered by air pollution.[11] Another recent study suggests that the welfare loss caused by ozone and particulate-matter pollution in 2005 is about 112 billion of 1997 U.S. dollars (Matus et al., 2012). The monetized health

[8]National Bureau of Statistics, 2012. China Statistical Yearbook. http://www.stats.gov.cn/tjsj/ndsj/ (Retrieved on Oct 7, 2012).

[9]National Bureau of Statistics, 2003-2011. Statistical Communique on the National Economic and Social Development. http://www.stats.gov.cn/tjgb/ (Retrieved on Oct 7, 2012).

[10]Ministry of Environmental Protection. 2012. "Report on the State of the Environment in China in 2011." http://jcs.mep.gov.cn/hjzl/zkgb/.

[11]Kevin Holden Platt. July 9, 2007. Chinese Air Pollution Deadliest in World, Report Says. http://news.nationalgeographic.com/news/2007/07/070709-china-pollution.html.

costs of air pollution alone are estimated to be between 1.2% and 3.8% of GDP (World Bank, 2007).[12] In addition, pollution has stirred widespread discontent among the emerging middle class in urban areas, resulting in what the Chinese government defines as "mass incidents." These mass incidents have threatened social stability that is regarded as a top priority for the Chinese central government.[13] They have also created bottom-up pressures for local governments to clean up the environment.

In the wake of serious air pollution, China has been constructing a national system of atmospheric air pollution standards since 1982 (Natural Resources Defense Council, 2009). As a national strategy to improve ambient air quality, 113 key cities are required to disclose their once classified air quality data. The mandate began with weekly reporting in 1998 and advanced to daily reporting in 2000. The central government uses information disclosure to create an incentive for local governments to engage in air pollution reduction more actively. Disclosed air pollution data are used not only to inform the public but also to evaluate city officials' environmental performance.

However, China's air pollution regulations have faced a fair amount of critique: the regulations are relatively lax compared to the standards recommended by WHO or those adopted by other developed countries; certain pollutants are not included; and in some cases the standards have been revised downward to increase compliance. The most pronounced case took place in the 2000 revision. In response to non-compliance due to the increase in automobile usage, the regulator removed NO_x from the list of the criteria pollutants. The standards for NO_2 and ozone

[12]The World Bank used the adjusted human capital (AHC) approach to estimate the forgone earnings due to pollution at 1.2% of GDP. They used the value of a statistical life (VSL) approach to estimate the mortality risks at 3.8% of GDP.

[13]Junjie Zhang. November 9th, 2012. How Will China's New Leaders Approach Rising Tide of Environmental Protests? http://asiasociety.org/blog/asia/how-will-chinas-new-leaders-approach-rising-tide-environmental-protests.

(O_3) were lowered as well. Due to the lax standard, ambient NO_2 concentrations are seldom considered the primary air pollutant. Another important case is $PM_{2.5}$, a fine particulate with major health consequences, which was not included in the standards until 2012. Fortunately, during the period of our study, 2001-2010, the standards remained consistent.

4.2.2 Costly Action vs. Effortless Perfection

Although China has a relatively comprehensive system of air pollution regulation, implementation of the standards at the local level is a major problem. In order to motivate local officials to reduce pollution, environmental compliance has entered the cadre promotion system.[14] Specifically, 113 key cities have been ranked in the annual Quantified Assessment of Urban Environmental Improvement (*Chenkao*) since 1989. Air quality is the single most important indicator in the assessment, which accounts for 20 percent of a city's environmental quality grade. During the 11th Five-Year Plan period (2006-2010), a city with automated monitoring systems receives 20 points if the annual count of blue-sky days is greater than 85% of a year and 0 points if the share is less than 30%. In other cases, the city's grade is determined by $20 \times (p - 30\%)/55\%$, where p is the proportion of blue-sky days.[15] Therefore, city officials strive to achieve 85% of blue-sky days in a year in order to obtain a full score on air quality.

Local officials are expected to comply with the environmental standards because their prospects for career advancement are linked to their ability to meet

[14]Ian Johnson. June 3, 2011. China Faces "Very Grave" Environmental Situation, Officials Say. http://www.nytimes.com/2011/06/04/world/asia/04china.html (Retrieved on November 29, 2012).

[15]More details about the *Chenkao* (in Chinese) are available at: http://wfs.mep.gov.cn/chengkao/ckzb/200612/P020061229356985550756.pdf (Retrieved on February 6, 2013).

the targets set by the higher-level offices. In addition, local officials compete with each other on observable performance measures including economic output and social stability, creating a promotion tournament (Chen, Li, and Zhou, 2005; Li and Zhou, 2005; Shih, Adolph, and Liu, 2012). The rankings of environmental performance by the *Chenkao* were intended to award the title of "Environmental Protection Model City" to top performing cities. The yardstick competition that it creates among city mayors was intended to improve the environment. Zheng et al. (2013) provide empirical evidence that Chinese mayors' likelihood of promotion is affected by both economic growth rate and environmental performance among other things.

However, some performance indicators are difficult to monitor and verify. The central government must often rely on data that are self-reported by local governments. Under asymmetric information, local officials have an incentive to use inappropriate behavior if their interests are not aligned with those who grant promotions. In a worst-case scenario, those officials may engage in data manipulation. Credibility of official statistical data in China has already been under international scrutiny, most notably the overstated economic growth rate (Rawski, 2001; Koech and Wang, 2012). This has led to similar concerns about the integrity of the environmental data.

Although environmental compliance has been explicitly written into the contract between the central and local governments, economic growth is still regarded as the top priority in China (Zhang, 2012). Local officials have unparalleled enthusiasm for growing the economy because of the dual incentives of financial rewards and political futures. This is consistent with the argument that authoritarian leaders opt for less environmental goods in return for faster economic growth (Congleton, 1992). Local governments might lower environmental standards in order

to appeal to investors and raise competitiveness, creating a "race to the bottom." Even worse, taking advantage of the asymmetric information between the central and local governments, self-interested local officials might overstate economic achievement and understate environmental pollution.

4.3 Data Description

We have assembled a unique data set for the empirical analysis, which integrates a confidential air pollution data set with visibility and other weather conditions.

4.3.1 Daily Air Pollution

The air pollution data are provided by the China National Environmental Monitoring Center (CNEMC), which is affiliated with the Ministry of Environmental Protection of China. Note that CNEMC faithfully compiled the air pollution data reported by the local governments during 2001-2010. Since CNEMC is neutral with respect to local interests, we are relatively certain that any anomalies in the data are attributable to the local governments.[16]

The air pollution data contain two parts. The first part is public, which includes daily API score and primary pollutant. The main shortfall of the public data is that pollutant concentrations are not reported. Although the primary pollutant concentration can be inferred from its API score, non-primary pollutants' information is unknown. In addition, the primary pollutant is not reported in a non-pollution day (API $\leq$ 100). Fortunately, we have obtained the confidential part that includes concentrations of all three criteria pollutants: PM_{10}, SO_2, and

[16]The judgment is based on personal communication with officials from CNEMC.

NO_2. The confidential pollutant concentration information has never been used in previous studies.

The air quality data cover 113 cities from 2001 to 2010. The spatial distributions of air pollution in terms of API and three criteria pollutant concentrations are illustrated in Figure A.2. The spatially interpolated air pollution levels shown by the filled contour plots are generated by inverse distance weighting approach based on the city-level daily air pollution data. Air pollution, particularly PM_{10} pollution, is generally worse in North and Northwest China. It is caused by a combination of pollution, geographic and meteorological conditions.

The summary statistics are reported in Table A.12. PM_{10} was the dominant primary pollutant in all cities, responsible for 73.7% of non-blue-sky days (API $\geq$ 100). SO_2 caused less than 10% of non-blue-sky days. NO_2 was almost never responsible for non-attainment because of its lax standard. The mean API is 76.32, implying the average air quality meets the requirement of blue-sky days. On average, blue-sky days account for 84.6% days in the last decade, which interestingly coincides with the target set by the central government. Days with Grade II air quality accounted for 67.7%, that is, most days had good air quality with moderate health consequences.

4.3.2 Air Pollution Index (API)

Air quality is reported in the form of both pollutant concentrations and Air Pollution Index (API). API converts concentrations of three criteria air pollutants (PM_{10}, SO_2, and NO_2) to a single index by a set of piece-wise linear transformations. The index I for one pollutant with concentration C is defined as the linear

interpolation between two index classes such that

$$I = \frac{I_u - I_l}{C_u - C_l}(C - C_l) + I_l. \tag{4.1}$$

In this form, C_u and C_l are the upper and lower boundaries of concentrations for each air quality level, and I_u and I_l are the corresponding upper and lower index classes. Although nine air pollutants were regulated during this period, only three pollutants enter the daily air pollution report system because of technical and cost constraints. The normalized index for each pollutant is computed based on its daily average concentration.

API on a given day is determined by the pollutant that has the highest index. The corresponding air pollutant is referred to as the primary pollutant.

$$API = \max\{I_{\mathrm{SO}_2}, I_{\mathrm{NO}_2}, I_{\mathrm{PM}_{10}}\}. \tag{4.2}$$

API varies between 0 and 500 with a large number indicating poor air quality. Different API categories are associated with different pollution levels and health consequences. Officially, a "blue-sky day" is defined as a day with the value of API less than 100, that is, the air quality is either excellent or good.[17] The compliance with the air quality standards is then simplified by just counting the number of blue-sky days.

Pollutant concentrations are measured and averaged across stations and over a 24-hour period. In order to release pollution information in the afternoon, daily report uses the data from the previous noon to current noon. To summarize, API calculation is implemented in four steps: First, a 24-hour average pollutant

[17]Note that a blue-sky day is a just technical definition. It does not necessarily mean that the sky is literally blue.

concentration is calculated for each station. Secondly, city average pollutant concentration is derived from multiple station averages. Thirdly, individual pollutant index I is calculated according to equation (4.1). And finally, API is the maximum of individual pollutant indices according to equation (4.2).

Pollutant concentrations are measured and averaged across stations and over a 24-hour period. In order to release pollution information in the afternoon, daily report uses the data from the previous noon to current noon. To summarize, API calculation is implemented in four steps: First, a 24-hour average pollutant concentration is calculated for each station. Secondly, city average pollutant concentration is derived from multiple station averages. Thirdly, individual pollutant index I is calculated according to equation (4.1). And finally, API is the max of individual pollutant index according to equation (4.2).

Manipulation happens in the process of calculating daily average pollutant concentrations at station- or city-level. It could be caused by data falsification, which is against the law. It could also be caused by regulatory loopholes. Specifically, the minimum data requirement to calculate daily averages for gaseous pollutants (SO_2 and NO_2) is 18 hours of effective monitoring and that for particulate matter is 12 hours.[18] This creates an incentive for cities to discard the observations with bad pollution on the excuse of faulty equipment. In addition, since the data requirement for PM_{10} is lower than the other two pollutants, this becomes another reason why PM_{10} is more vulnerable to manipulation besides PM_{10} being the dominant primary pollutant.

[18] Automated Methods for Ambient Air Quality Monitoring HJ/T193-2005. http://www.zhb.gov.cn/image20010518/5523.pdf

4.3.3 Visibility and Other Weather Variables

The meteorological data are obtained from the National Climatic Data Center under the National Oceanic and Atmospheric Administration (NOAA) of the United States. The weather data are collected by the weather stations under the China Meteorological Administration (CMA). Since weather stations are not prone to political interference, the human operators of weather stations do not have an incentive to manipulate the results. This data set that we use records weather information for 499 weather stations every three hours from 2001-2010. Note that we dropped the weather stations that have less than 10,000 records.

Meteorological factors are correlated with air pollution. The variables that we use in the paper include visibility (VSB, in statute miles), temperature (TEMP, in Fahrenheit), atmospheric pressure (STP, in millibars), precipitation (PCP, in inches), and wind speed (SPD, in miles per hour). The weather variable of central interest is visibility, which is used as a proxy for API. Visibility is historically defined as "the greatest distance at which an observer can just see a black object viewed against the horizon sky" (Malm, 1999). It has been shown that particulate matter and gaseous pollution can cause visibility impairment. All these variables are daily averages in order to match the time scale of API data.

4.4 Irregularities at the Cutoff for Blue-Sky Days

One expects that manipulation around the cutoff for blue-sky days is the most likely form of manipulation to occur. The reasoning here is straightforward. Local governments report their API scores on a daily basis and this data set is publicly available. If the API scores are tweaked by large amounts, citizens and central government officials may doubt the information reported by local govern-

ments. Manipulation right around the cut-off is less likely to be detected because the difference in air quality between API values at 100^- and 100^+ may be indiscernible. Hence, we may rightly predict that cities manipulate blue-sky days by examining the discontinuity in the probability density function (pdf) of air quality data around the cutoff.

4.4.1 API vs Concentration

McCrary (2008) proposes a test for the manipulation of the running variable in regression discontinuity design. An implication of the manipulation of the running variable is that its pdf would have a discontinuity. The main assumption for the validity of this test is the continuity of the pdf of the underlying variable under the null hypothesis of no manipulation. Polluters and regulators have imprecise control of the waste load output because it is subject to random shocks (Brannlund and Lofgren, 1996). The pdfs of pollutant concentrations are hence expected to be continuous. The pdf of the API, on the other hand, is not continuous because it is a piece-wise linear transformation of the underlying pollutant concentrations. Figure A.3 illustrates how the relationship between pollutant concentration and API has kinks at some boundaries. This is why we apply this test to the pollutant concentrations. Thus, detection of manipulation boils down to a test of whether the pdf of a pollutant concentration exhibits a discontinuity around the cut-off for blue-sky days.

In the rest of the section, we show how the piece-wise linear transformation may lead to discontinuities in the pdf of API. According to equation (4.1), the relationship between pollution index I and pollutant concentration C can be

simplified as the following linear function:

$$I = k_1 C + k_2, \tag{4.3}$$

where $k_1 = (I_u - I_l)/(C_u - C_l)$ and $k_2 = I_l - C_l(I_u - I_l)/(C_u - C_l)$. In this form, pollutant concentration C is a continuous random variable and index I is piece-wise linear in C.

The cumulative distribution function (cdf) for a pollutant concentration is denoted by $F_C()$. The cdf for the corresponding pollution index $F_I()$ is then:

$$F_I(x) = Pr\{I(C) \leq x\} = Pr\{k_1 C_{it} + k_2 \leq x\} = F_C\left(\frac{x - k_2}{k_1}\right). \tag{4.4}$$

Because k_1 and k_2 depend on pollution index classes, the cdf of pollution index is only differentiable piece-wisely. We can derive the pdf for the pollution index around a threshold x_c and examine the difference in probability densities between:

$$f_I(x_c^-) = \frac{1}{k_1^-} f_C\left(\frac{x^- - k_2^-}{k_1^-}\right) \text{ and } f_I(x_c^+) = \frac{1}{k_1^+} f_C\left(\frac{x^+ - k_2^+}{k_1^+}\right). \tag{4.5}$$

Equation (4.5) shows that discontinuity can occur even if the pollution data are faithfully reported. The discontinuity can be caused by the piece-wise linear transformation of pollution concentration. Let's take SO_2 as an example. The probability densities of the pollution index around 100 are $f_I(100^-) = 0.002 f_C(0.15)$ and $f_I(100^+) = 0.006 f_C(0.15)$ respectively. It is apparent that the right density is higher than the left density but this is not attributed to manipulation. The same situation is also true for SO_2. However, the pollution index for PM_{10} should be continuous around the 100 cut-off (see Figure A.3). Besides that the normalized pollution index is a piece-wise linear transformation of con-

centration, API is the maximum of three pollutant indexes according to equation (4.2), which also leads to kinks at the cut-off. Therefore, applying the McCrary test to API directly will yield inconsistent results. Utilizing the confidential air quality data, we apply the discontinuity test to pollutant concentrations instead of API since the former is expected to have a continuous pdf in the absence of manipulation.

4.4.2 Test of Discontinuity

Since PM_{10} is the primary culprit for non-blue-sky days for the majority of Chinese cities, the main component here is the application of the McCrary test to the PM_{10} concentration. Since SO_2 and particularly NO_2 seldom cause non-blue-sky days, manipulation of these two pollutants increase the risk of being caught without improving the environmental record of local officials. It is reasonable to assume that their pollution levels should be more trustworthy. Therefore, we also apply the test to the SO_2 and NO_2 concentrations as part of our robustness checks.

The test statistic proposed by McCrary (2008) is an estimator of the log difference in height between the left and right limit of the density of the variable of interest at the cut-off, c:

$$\theta = \ln \lim_{r \downarrow c} f(r) - \ln \lim_{r \uparrow c} f(r) = \ln f^{+} - \ln f^{-}, \tag{4.6}$$

where f denotes the density of variable r, and f^{+} and f^{-} denote the right and left limit, respectively. In our case, r is the pollutant concentration and c is the cutoff for API = 100. The estimator $\hat{\theta}$ is given in the appendix. It is asymptotically

normal:

$$\sqrt{nh}(\hat{\theta}-\theta) \xrightarrow{d} N\left(B, \frac{24}{5}\left(\frac{1}{f^+}+\frac{1}{f^-}\right)\right)$$
$$B = \frac{H}{20}\left(\frac{-f^{+"}}{f^+}-\frac{-f^{-"}}{f^-}\right),$$

where h is the bandwidth and $H = \lim_{n\to\infty,h\to 0} h^2\sqrt{nh}$. We perform the one-sided lower-tailed version of the test, since we are interested in the shifting of probability mass from above the cut-off to below it, which will yield the left limit, f^- to be higher than the right limit, f^+.

The test statistic is calculated in two steps. First, using the estimated variance of the data, a bin size, b, is chosen to discretize the data and plot the first-step histogram. After that, the discretized data is used to estimate the left and right limit of the pdf at the cut-off using a bandwidth h. McCrary (2008) recommends that the ratio of bandwidth and bin size $a \equiv h/b$ shall be greater than 10. We use $a = 15$ as the benchmark. See Section 4.9 for the estimation process.

We use the t-statistics to infer whether a city is a manipulator. Because manipulation involves under-reporting of pollution, it causes bunching below the cut-off. Therefore, a significantly negative t-statistic is an indicator of manipulation. We also use the t-statistics to compare the degree of manipulation across cities. We rank cities by the significance of the discontinuity. The t-statistic is normalized by its variance and hence is more comparable than the actual magnitude of the discontinuity that depends on the shape of the pdf.[19] It is however important to note that a larger t-statistic does not necessarily imply a higher level

[19]To make this point clear, note that for a cdf we all understand what a discontinuity of 0.1 means. However, the McCrary statistic is the difference between the logs of the left and right limit and hence is not easy to compare if the pdfs are different.

of manipulation in the sense of a larger discontinuity in the pdf. Rather, it signifies a higher degree of confidence in the presence of manipulation.

4.4.3 Baseline Results

We apply the McCrary test to each city using 10 years of daily pollutant concentration data. PM_{10} is the dominant pollutant in all cities, which accounts for 74% of non-blue-sky days. Hence, we expect it to be manipulated by a larger number of cities. Our baseline result, the t-statistic of the McCrary test using $a = 15$, is illustrated in Figure A.4. It shows different degrees of manipulation in the PM_{10} data across cities. We categorize three levels of manipulation based on t-statistics: above -1.5 as non-manipulators, between -1.5 and -2 as borderline manipulators, and smaller than -2 as manipulators. A visual observation of this graph does not reveal obvious spatial patterns of manipulation.

The city-specific McCrary test result confirms heterogeneous manipulation behavior. Table A.13 exhibits the cities that report dubious PM_{10} pollution. These cities are flagged because their pdfs of PM_{10} concentrations exhibit a statistically significant discontinuity around the cut-off. More specifically, there is a bunching below the cut-off. The baseline result, the column with $a = 15$, suggests that 61 cities reported dubious PM_{10} pollution data in the last decade. Please note that our estimation results are relatively conservative. Fifty cities show no signs of manipulation in the McCrary test.

In the introduction section, we use the API histograms of four municipalities to motivate the research question (see Figure A.1). However, our formal empirical test uses pollutant concentration instead of API. The McCrary test is illustrated in Figure A.5. The results confirm that Beijing has the highest likelihood of manipulation, followed by Tianjin and Chongqing. However, we cannot reject the

null hypothesis that Shanghai is not a manipulator.

To summarize the McCrary test results, we plot McCrary t-statistics for the three criteria pollutants against GDP per capita, population density, and value added of the secondary industry per capita in Figure A.6. These plots show that for PM_{10} the relationship between the McCrary t-statistic and the three economic variables is negative. Since a more negative McCrary t-statistic implies more significant manipulation, a negative slope implies that the higher the GDP per capita, population density, or industrial value added per capita, the more significant the manipulation of PM_{10} concentration data. This correlation is intuitive, since cities that are larger demographically or economically are more likely to have pollution problems due to PM_{10} and hence are more likely to manipulate their data. For SO_2 and NO_2, we find that there is no correlation between the relevant McCrary t-statistics and the economic and demographic variables.

4.5 Patterns of Manipulation

The McCrary test does not inform under what situations cities may report dubious air pollution data. To address this issue, we propose an alternative identification strategy to investigate the patterns of manipulation. The identification relies on the fact that two cities with identical air quality should have the same API. Otherwise, the discrepancy is attributed to manipulation. This motivates our panel matching approach.

4.5.1 Identification Strategy

In order to detect particular patterns of manipulation, we cannot simply look at one city. The ideal experiment in this setup would be to have twin cities in

the sense that they have the same distribution of true air quality. In the absence of manipulation, these twin cities, city 1 and 2, have identical distributions of API. This implies that to test whether one of the two cities manipulate, one can test the following hypothesis:

$$H_0 \quad : \quad API_{t1} \stackrel{d}{=} API_{t2}. \tag{4.7}$$

In this form, $\stackrel{d}{=}$ denotes equality of distribution. Since in practice we do not have twin cities, we have to form city pairs that are as close as possible in terms of true air quality.

We utilize visibility as a proxy for true air quality. Visibility measures the distance at which an object can just be discerned against a light sky (Sloane and White, 1986). Visibility degradation is closely related to air pollution because sunlight is absorbed or scattered by pollution particles in the air (Guo et al., 2009). Since visibility is reported by weather stations that are not prone to political pressures, it can serve as a reliable proxy for air pollution. However, the visibility-pollution connection is also affected by natural variations such as humidity. Therefore, we control for other weather variables including wind speed, temperature, and precipitation. These weather variables are perceived as exogenous, since they are determined by "nature" outside the economy, or for our purposes the political economy.

Our identification strategy is described as follows. Air quality is unobservable; however, it affects both API and visibility. API is determined by both true air quality and manipulation, where the latter is the latent variable to be identified. Visibility is determined by true air quality and exogenous shocks by "nature." "Nature" is partially observable such as weather. For the unobservable "nature,"

we spatially cluster cities since neighboring cities share some common geographical characteristics that affect the visibility-pollution connection.

Air quality can be proxied by visibility, weather conditions, and geographical location. This allows us to compare the distributions of API of two cities in a pair on days where both face the same visibility and other weather conditions. So our revised hypothesis is:

$$H_0 \quad : \quad API_{t1}|W_{t1} = w \stackrel{d}{=} API_{t2}|W_{t2} = w. \tag{4.8}$$

In this form, W designates the vector of visibility and other weather variables, and w designates a value that W takes. The above hypothesis suggests that two cities have identical distributions of API conditional on visibility, weather, and geographical locations.

4.5.2 Why Not a Linear Model?

Now one implication of our hypothesis of no manipulation in equation (4.8) is that the conditional mean of API is equal for the two cities on days with the same weather variables:

$$\text{Eq. (4.8)} \quad \Rightarrow \quad E[API_{t1}|W_{t1} = w] = E[API_{t2}|W_{t2} = w]. \tag{4.9}$$

Now we can impose a linear model on the relationship between API and weather variables, as follows:

$$API_{ti} = W'_{ti}\beta + \alpha_i + \epsilon_{ti}, \tag{4.10}$$

for city $i = 1, 2$. In this situation, the testable implication would be:

$$\text{Eq. (4.8) and (4.10)} \quad \Rightarrow \quad \alpha_1 = \alpha_2. \tag{4.11}$$

Hence, the intuition behind our approach here is to test that there is no heterogeneity in the relationship between weather variables and API, once we control for geographic and provincial characteristics. For the linear model, this translates to the equality of the city-specific intercepts (city fixed effects). It is important to note that the type of manipulation that a linear model can detect is very restrictive. This is our motivation for using a very general nonparametric approach that can allow for nonlinear dependence between API and weather variables.

The nonlinear dependence between API and weather variables stems from two main sources: 1) manipulation of API, and 2) the nonlinear relationship between true air quality and weather conditions. First, the type of manipulation that can be modeled linearly only changes the mean of API. It imposes that manipulation leads the mean of API to change by the same amount whatever the weather conditions are. So it rules out a situation where manipulation occurs only under certain weather conditions. The latter is an implication of linearity. More specifically, it is due to the separability of W_{ti} and α_i in (4.10). We suspect that manipulation may occur without changing the mean of API. Furthermore, it is more likely to occur under weather conditions that make it harder to detect manipulation. Hence, we do not expect manipulation to be orthogonal to weather conditions. As for the second issue, the relationship between true air quality and weather conditions, especially visibility, is inherently nonlinear. For instance, visibility is a censored variable since its measure is limited to 7 or 10 miles. Hence, we expect its relationship with true air quality to be nonlinear. Given the fact

that API is a nonlinear transformation of measures of true air quality, this further strengthens the argument for using a general nonlinear approach to model the relationship between API and weather variables that imposes no parametric restrictions. This is one of the primary motivations for us to use a fully nonparametric approach here.

4.5.3 Panel Matching Approach

Following the above identification strategy, we propose a panel matching approach to study the pattern of manipulation. Now we formalize our above discussion and show the key assumptions that lead to our hypothesis. For city pair k with cities $i = 1, 2$, we have the following relationship,[20]

$$API_{tki} = \xi_k(W_{tki}, \mathcal{A}_k, \mathcal{U}_{tki}), \tag{4.12}$$

where API_{tki} is the API score on day t of city i which belongs to city pair k, W_{tki} are weather variables including visibility, wind speed, temperature, and precipitation, $\mathcal{A}_k$ are unobservable factors that are time-invariant and are specific to a city-pair, and $\mathcal{U}_{tki}$ represents idiosyncratic shocks. Now our key identifying assumption here is that:

$$\mathcal{U}_{tki}|W_{tk1} = w, \mathcal{A}_k = a \overset{d}{=} \mathcal{U}_{\tau k2}|W_{\tau k2} = w, \mathcal{A}_k = a, \tag{4.13}$$

where a is a realized value for the unobservable city-pair attribute $\mathcal{A}_k$. Please note that t is not necessarily equal to τ.[21]

[20] Note that (4.12) is not a structural relationship per se.

[21] This is not a restriction, because the two cities do not have to face the same weather conditions on the same day. We just need to compare days with the same weather conditions, not the same days with the same weather conditions. We do this mainly because of constraints on sample size.

Equation (4.13) is a homogeneity assumption, such as the one made in Chernozhukov et al. (2013) and Chapter 3 of this dissertation, where similar assumptions are used to identify average partial effects in nonseparable panel models. The assumption is employed here to test the existence of manipulation. More importantly, the content of (4.13) is that once we control for weather conditions and unobservable factors specific to the city-pair, other unobservable factors should have the same distribution across the two cities.

Now (4.12) and (4.13) imply our testable hypothesis from above:

$$H_0 \quad : \quad API_{tk1}|W_{tk1} = w \overset{d}{=} API_{\tau k2}|W_{\tau k2} = w. \tag{4.14}$$

We are essentially matching days based on weather conditions. On days where cities 1 and 2 in pair k face the same weather conditions, their API should have the same distribution.

In order to test the equality of distribution, we apply two tests, the KS and the t-test. The two-sample Kolmogorov-Smirnov test is a natural statistic in this setup since it detects any deviation from the equality of distributions. However, it may be conservative, when we do not have two independent samples with *i.i.d.* observations. We then run the t-test for the equality of means, which is robust to deviations from the *i.i.d.* assumptions. However, it only tests one implication of the equality of distributions, which is the equality of means.

Now for every city-pair, we test the equality of the distribution of API on days with similar weather variables including precipitation, wind speed, temperature, and visibility. We discretize our weather variables and implement the KS-test on each possible combination of weather variables.[22] One can motivate

[22]For details on discretization, see Section 4.11.

the approach here as an extension of matching methods, where we compare the distribution of API on days where the two cities face similar weather conditions. Recall that in propensity score matching, one matches individuals according to the propensity score, i.e. the predicted probability of treatment. In this paper, we match days based on weather conditions. Then, we test the equality of the API distributions of the two cities for those days.

Since we apply the same tests for all different weather combinations for every city pair, we correct for multiple testing using Romano and Shaikh (2006). For details on the specific procedure that we use, please see Appendix 4.10.

4.5.4 Baseline Results

First of all, we need to form city pairs before using the panel matching method. It is implemented in two steps. First, we find nearest neighbors in terms of geographical distance for each city, and define a candidate city pair if both cities are each other's nearest neighbors.[23] Then, we remove all candidate city pairs that are not in the same province to ensure that each city pair is faced with the same provincial environmental goals. Our method results in 14 city pairs. We discard one of these pairs, Kelamayi-Wulumuqi, since the geographical distance between them is quite large. Hence, we are then left with 13 city pairs.

Tables SA.1-SA.15 in the Supplementary Appendix show the results for our panel matching approach.[24] Among the 13 city pairs that we examine, we have four city pairs only that do not show any evidence of manipulation, specifically Wuhu-Maanshan, Zhenjiang-Yangzhou, Changzhou-Wuxi and Jilin-Changchun. We find at least some rejections for most city pairs, including Kaifeng-Zhengzhou,

[23]Nearest neighbor matching does not result in unique matches, this is why we impose this condition.

[24]The supplementary appendix is available upon request.

Quanzhou-Xiamen, Hangzhou-Shaoxing, Shenyang-Fushun, Yinchuan-Shizuishan, Xian-Xianyang, and Zhuzhou-Xiangtan.

With the exception of Xian-Xianyang, the rejections seem to occur mostly for higher levels of visibility and low wind speed. This is intuitive for two reasons. First, since poor visibility is associated with high levels of pollution, it is harder to detect manipulation. This is an important concern for the local governments because the API data are published on a daily basis and the citizens can detect whether it is reasonable to think that a particular day is a blue-sky day or not. Secondly, it is more likely for manipulation to occur when it can make a difference, i.e. when it would turn a pollution day to a blue-sky day. In addition, to make it less detectable, we are more likely to see that manipulation occurs closer to the cut-off, i.e. the pollution levels should not be that severe. This again confirms our intuition that manipulation is more likely to occur with higher levels of visibility. It is also intuitive why manipulation would occur when wind speed is low. Note that if wind speed is high, the pollutants could be "gone with the wind." However, if wind speed is low, then nature is not helping reduce pollution. As a result, cities manipulate their API to meet the target.

It is important to note here that the fact that we find manipulation occurs for certain weather conditions but not others indicates that a linear specification is not appropriate. Recall from above, that a linear specification implies that manipulation is orthogonal to the weather conditions. Our results indicate that this is not the case. Furthermore, they illustrate that the measurement error resulting from manipulation is expected to violate the classical measurement-error assumptions.

To illustrate the formal results in Tables SA.1-SA.15 of the Supplementary Appendix and to gain some intuition for the approach, we also include Figures

SA.1-SA.13. Figure A.7 is an example of illustration for the city pair Zhejiang and Yangzhou. Each figure contains 6 plots for different weather variable combinations for each city pair. VSB denotes visibility, WSP wind speed, and TEMP denotes temperature. All figures are for precipitation equal to zero after discretization, see Section 4.11. Each plot includes two empirical cumulative distribution functions (cdfs) in solid lines, one for each city in a pair. We also include point-wise 95% confidence bands for each empirical cdf using dotted lines. It is important to note that these confidence bands are only to be used as a reference point and may not be interpreted formally, since the significance level that is used for the formal test is adjusted to correct for multiple testing. Furthermore, since we are comparing two functions, we ought to use uniform confidence bands, which would be wider than the point-wise ones. Hence, caution is needed in interpreting the point-wise confidence bands.

Figures SA.1, SA.2, SA.3, and SA.13 show how the city pairs Zhenjiang-Yangzhou, Changzhou-Wuxi, Jilin-Changchun, and Wuhu-Maanshan, respectively, do not reflect manipulation under various weather conditions. This of course gives us confidence in our approach that it can detect the absence of manipulation.

For the other city pairs, we do find manipulation, but what the figures show is that different city pairs manipulate in different ways. In most cases, the empirical cdf of API of the non-manipulating city first-order stochastically dominates that of the manipulating city. The reasoning here is that the manipulating city will tend to shift mass to the lower values of API throughout the entire distribution. For higher levels of visibility, this occurs for Kaifeng-Zhengzhou, Zhuzhou-Xiangtan, Quanzhou-Xiamen, Hangzhou-Shaoxing, Shenyang-Fushun, Jinan-Taian, and Huhehaote-Baotou. It is also important to note that the degree of manipulation, which is represented graphically by the vertical distance between the two cdfs, may

be very different according to the weather conditions. For instance, for Huhehaote-Baotou, when wind speed and temperature is low, manipulation is much more severe than with higher temperature and wind speed.

The manipulation that occurs for city pair Xian-Xianyang, however, does not lead to first-order stochastic dominance. The upper-middle plot in Figure SA.10 shows that the empirical cdf of Xian is flat between 100 and 125, which is clear evidence of manipulation. For this case, it seems that manipulation occurs mostly right around the cut-off.

4.6 Robustness Checks and Discussion

Since we are dealing with observational data, we have to check the robustness of our approach and also discuss potential caveats that may affect how we interpret our results. This section is devoted to this end.

4.6.1 Robustness Checks

First, we investigate the impact of a, the ratio of bandwidth to bin size, on the McCrary test results. According to McCrary (2008), the choice of bin size has no consequences on the test statistic if $a > 10$. Now in order to ensure that the asymptotic approximation delivers correct inference in finite samples, we need h to be fairly small, hence we choose to rank the cities' statistics using $a = 15$. In the robustness checks, we allow different values for a. The McCrary test results with $a = 10$ or 20 are reported alongside with the baseline results in Tables A.13 - A.16. The choice of bandwidth and bin size matters in the test results. It affects not only the significance but also the ranking of the cities in terms of the likelihood of manipulation. The results show that the larger the bandwidth or the smaller

the bin size the more likely it is a city to be tagged as a manipulator. This is a standard issue in nonparametric methods.

The second robustness check is concerned with the manipulation of different pollutants. Since PM_{10} caused 73.7% of the total pollution days, it is the most vulnerable target. We find that 55% of cities reported dubious PM_{10} data. SO_2 and NO_2 account for 9.2% and 0.2% of primary pollutants respectively. There should be less cities manipulating SO_2 and NO_2 concentrations. We report the results for the SO_2 concentrations in Table A.15. We find that , 26 cities, or 23%, are flagged because of the bunching below the cut-off. As for NO_2, we include its results in the Supplementary Appendix. NO_2 seldom leads to API greater than the cut-off for blue-sky days. Hence, there is very little mass above that cut-off to be able to estimate the right limit of the pdf, which makes the results of the McCrary test unreliable.

The third robustness check involves implementing the McCrary test for "artificial" cut-offs. We change the cut-off to $c = 0.1$ and $c = 0.2$ in lieu of the cut-off for blue-sky day. If our hypothesis is true that manipulation leads to a discontinuity, then we do not expect to find any rejection for the artificial cut-offs. We implement the test for both PM_{10} and SO_2 and report the results in the Supplementary Appendix. We do not find evidence of a discontinuity for either of the pollutants.[25]

The fourth robustness check compares the application of the McCrary test to pollutant concentrations and API. We have demonstrated in the identification section that applying the test to API directly might lead to inconsistent estimates because some discontinuities in the API distribution are inevitable because it is a piece-wise linear transformation of pollutant concentration. In order to empirically

[25]There are very few significant results that are driven by little mass above the cut-off which leads to unreliable estimates of the right limit of the pdf.

illustrate this argument, we also apply our test to API and report the implications for our key results in Table A.17. Because there is no kink in the transformation from PM_{10} to API at API = 100, see Figure A.3, the McCrary test shall yield similar results for the concentration and API. The major difference is in SO_2. The concentration model suggests 19 manipulators while the API model only identifies 4 manipulators. The results on NO_2 are close and we attribute it to the fact that the number of observations is small because NO_2 seldom showed up as a primary pollutant.

Fifth, as a robustness check for the panel matching approach, we compare its result with the McCrary result in Table A.18. "YES" implies that the relevant test finds evidence for manipulation, "NO" implies the contrary, and "Borderline" implies that it depends on the bandwidth. If the panel matching approach finds rejections, then this may indicate that either one city in the pair is manipulating or both cities manipulate in different ways. If we do not find rejections, then most likely both cities should not be manipulating. It is however possible, though unlikely, that those cities manipulate in exactly the same way. Finally, we may find manipulation in the panel matching approach but not in the McCrary test. This would be the case, if manipulation behavior does not lead to a discontinuity in the pdf of the pollutant concentrations at the cut-off.

As we expect, we do not find a city pair, where both cities manipulate according to the McCrary test but we find no rejections in the panel matching approach. However, we find three pairs, where only one city in the pair manipulates according to the McCrary test, but our panel matching approach finds no rejections. This is the case for Zhenjiang-Yangzhou, Changzhou-Wuxi and Jilin-Changchun. This may be due to the relatively small subsample size for these three city-pairs. It also may be because the panel matching approach does not test for

discontinuities so by construction it is less powerful at detecting this form of manipulation. This phenomenon may be explained from a political-economy perspective. For instance, it is possible that a city manipulates in a way to keep up with neighboring cities, if local government officials in these cities compete over promotions. It is also important to point out that the KS statistic may be conservative in our setting, since it may violate the classical assumptions under which the KS asymptotic distribution is derived. Since we have a time-series cross-section, there may be time-series dependence that would render the KS statistic conservative.

For the rest of the city-pairs, we find that our panel matching approach and the McCrary test yield consistent results. This is in line with our expectations that both approaches should confirm each other once we exclude some unlikely cases. As noted above, if there is manipulation that does not lead to a discontinuity at the cut-off, the McCrary test cannot detect it. However, the panel matching approach can detect manipulation regardless of the presence of a discontinuity. This is exactly the case for the Hangzhou-Shaoxing city pair. The McCrary test does not indicate manipulation for either city, yet we find evidence for manipulation in the panel matching approach. It is important to note here that for manipulation to lead to a discontinuity in the pdf of the pollutant concentrations there needs to be very strong manipulation around the cut-off. Since both of these cities are coastal, it is conceivable that they only manipulate their data slightly such that one cannot detect a discontinuity in the pdf of their pollutant concentrations.

4.6.2 Further Discussion

It is important to stress that our empirical results are suggestive. What we refer to as manipulation may not be actual manipulation if our assumptions do not

hold. For the McCrary test, the main caveat is that one can only detect the types of manipulation that lead to a discontinuity. For instance, if manipulation leads to a mean shift in the distribution of the concentration, then the test has no power against this alternative by construction. There are other ways of manipulation that may not result in a discontinuity, but we do not find these alternatives very likely in practice. For cities to manipulate without leading to a discontinuity at the cut-off for blue-sky days, they must have knowledge of the distribution of the concentration for the entire period that we are studying. However, cities have to report their data on a daily basis and hence it is rather unlikely that they can manipulate without leading to a discontinuity at the cut-off.

Discontinuity is an alarming signal but clustering of pollutant distribution around the cutoff is not necessarily due to manipulation. Risk-averse polluters might over-comply with the standards and cause bunching of pollution data (Bandyopadhyay and Horowitz, 2006; Earnhart, 2007; Shimshack and Ward, 2008). In our case, in order to achieve a certain number of blue-sky days, some local governments temporarily shut down factories, reduce energy supply, or require firms to use high-quality fuels.[26] These activities will also cause clustered pollutant distributions. Although these short-run policies are not efficient, we cannot label these local officials as manipulators. However, as long as air quality cannot be precisely controlled, clustering should not lead to a discontinuity. Therefore, our discontinuity test is still a valid means of detecting manipulation. In addition, clustering does not affect the relationship between API, visibility and other weather variables. In this case, our panel matching approach is also robust.

The panel matching approach uses the KS test, which does not directly identify which city in a pair manipulates. We do not think this is a downside to

[26]Personal communication with the officials from the Ministry of Environmental Protection of China.

this approach, since our objective here is to learn about patterns of manipulation rather than detecting which city is a manipulator. The McCrary test is more geared toward pointing out manipulators. Hence, we see the two approaches as complements rather than substitutes.

The main caveat of the panel matching approach is that it relies heavily on the assumption that once we condition on weather variables, the *entire* distribution of API should be the same for both cities. This is of course a strong assumption. If we thought that there is still a geographical difference between the two cities that may lead their distribution of API to be different under certain weather conditions, such as high visibility, but not otherwise, then this would confound the results of our panel matching approach. However, the manipulation that we find using the panel matching approach is consistent with the results from the McCrary test. Even though the formal tests of the panel matching approach do not indicate which city manipulates, examining the figures of the empirical cdfs can suggest which city is the manipulator as we discussed above. For instance, Figure SA.4 shows that Kaifeng is the one that moves mass toward the lower values, hence suggesting that it is the manipulator. This is consistent with the McCrary test results. Similar examination of Figures SA.5, SA.6, SA.8 and SA.10 show that the panel matching approach is detecting manipulation.

For the pair of Hangzhou-Shaoxing, we do find very few instances of manipulation using the panel matching approach, even though the McCrary test does not detect manipulation for either city. This is not surprising, because if manipulation is mild, then it may not lead to a discontinuity in the unconditional distribution of the pollutant concentration. Hence, it cannot be detected by the McCrary test, but may be detected by the panel matching approach.

4.7 Conclusion

We find that the daily air pollution data in China are not well-behaved. The assertion is based on the analysis that applies a discontinuity test and a panel matching approach to a unique data set that covers all major cities over a decade. These results are relevant for empirical researchers and policy makers who may use such data to learn about the effects of pollution on various types of outcomes. We suggest that thorough robustness checks need to be done to examine the impact of the cut-offs. Particular attention should be paid to the cities that are on our suspect list. In addition, we find a fair amount of heterogeneity and non-linearity in the data reporting behavior. Thus, the resulting data are unlikely to reflect the classical measurement error assumptions. Our results indicate that the use of standard methods that rely on the classical measurement-error assumptions would not be appropriate. Hence, the use of such data requires caution and care in the choice of estimation strategy and assumptions on the measurement error.

Our methodology can help the monitors ferret out the cities that report dubious data. In particular, we have discovered the meteorological conditions under which local officials are more likely to manipulate. However, the conviction of a manipulator requires an independent direct measure of pollution because ours only provides suggestive evidence. Our results suggest that situations where government officials report data that are used in their own performance evaluation lead to strong incentives for manipulation, as we would expect. Therefore, our models are applicable not only to daily air pollution reports but also to other self-reported data.

This paper has focused on the identification of potential manipulation behavior. Although we have done some preliminary analyses on the patterns of manipulation, we did not provide a political economy interpretation why manipu-

lation is more likely to occur in some cities but not others. This question will be left for future studies.

4.8 Acknowledgement

This chapter is co-authored with Junjie Zhang at the School of International Relations/Pacific Studies, UCSD.

4.9 Appendix: McCrary (2008) Test Statistic

4.9.1 Implementation given b and h

There are two steps to implementing the McCrary (2008) test statistic on a variable Z_i:

1. First-step Histogram of the Discretized X_i

   Using a binsize b,

$$g(Z_i) = \lfloor \frac{Z_i - c}{b} \rfloor b + \frac{b}{2} + c \in \left\{ ..., c - 5\frac{b}{2}, c - 3\frac{b}{2}, c - \frac{b}{2}, c + \frac{b}{2}, c + 5\frac{b}{2}, \right\} \quad (4.15)$$

   where $\lfloor a \rfloor$ is the greatest integer function.

   Now $\{X_j\}_{j=1}^{J}$ is the equi-spaced grid of width b covering the support of $g(Z_i)$ and

$$Y_j = \frac{1}{nb} \sum_{i=1}^{n} 1\{g(R_i) = X_j\} \quad (4.16)$$

   A scatter plot of X_j and Y_j gives the first-step histogram.

2. Calculation of $\hat{\theta}$

$$\begin{aligned}\hat{\theta} &= \ln \hat{f}^{+} - \ln \hat{f}^{-} \\ &= \ln \left\{ \sum_{X_j > c} K\left(\frac{X_j - c}{h}\right) \frac{S^{+}_{n,2} - S^{+}_{n,1}(X_j - c)}{S^{+}_{n,2} S^{+}_{n,0} - (S^{+}_{n,1})^2} Y_j \right\} \\ &\quad - \ln \left\{ \sum_{X_j < c} K\left(\frac{X_j - c}{h}\right) \frac{S^{-}_{n,2} - S^{-}_{n,1}(X_j - c)}{S^{-}_{n,2} S^{-}_{n,0} - (S^{-}_{n,1})^2} Y_j \right\},\end{aligned}$$

where $S^{+}_{n,k} = \sum_{X_j > c} K((X_j - c)/h)(X_j - c)^k$, $S^{-}_{n,k} = \sum_{X_j < c} K((X_j - c)/h)(X_j - c)^k$, and $K(t) = max\{0, 1 - |t|\}$.

4.9.2 Selection of b and h

McCrary (2008) points out that the binsize does not matter provided that $h/b > 10$. We use $\hat{b} = 2\hat{\sigma}n^{-1/2}$ proposed in the bandwidth selection guide, where $\hat{\sigma}$ is the standard deviation of Z_i. In terms of bandwidth selection, McCrary (2008) proposes a method of bandwidth selection based on the rule of the thumb of Fan and Gijbels (1996). The method is based on a global 4th order polynomial approximation on either side of the cutoff. Unfortunately, the regression is sparse for our case. Hence we choose $h = c * b$ where $c \in \{10, 15, 20\}$. We use relatively small bandwidths, since the normal approximation behaves poorly when the bandwith is large due to the bias term.[27]

[27] We may want to estimate bias to double-check that our results are not affected by this issue.

4.9.3 Local Linear Estimator for Density Plots

The density plots are local linear estimators of the density on the right and left of the cutoff, as follows

$$\begin{aligned}(\hat{\phi}_1, \hat{\phi}_2) &\equiv \arg\min L(\phi_1, \phi_2, r) \\ &= \sum_{j=1}^{J} \{Y_j - \phi_1 - \phi_2(X_j - r)\}^2 K\left(\frac{X_j - r}{h}\right) \\ &\quad (1\{X_j > c\}1\{r \geq c\} + 1\{X_j < c\}1\{r < c\})\end{aligned}$$

4.10 Correction for Multiple Testing

To correct for multiple testing, for each city pair we use a nominal value of $\alpha = 0.05$ and then apply Romano and Shaikh (2006) (Romano and Shaikh, 2006) step-up procedure to control the k-FWER with $k = 1$ and the initial α_i as defined in (13) in Romano and Shaikh (2006).

$$\alpha_i = \begin{cases} \frac{k}{s}, & \text{if} i \leq k, \\ \frac{k}{s+k-i}, & \text{if} i > k, \end{cases}, \tag{4.17}$$

where s denotes the total number of hypotheses, $H_1, H_2, ..., H_s$, and for our case this is equivalent to the total number of weather variable combinations. Together with $D_1(k, s)$ defined as

$$\begin{aligned}D_1(k, s) &= \max_{k \leq |I| \leq s} S_1(k, s, |I|), \\ S_1(k, s, |I|) &= |I| \frac{\alpha_{s-|I|+k}}{k} + |I| \sum_{k<j<|I|} \frac{\alpha_{s-|I|+j} - \alpha_{s-|I|+j-1}}{j},\end{aligned} \tag{4.18}$$

we can construct the critical values for our p-values,which we denote by p_i^c,[28]

$$p_i^c = \frac{\alpha\alpha_i'}{D_1(k,s)} \tag{4.19}$$

To apply the step-up procedure, we order the observed p-values $\hat{p}_i$ in an ascending order, and let $\hat{p}_{(i)}$ denote the i^{th} smallest p-value. Now the step-up procedure start with the largest p-value, $\hat{p}_{(s)}$. If $\hat{p}_{(s)} \leq p_s^c$, then reject all hypotheses $H_1, ..., H_s$. Otherwise, reject $H_1, ..., H_j$, such that j is the smallest integer for which,

$$\hat{p}_{(s)} > p_s^c, ..., \hat{p}_{(j+1)}^c > p_{j+1}^c. \tag{4.20}$$

4.11 Discretization of Weather Variables

Visibility is rounded to the nearest integer. Precipitation is divided into three equi-space bins $[0, 0.33]$, $(0.33, 0.67]$ and $(0.67, 1]$, which we refer to in the table by 0, 0.5 and 1, respectively. Windspeed is divided into 5 bins, $[0, 5]$, $(5, 10]$, $(10, 15]$, $(15, 20]$, $(20, 25]$. Finally, temperature is also divided into 5 bins ≤ 20, $(20, 40]$, $(40, 60]$, $(60, 80]$, and $(80, 100]$.

[28] Romano and Shaikh (2006) denote it by $\tilde{\alpha}_i$.

Appendix A

Tables and Figures

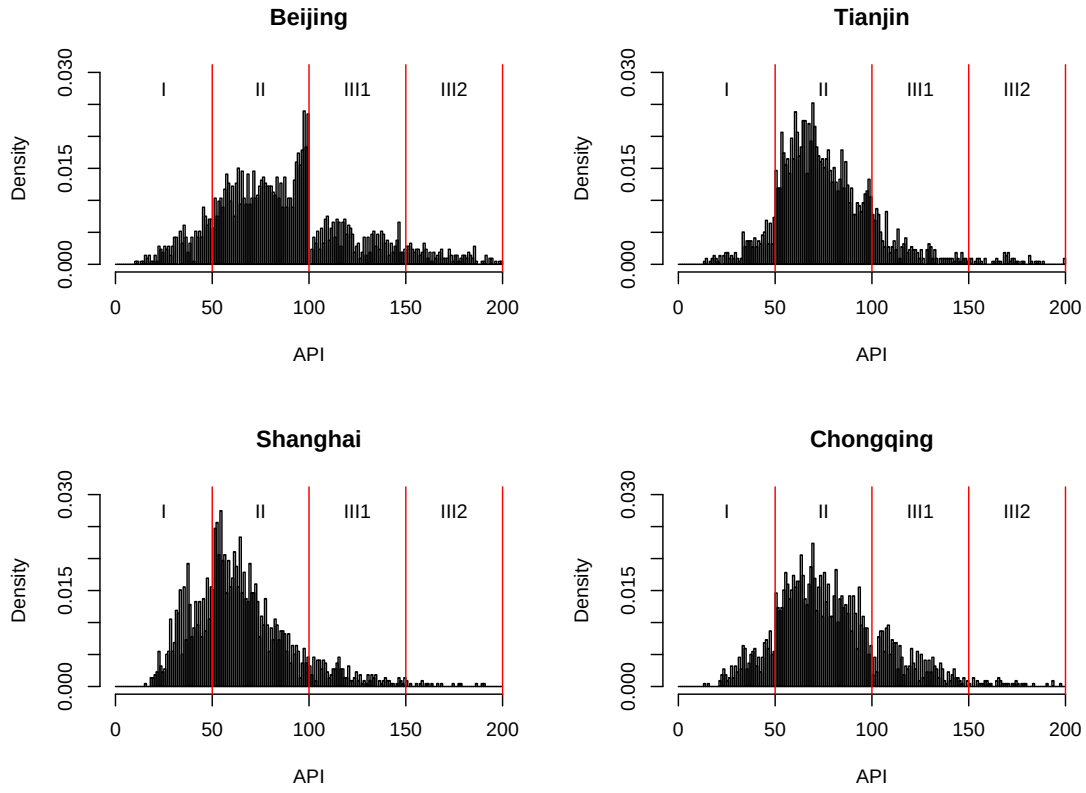

Figure A.1: Frequency of API scores in four capital cities: Beijing (upper left), Tianjin (upper right), Shanghai (lower left), and Chongqing (lower right). The histograms are based on the city-level daily API scores during 2001-2010. Scores greater than 200 are not illustrated. The red lines denote the cut-off points for each pollution level. Levels I and II are defined as "blue-sky days."

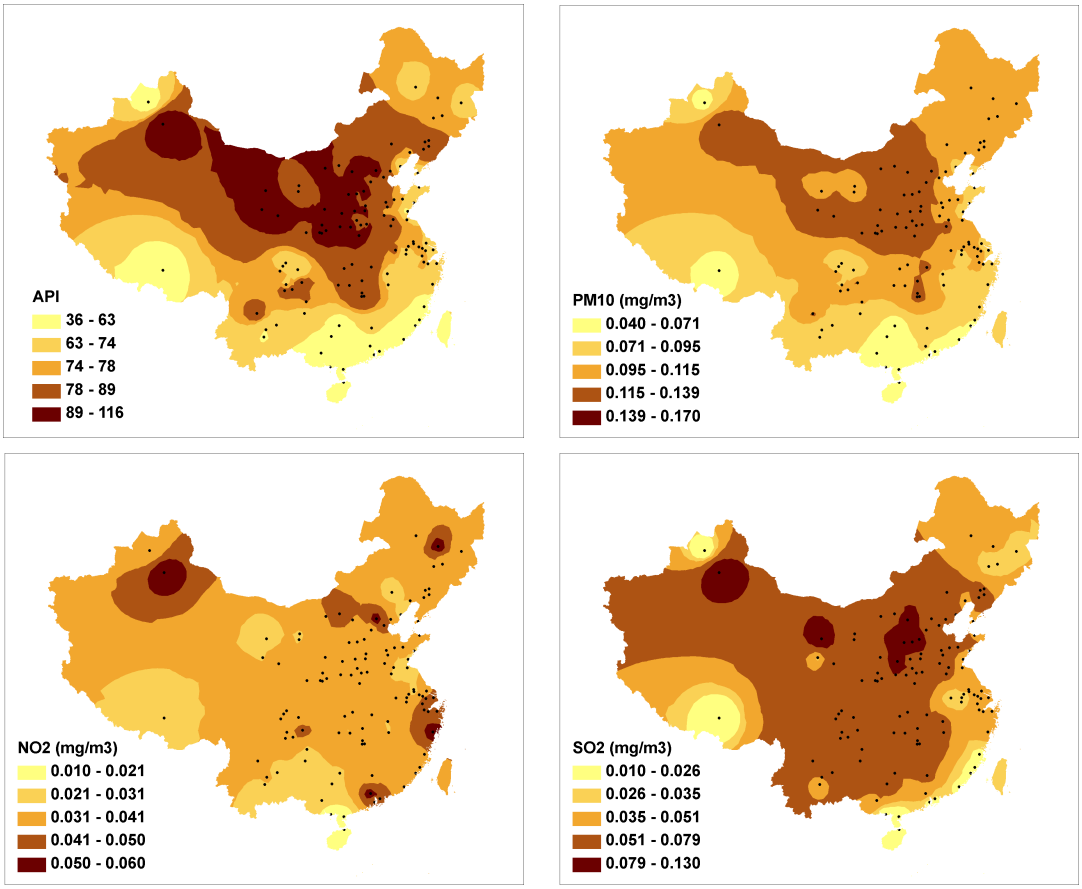

Figure A.2: Average daily air pollution levels during 2001-2010: air pollution index (API, upper left), particulate matter concentration (PM_{10}, upper right), nitrogen dioxide concentration (NO_2, lower left), and sulfur dioxide concentration (SO_2, lower right). The filled contour plot shows spatially interpolated air pollution levels, which is generated by inverse distance weighting approach based on the city-level daily air pollution data. The dots represent cities that disclose daily API scores.

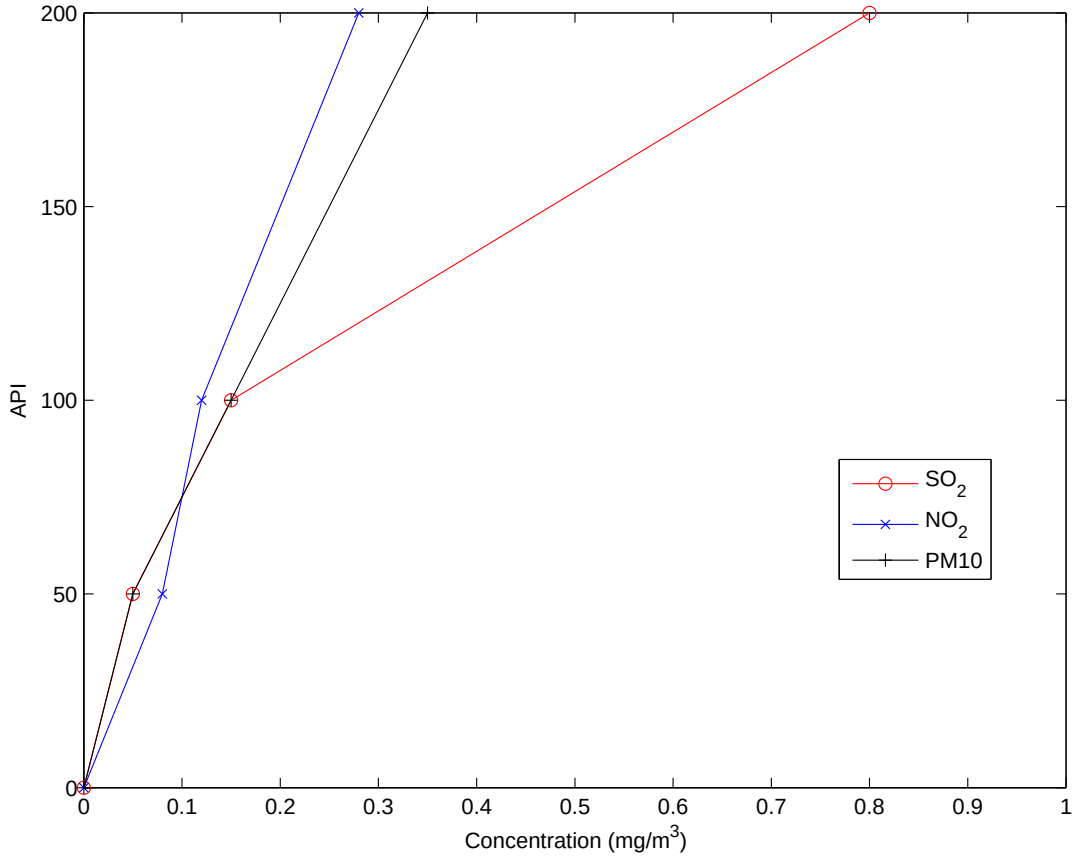

Figure A.3: API is a nonlinear transformation of pollutant concentrations. In particular, there are kinks corresponding to API=100 for SO_2 and NO_2 respectively

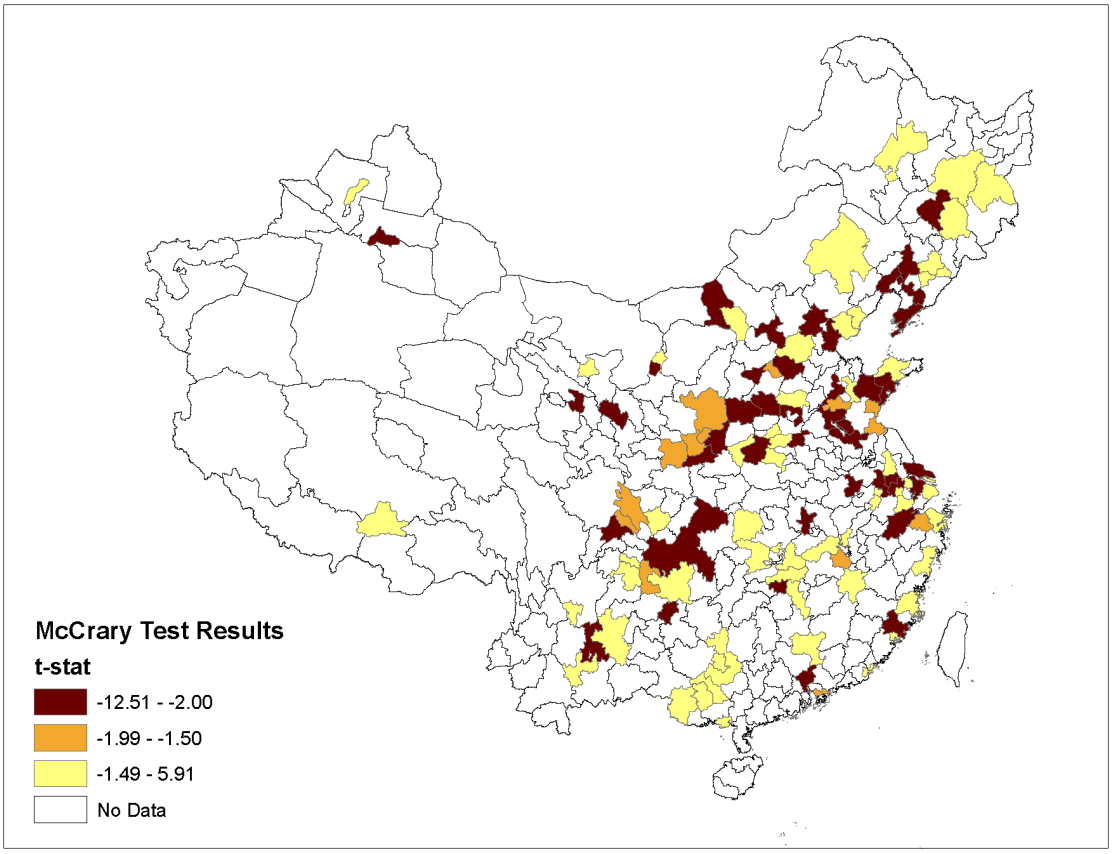

Figure A.4: Results of McCrary (2008) tests show different degrees of manipulation in the PM_{10} data. Three levels of manipulation are illustrated: t-stats below -1.5 as non-manipulators, t-stats between -1.5 and -2 as borderline manipulators, and finally, t-stats smaller than -2 as manipulators. PM_{10} is selected because it is the dominant primary pollutant. A polygon is a prefecture that has the city as its capital.

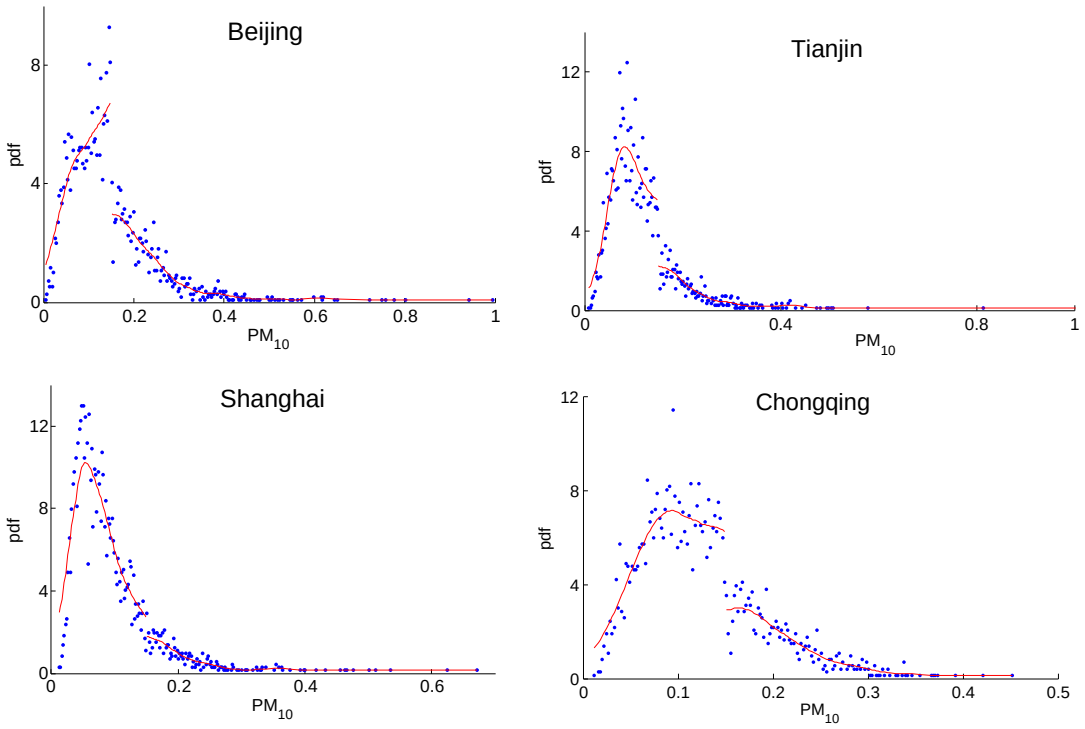

Figure A.5: The McCrary (2008) test on four major cities. Beijing (upper left), Tianjin (upper right), Shanghai (lower left), and Chongqing (lower right). The blue dots are the first-step histogram obtained from implementing McCrary (2008) on the city-level daily PM_{10} concentrations during 2001-2010. The red curve is the Kernel estimator of the pdf to the right and left of the cut-off.

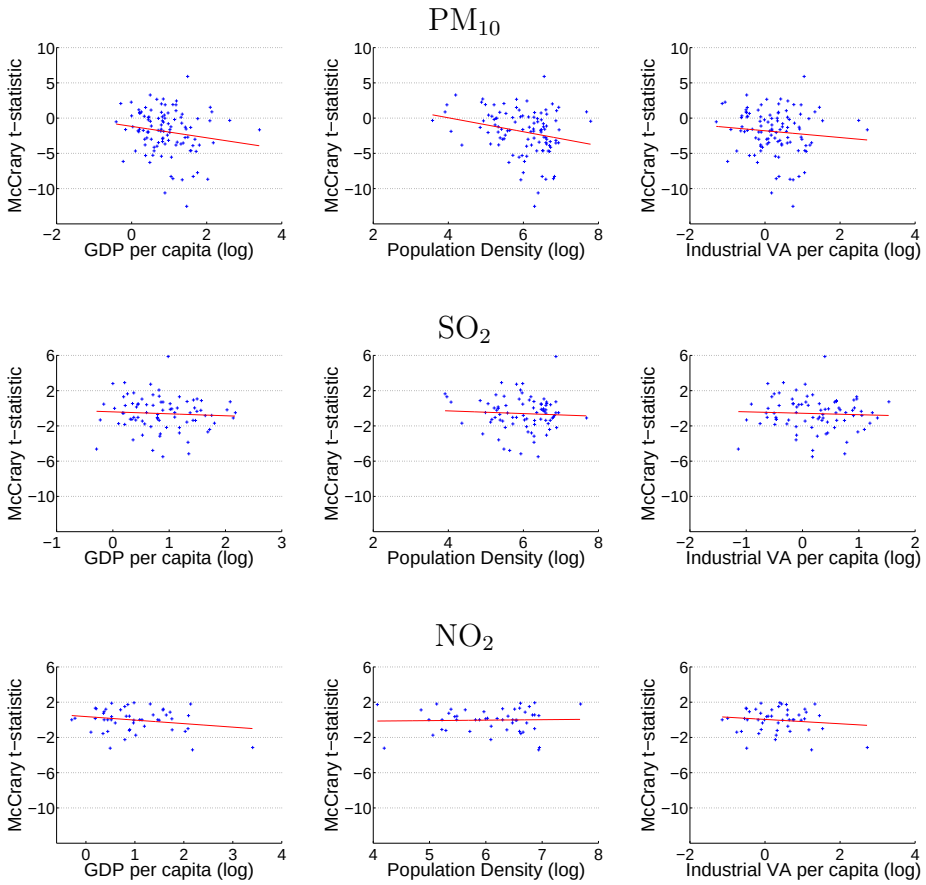

Figure A.6: GDP per capita, population density, industrial value added per capita and McCrary results. The scatter plots graph the relationship between the log of GDP per capita (column 1), log of population density (column 2), log of value added of the secondary industry per capita (column 3) and the McCrary t-statistics for PM_{10} (upper), SO_2 (middle), and NO_2 (lower).

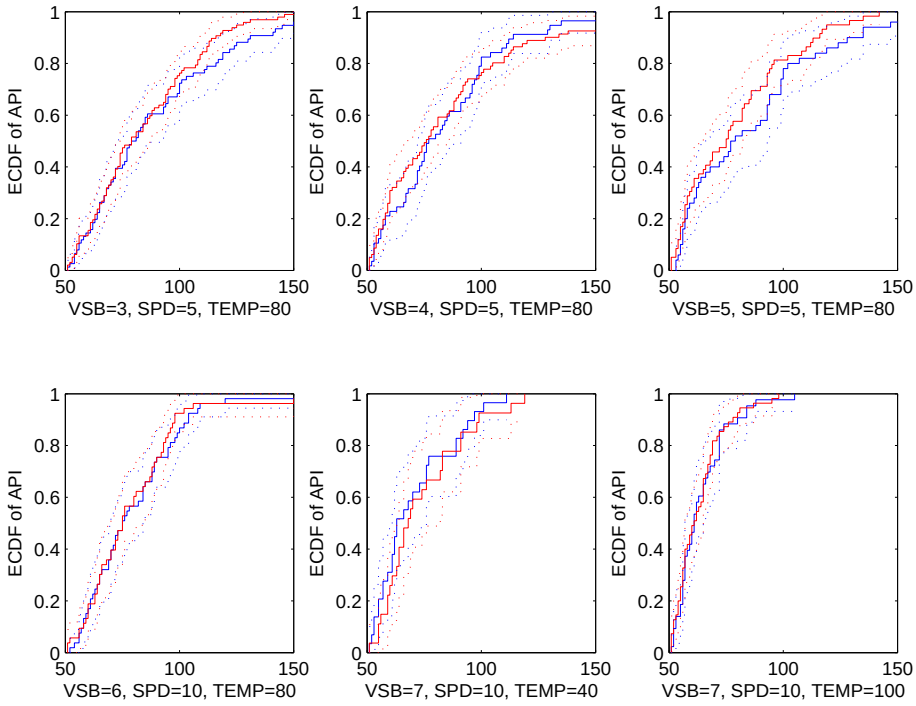

Figure A.7: Panel Matching Approach: Zhenjiang-Yangzhou. The red lines represent Zhenjiang, the blue Yangzhou. The solid lines are the empirical cdfs of API conditional on the weather variables specified in the plot, the dotted lines give 95% point-wise confidence intervals for the empirical cdf. VSB denotes visibility, WSP wind speed, and TEMP temperature. The values for the weather variables are the discretized versions of the actual data for details on the discretization see Appendix 4.11.

Table A.1: Shrinkage Estimator: Monte Carlo Results for the Common Parameter $\theta_0 = 1$, $\alpha_i \sim N(0, b)$, $S = 1000$, $N = 100$

				T=4					
$\alpha_i \sim$	Estimator	Mean	Median	SD	RMSE	MAE	$\hat{p}; 0.10$	$\hat{p}; 0.05$	S.E./SD
N(0,1)	MLE	1.42	1.41	0.382	0.570	0.412	0.379	0.263	0.851
	Analytical	1.12	1.11	0.293	0.316	0.205	0.095	0.0354	1.069
	Shrinkage	1.08	1.08	0.310	0.321	0.214	0.106	0.045	1.008
N(0,2)	MLE	1.44	1.43	0.497	0.663	0.456	0.304	0.206	0.824
	Analytical	1.12	1.12	0.377	0.397	0.262	0.085	0.033	1.040
	Shrinkage	1.07	1.08	0.407	0.413	0.283	0.104	0.051	0.960
N(0,3)	MLE	1.47	1.45	0.630	0.788	0.495	0.255	0.166	0.789
	Analytical	1.14	1.13	0.444	0.465	0.310	0.066	0.027	1.063
	Shrinkage	1.06	1.06	0.487	0.491	0.344	0.106	0.039	0.963
				T=8					
$\alpha_i \sim$	Estimator	Mean	Median	SD	RMSE	MAE	$\hat{p}; 0.10$	$\hat{p}; 0.05$	S.E./SD
N(0,1)	MLE	1.19	1.18	0.148	0.238	0.177	0.367	0.254	0.917
	Analytical	1.06	1.05	0.131	0.143	0.088	0.122	0.066	1.003
	Shrinkage	1.05	1.04	0.133	0.142	0.089	0.125	0.067	0.989
N(0,2)	MLE	1.20	1.19	0.178	0.267	0.196	0.320	0.223	0.925
	Analytical	1.07	1.06	0.157	0.171	0.111	0.112	0.057	1.011
	Shrinkage	1.06	1.05	0.160	0.171	0.110	0.111	0.053	0.990
N(0,3)	MLE	1.22	1.21	0.233	0.317	0.219	0.286	0.189	0.900
	Analytical	1.08	1.07	0.204	0.219	0.139	0.131	0.073	0.925
	Shrinkage	1.07	1.06	0.209	0.220	0.137	0.142	0.076	0.9004

Table A.2: Shrinkage Estimator: Monte Carlo Results for the Common Parameter $\theta_0 = 1$, $\alpha_i \sim N(a, 1)$, $S = 1000$, $N = 100$

				T=4					
$\alpha_i \sim$	Estimator	Mean	Median	SD	RMSE	MAE	$\hat{p}; 0.10$	$\hat{p}; 0.05$	S.E./SD
N(0,1)	MLE	1.42	1.41	0.382	0.570	0.412	0.379	0.263	0.851
	Analytical	1.12	1.11	0.293	0.316	0.205	0.095	0.0354	1.069
	Shrinkage	1.08	1.08	0.310	0.321	0.214	0.106	0.045	1.008
N(1,1)	MLE	1.47	1.44	0.501	0.690	0.472	0.315	0.216	0.820
	Analytical	1.15	1.13	0.372	0.400	0.254	0.089	0.043	1.056
	Shrinkage	1.10	1.09	0.400	0.413	0.260	0.108	0.051	1.976
N(2,1)	MLE	1.66	1.58	0.933	1.145	0.665	0.223	0.132	0.773
	Analytical	1.21	1.20	0.595	0.632	0.428	0.059	0.023	1.108
	Shrinkage	1.11	1.09	0.649	0.658	0.454	0.069	0.022	1.004
				T=8					
$\alpha_i \sim$	Estimator	Mean	Median	SD	RMSE	MAE	$\hat{p}; 0.10$	$\hat{p}; 0.05$	S.E./SD
N(0,1)	MLE	1.19	1.18	0.148	0.238	0.177	0.367	0.254	0.917
	Analytical	1.06	1.05	0.131	0.143	0.088	0.122	0.066	1.003
	Shrinkage	1.05	1.04	0.133	0.142	0.089	0.125	0.067	0.989
N(1,1)	MLE	1.24	1.22	0.215	0.319	0.235	0.353	0.238	0.868
	Analytical	1.10	1.08	0.189	0.211	0.134	0.140	0.085	0.951
	Shrinkage	1.09	1.08	0.193	0.211	0.135	0.139	0.083	0.929
N(2,1)	MLE	1.37	1.32	0.437	0.571	0.330	0.238	0.133	0.802
	Analytical	1.18	1.14	0.343	0.390	0.221	0.107	0.053	0.9537
	Shrinkage	1.16	1.12	0.363	0.398	0.232	0.116	0.063	0.895

Table A.3: Shrinkage Estimator: Monte Carlo Results for the Overall Average Partial Effect, $\mu/\mu_0 = 1$, $\alpha_i \sim N(0, b)$, $S = 1000$, $N = 100$

				T=4					
$\alpha_i \sim$	Estimator	Mean	Median	SD	RMSE	MAE	$\hat{p}; 0.10$	$\hat{p}; 0.05$	S.E./SD
N(0,1)	MLE	0.989	0.991	0.232	0.232	0.157	0.191	0.124	0.834
	Analytical	0.996	0.999	0.249	0.249	0.171	0.202	0.131	0.779
	Shrinkage	0.950	0.957	0.267	0.271	0.191	0.238	0.163	0.728
N(0,2)	MLE	0.970	0.966	0.299	0.300	0.196	0.221	0.141	0.782
	Analytical	0.978	0.972	0.319	0.319	0.207	0.221	0.147	0.736
	Shrinkage	0.910	0.912	0.350	0.361	0.235	0.277	0.186	0.674
N(0,3)	MLE	1.002	0.991	0.358	0.358	0.234	0.215	0.137	0.737
	Analytical	1.010	0.996	0.375	0.375	0.248	0.213	0.139	0.709
	Shrinkage	0.923	0.925	0.412	0.419	0.283	0.263	0.185	0.651
				T=8					
$\alpha_i \sim$	Estimator	Mean	Median	SD	RMSE	MAE	$\hat{p}; 0.10$	$\hat{p}; 0.05$	S.E./SD
N(0,1)	MLE	0.992	0.990	0.103	0.103	0.070	0.200	0.125	0.737
	Analytical	1.010	1.007	0.109	0.109	0.073	0.208	0.139	0.702
	Shrinkage	1.009	1.006	0.110	0.110	0.075	0.214	0.140	0.696
N(0,2)	MLE	0.993	0.990	0.122	0.123	0.085	0.203	0.139	0.655
	Analytical	1.011	1.008	0.130	0.130	0.086	0.206	0.139	0.626
	Shrinkage	1.010	1.008	0.132	0.132	0.087	0.210	0.150	0.619
N(0,3)	MLE	0.993	0.990	0.122	0.123	0.085	0.203	0.139	0.655
	Analytical	1.011	1.008	0.130	0.130	0.086	0.206	0.139	0.626
	Shrinkage	1.010	1.008	0.132	0.132	0.087	0.210	0.150	0.619

Table A.4: Shrinkage Estimator: Monte Carlo Results for Overall Average Partial Effect, $\mu/\mu_0 = 1$, $\alpha_i \sim N(a, 1)$, $S = 1000$, $N = 100$

				T=4					
$\alpha_i \sim$	Estimator	Mean	Median	SD	RMSE	MAE	$\hat{p}; 0.10$	$\hat{p}; 0.05$	S.E./SD
N(0,1)	MLE	0.989	0.991	0.232	0.232	0.157	0.191	0.124	0.834
	Analytical	0.996	0.999	0.249	0.249	0.171	0.202	0.131	0.779
	Shrinkage	0.950	0.957	0.267	0.271	0.191	0.238	0.163	0.728
N(1,1)	MLE	0.969	0.967	0.291	0.292	0.197	0.209	0.128	0.778
	Analytical	0.976	0.966	0.308	0.309	0.210	0.210	0.136	0.736
	Shrinkage	0.911	0.908	0.337	0.348	0.234	0.277	0.196	0.678
N(2,1)	MLE	0.957	0.943	0.457	0.459	0.301	0.240	0.165	0.724
	Analytical	0.930	0.922	0.464	0.470	0.312	0.244	0.153	0.704
	Shrinkage	0.811	0.807	0.499	0.533	0.373	0.339	0.237	0.661
				T=8					
$\alpha_i \sim$	Estimator	Mean	Median	SD	RMSE	MAE	$\hat{p}; 0.10$	$\hat{p}; 0.05$	S.E./SD
N(0,1)	MLE	0.992	0.990	0.103	0.103	0.070	0.200	0.125	0.737
	Analytical	1.010	1.007	0.109	0.109	0.073	0.208	0.139	0.702
	Shrinkage	1.009	1.006	0.110	0.110	0.075	0.214	0.140	0.696
N(1,1)	MLE	0.986	0.983	0.144	0.145	0.096	0.249	0.164	0.633
	Analytical	1.002	0.999	0.151	0.151	0.096	0.260	0.172	0.608
	Shrinkage	1.001	0.999	0.153	0.153	0.098	0.265	0.175	0.602
N(2,1)	MLE	0.960	0.959	0.248	0.251	0.164	0.325	0.229	0.557
	Analytical	0.967	0.962	0.253	0.255	0.168	0.322	0.236	0.543
	Shrinkage	0.965	0.966	0.260	0.262	0.168	0.327	0.243	0.532

Table A.5: Shrinkage Estimator: Monte Carlo Results for Average Partial Effect at the Average, $\mu/\mu_0 = 1$, $\alpha_i \sim N(0, b)$, $S = 1000$, $N = 100$

				T=4					
$\alpha_i \sim$	Estimator	Mean	Median	SD	RMSE	MAE	$\hat{p}; 0.10$	$\hat{p}; 0.05$	S.E./SD
N(0,1)	MLE	0.992	0.993	0.242	0.242	0.165	0.199	0.125	0.788
	Analytical	0.983	0.983	0.250	0.250	0.174	0.199	0.123	0.794
	Shrinkage	0.807	0.804	0.245	0.312	0.228	0.316	0.211	0.810
N(0,2)	MLE	0.981	0.977	0.310	0.311	= 0.200	0.216	0.158	0.737
	Analytical	0.972	0.965	0.318	0.319	0.207	0.209	0.141	0.755
	Shrinkage	0.759	0.746	0.297	0.382	0.282	0.322	0.221	0.812
N(0,3)	MLE	1.018	1.011	0.369	0.369	0.239	0.225	0.156	0.691
	Analytical	1.004	0.994	0.369	0.369	0.248	0.196	0.134	0.732
	Shrinkage	0.762	0.749	0.329	0.407	0.296	0.272	0.170	0.823
				T=8					
$\alpha_i \sim$	Estimator	Mean	Median	SD	RMSE	MAE	$\hat{p}; 0.10$	$\hat{p}; 0.05$	S.E./SD
N(0,1)	MLE	1.019	1.018	0.119	0.121	0.080	0.201	0.141	0.711
	Analytical	1.022	1.021	0.121	0.123	0.081	0.201	0.140	0.712
	Shrinkage	0.963	0.963	0.121	0.126	0.087	0.221	0.145	0.709
N(0,2)	MLE	1.035	1.031	0.140	0.144	0.092	0.235	0.151	0.634
	Analytical	1.035	1.032	0.141	0.145	0.091	0.225	0.141	0.644
	Shrinkage	0.955	0.951	0.141	0.148	0.101	0.235	0.146	0.648
N(0,3)	MLE	1.035	1.031	0.140	0.144	0.092	0.235	0.151	0.634
	Analytical	1.035	1.032	0.141	0.145	0.091	0.225	0.141	0.644
	Shrinkage	0.955	0.951	0.141	0.148	0.101	0.235	0.146	0.648

Table A.6: Shrinkage Estimator: Monte Carlo Results for Average Partial Effect at the Average, $\mu/\mu_0 = 1$, $\alpha_i \sim N(a, 1)$, $S = 1000$, $N = 100$

				T=4					
$\alpha_i \sim$	Estimator	Mean	Median	SD	RMSE	MAE	$\hat{p}; 0.10$	$\hat{p}; 0.05$	S.E./SD
N(0,1)	MLE	0.992	0.993	0.242	0.242	0.165	0.199	0.125	0.788
	Analytical	0.983	0.983	0.250	0.250	0.174	0.199	0.123	0.794
	Shrinkage	0.807	0.804	0.245	0.312	0.228	0.316	0.211	0.810
N(1,1)	MLE	0.983	0.973	0.300	0.301	0.205	0.224	0.136	0.732
	Analytical	0.972	0.961	0.306	0.307	0.214	0.203	0.133	0.758
	Shrinkage	0.762	0.746	0.286	0.372	0.283	0.332	0.223	0.811
N(2,1)	MLE	0.978	0.969	0.451	0.451	0.298	0.276	0.199	0.658
	Analytical	0.935	0.939	0.454	0.458	0.297	0.228	0.137	0.717
	Shrinkage	0.685	0.671	0.378	0.492	0.375	0.276	0.166	0.865
				T=8					
$\alpha_i \sim$	Estimator	Mean	Median	SD	RMSE	MAE	$\hat{p}; 0.10$	$\hat{p}; 0.05$	S.E./SD
N(0,1)	MLE	1.019	1.018	0.119	0.121	0.080	0.201	0.141	0.711
	Analytical	1.022	1.021	0.121	0.123	0.081	0.201	0.140	0.712
	Shrinkage	0.963	0.963	0.121	0.126	0.087	0.221	0.145	0.709
N(1,1)	MLE	1.043	1.038	0.157	0.162	0.105	0.286	0.219	0.579
	Analytical	1.037	1.037	0.154	0.159	0.102	0.257	0.183	0.608
	Shrinkage	0.943	0.944	0.153	0.163	0.112	0.271	0.189	0.620
N(2,1)	MLE	1.058	1.061	0.247	0.254	0.177	0.439	0.365	0.416
	Analytical	1.023	1.024	0.232	0.234	0.162	0.345	0.283	0.472
	Shrinkage	0.937	0.938	0.246	0.253	0.161	0.401	0.311	0.486

Table A.7: Testing Identifying Assumptions Models (A)-(D): Monte Carlo Results for KS and CM Tests, $S = 1000$

	n	1000			1500			2000		
	α	0.025	0.05	0.10	0.025	0.05	0.10	0.025	0.05	0.10
		Model (A)								
KS	nt	0.024	0.052	0.110	0.026	0.054	0.104	0.036	0.068	0.103
	pt	0.013	0.022	0.055	0.020	0.041	0.068	0.020	0.031	0.074
	gt	0.009	0.021	0.048	0.016	0.022	0.047	0.012	0.023	0.045
	$excl$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
CM	nt	0.025	0.053	0.113	0.026	0.046	0.086	0.029	0.045	0.116
	pt	0.022	0.042	0.091	0.026	0.052	0.084	0.028	0.052	0.094
	gt	0.024	0.049	0.094	0.024	0.039	0.082	0.031	0.048	0.091
	$excl$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		Model (B)								
KS	nt	0.985	0.996	0.998	0.998	1.000	1.000	1.000	1.000	1.000
	pt	0.014	0.023	0.054	0.020	0.041	0.067	0.020	0.030	0.074
	gt	0.009	0.021	0.046	0.015	0.022	0.046	0.011	0.021	0.044
	$excl$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
CM	nt	0.861	0.919	0.951	0.971	0.991	0.997	0.993	0.998	1.000
	pt	0.023	0.043	0.090	0.030	0.053	0.081	0.028	0.050	0.092
	gt	0.025	0.052	0.095	0.030	0.043	0.083	0.030	0.047	0.092
	$excl$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		Model (C)								
KS	nt	0.979	0.987	0.997	1.000	1.000	1.000	1.000	1.000	1.000
	pt	0.941	0.966	0.982	0.992	0.996	0.998	1.000	1.000	1.000
	gt	0.009	0.021	0.048	0.015	0.020	0.045	0.012	0.023	0.045
	$excl$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
CM	nt	0.428	0.588	0.775	0.669	0.805	0.921	0.871	0.937	0.977
	pt	0.474	0.626	0.762	0.734	0.848	0.930	0.893	0.952	0.979
	gt	0.025	0.048	0.096	0.030	0.039	0.081	0.031	0.048	0.092
	$excl$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		Model (D)								
KS	nt	0.990	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	pt	0.672	0.771	0.861	0.894	0.941	0.975	0.974	0.986	0.995
	gt	0.416	0.521	0.658	0.731	0.821	0.920	0.901	0.954	0.985
	$excl$	0.038	0.073	0.124	0.042	0.064	0.103	0.036	0.065	0.125
CM	nt	0.965	0.979	0.994	1.000	1.000	1.000	1.000	1.000	1.000
	pt	0.047	0.073	0.121	0.048	0.092	0.177	0.073	0.125	0.212
	gt	0.051	0.111	0.203	0.110	0.185	0.309	0.157	0.259	0.423
	$excl$	0.035	0.064	0.116	0.030	0.051	0.103	0.025	0.048	0.099
	s.e.	0.005	0.007	0.009	0.005	0.007	0.009	0.005	0.007	0.009

Table A.8: National Longitudinal Survey of Youth 1983-87 (revised since Angrist and Newey (1991)): Descriptive Statistics

	1983	1984	1985	1986	1987
Race	0.12				
Age	21.84				
	(2.22)				
HGC	12.34	12.45	12.57	12.57	12.61
	(1.77)	(1.83)	(1.94)	(1.94)	(1.98)
South	0.29	0.30	0.30	0.30	0.30
Urban	0.76	0.77	0.76	0.77	0.76
Log Hourly Wage	6.31	6.39	6.50	6.61	6.72
	(0.48)	(0.49)	(0.49)	(0.49)	(0.50)

Table A.9: P-Values for Testing the Time Homogeneity up to a Time Effect: Revisiting Angrist and Newey (1991)

	KS	CM	F
LOG Hourly Wage			
1983-84	0.50	0.74	0.11
1984-85	0.11	0.87	0.09
1985-86	0.40	0.72	0.25
1986-87	0.43	0.17	0.46

Table A.10: Estimates of the Average Partial EFfect for Returns to Schooling: Revisiting Angrist and Newey (1991)

Grade		Subs	APE	S.E.	t-Stat
1983	1984				
11	12	26	0.052	0.089	0.576
12	13	24	-0.120	0.133	-0.904
13	14	28	-0.069	0.061	-1.124
14	15	19	-0.037	0.035	-1.045
15	16	20	0.141	0.131	1.073
All Movers		122	-0.012	0.043	-0.267
1984	1985				
11	12	6	-0.010	0.032	-0.318
12	13	14	-0.080	0.067	-1.191
13	14	12	0.295	0.247	1.194
14	15	17	0.011	0.063	0.178
15	16	17	0.263	0.111	2.379
All Movers		73	0.095	0.055	1.723
1985	1986				
11	12	3	0.457	0.278	1.644
12	13	12	0.148	0.113	1.311
13	14	11	0.206	0.170	1.214
14	15	14	0.052	0.106	0.496
15	16	11	0.601	0.112	5.345
All Movers		58	0.226	0.060	3.737
1986	1987				
13	14	9	0.030	0.119	0.250
14	15	8	-0.125	0.168	-0.743
15	16	11	0.167	0.099	1.695
All Movers		41	-0.012	0.068	-0.179

Note: Results are listed for subpopulations with subsample size of 5 and above.

Table A.11: Returns to Schooling: Replication of Angrist and Newey (1991) ANACOVA Results

	Full Sample			1983-1984		
	RF	X	A	RF	X	A
Grade	-0.0688	0.0772	0.0714	-0.2343	-0.0196	-0.0097
	(0.1308)	(0.0152)	(0.0148)	(0.1502)	(0.0194)	(0.0184)
$Grade^2$	-0.0070	-0.0005		-0.0013	0.0006	
	(0.0048)	(0.0004)		(0.0060)	(0.0005)	
Age^2	-0.0030		-0.0001	-0.0023		0.0006
	(0.0005)		(0.0003)	(0.0009)		(0.0004)
$Grade * Age$	0.0134			0.0113		
	(0.0015)			(0.0032)		
Union	0.1397	0.1423	0.1424	0.1642	0.1642	0.1643
	(0.0162)	(0.0163)	(0.0163)	(0.0152)	(0.0152)	(0.0152)

Table A.12: Air Pollution Index and Pollutant Concentrations: Summary Statistics at the Pollutant Levels

		Mean	Std	Min	Max
API (max)		76.315	37.680	0	625
Primary Pollutant					
	PM10	0.737			
	SO2	0.092			
	NO2	0.002			
Grade					
	I	0.169			
	II	0.677			
	III1	0.124			
	III2	0.021			
	IV1	0.003			
	IV2	0.002			
	V	0.004			
Pollution Concentration					
	PM10	0.100	0.066	0.001	2.721
	SO2	0.054	0.055	0.001	2.147
	NO2	0.036	0.020	0.001	0.353
API (pollutant level)					
	PM10	74.731	38.277	0	501
	SO2	43.262	27.544	0	409
	NO2	22.850	14.116	0	226

Table A.13: t-Statistics for McCrary (2008) Test: Cities Exhibiting Manipulation (PM10)

c	10	15	20	c	10	15	20
Shenyang	-10.318	-12.514	-14.091	Quanzhou	-3.012	-3.414	-4.043
Shijiazhuang	-8.620	-10.609	-11.865	Zaozhuang	-1.700	-3.330	-3.975
Anshan	-7.612	-8.748	-9.528	Nantong	-2.285	-3.160	-3.145
Beijing	-7.800	-8.673	-9.064	Dalian	-2.372	-2.998	-3.742
Chengdou	-6.407	-8.588	-10.079	Changzhi	-2.548	-2.992	-3.156
Nanjing	-7.291	-8.277	-8.862	Anyang	-3.075	-2.992	-3.161
Hefei	-6.648	-8.275	-9.239	Lanzhou	-2.069	-2.931	-4.144
Hangzhou	-6.235	-7.735	-8.441	Weifang	-0.821	-2.833	-4.159
Xining	-5.286	-6.294	-6.984	Datong	-2.935	-2.803	-2.693
Weinan	-4.828	-6.144	-6.899	Jinan	-1.356	-2.757	-3.841
Changchun	-3.602	-5.541	-6.750	Zhenjiang	-1.512	-2.697	-3.553
Luoyang	-5.198	-5.409	-5.307	Kunming	-2.226	-2.413	-2.327
Chongqing	-4.167	-5.307	-5.761	Yinchuan	-1.806	-2.045	-2.541
Jinzhou	-4.054	-5.275	-5.728	Baoji	-1.367	-1.979	-2.590
Tianjin	-3.759	-4.725	-5.510	Deyang	-1.063	-1.975	-2.406
Kaifeng	-3.604	-4.663	-5.622	Xianyang	-0.467	-1.930	-2.966
Xiangtan	-3.514	-4.649	-5.437	Yanan	-1.282	-1.882	-2.202
Qingdao	-3.300	-4.625	-4.985	Rizhao	-1.545	-1.841	-2.214
Suzhou	-3.066	-4.609	-5.728	Mianyang	-1.671	-1.803	-1.935
Xuzhou	-3.301	-4.133	-4.539	Lianyungang	-0.757	-1.738	-2.194
Guiyang	-3.072	-3.982	-4.812	Yangquan	-1.731	-1.736	-1.524
Changzhou	-3.574	-3.905	-4.079	Shenzhen	-0.572	-1.656	-2.540
Baotou	-2.856	-3.829	-4.532	Nanchang	-0.858	-1.633	-2.358
Taiyuan	-2.833	-3.820	-5.254	Tongchuan	-0.581	-1.576	-1.859
Linfen	-3.466	-3.803	-4.056	Luzhou	-1.671	-1.620	-1.278
Xian	-1.848	-3.672	-4.914	Taian	-0.681	-1.572	-2.472
Wulumuqi	-3.582	-3.672	-4.269	Yantai	-0.873	-1.288	-1.879
Wuhan	-3.242	-3.534	-3.355	Yangzhou	-0.673	-1.237	-1.663
Guangzhou	-3.113	-3.497	-3.324	Lasa	-0.711	-1.099	-1.678
Jining	-2.109	-3.475	-3.973	Shaoxing	-1.030	-1.544	-1.463
				Huhehaote	-1.066	-1.325	-1.614

Note: $c = h/\hat{b}$, where $\hat{b} = 2\hat{\sigma}/\sqrt{n}$, where $\hat{\sigma}$ of the pollutant concentration.
For more details, see Section 4.9.

Table A.14: t-Statistics for McCrary (2008) Test: Cities Not Exhibiting Manipulation (PM10)

c	10	15	20	c	10	15	20
Qujing	-1.271	-1.225	-1.583	Wenzhou	0.731	0.731	0.729
Handan	-0.744	-1.122	-1.486	Baoding	1.382	0.785	0.150
Yichang	-0.121	-0.974	-1.472	Sanmenxia	1.015	0.824	0.917
Wuhu	-0.178	-0.797	-1.474	Chifeng	1.175	0.877	0.319
Tangshan	-0.435	-0.714	-0.629	Shanghai	1.285	0.887	0.603
Fuzhou	-0.921	-0.562	-0.019	Zhengzhou	1.260	1.055	1.169
Nanchong	NaN	-0.524	-0.808	Yueyang	1.195	1.140	0.833
Shantou	-0.883	-0.447	-0.325	Zigong	1.719	1.177	0.665
Yuxi	-0.257	-0.392	-0.100	Changde	1.792	1.505	1.271
Wuxi	0.305	-0.372	-0.535	Xiamen	1.602	1.530	1.444
Kelamayi	-0.526	-0.337	-0.236	Jinchang	2.341	1.882	1.886
Pingdingshan	0.161	-0.239	-0.608	Haerbin	1.563	1.888	1.679
Benxi	-0.343	-0.194	-0.139	Panzhihua	2.114	1.918	2.004
Ningbo	0.306	-0.078	-0.340	Jiaozuo	2.289	2.019	1.859
Liuzhou	-0.267	-0.039	-0.080	Zunyi	2.400	2.065	2.065
Fushun	-0.342	-0.013	-0.147	Qiqihaer	2.220	2.262	2.470
Nanning	0.631	0.077	-0.084	Huzhou	2.411	2.434	2.649
Changsha	0.671	0.126	-0.362	Zhuzhou	2.295	2.666	2.804
Beihai	0.431	0.164	0.174	Shizuishan	3.074	2.712	2.130
Qinhuangdao	0.215	0.184	0.422	Mudanjiang	2.747	3.276	3.748
Shaoguan	0.382	0.266	0.237	Zibo	5.304	5.908	6.024
Yibin	0.591	0.319	-0.019	Guilin	-1.502	NaN	NaN
Jilin	1.186	0.431	0.144	Haikou	NaN	NaN	NaN
Maanshan	1.104	0.562	0.418	Zhuhai	NaN	NaN	NaN
Jiujiang	1.304	0.660	0.321	Zhanjiang	NaN	NaN	NaN

Note: $c = h/\hat{b}$, where $\hat{b} = 2\hat{\sigma}/\sqrt{n}$, where $\hat{\sigma}$ of the pollutant concentration. For more details, see Section 4.9.

Table A.15: t-Statistics for McCrary (2008) Test: Cities Exhibiting Manipulation of SO2

c	10	15	20
Shijiazhuang	-4.9154	-5.4971	-5.7896
Anshan	-4.3499	-5.1613	-5.5384
Yangquan	-3.7741	-4.7820	-5.4176
Zunyi	-3.1786	-4.6268	-5.1011
Tangshan	-2.8904	-3.8594	-4.2408
Linfen	-2.7670	-3.5816	-3.1446
Weifang	-2.1114	-3.1647	-4.0624
Handan	-3.4122	-3.0500	-2.6932
Shizuishan	-2.2226	-2.9100	-3.5842
Taiyuan	-1.2444	-2.7966	-3.9241
Dalian	-2.7057	-2.6835	-2.6973
Tianjin	-1.6428	-2.4047	-2.3984
Zhengzhou	-1.4686	-2.1941	-2.6340
Kunming	-1.9936	-2.1755	-2.0017
Guiyang	-1.5683	-2.0856	-2.5199
Baoding	-1.3858	-1.8321	-2.2295
Jinzhou	-1.3630	-1.7740	-1.7233
Kaifeng	-0.5515	-1.7361	-1.6730
Baotou	-1.4601	-1.6942	-1.7832
Wulumuqi	-1.3142	-1.5663	-1.9597
Changzhi	-1.0818	-1.4499	-1.6063
Shenyang	-0.3135	-1.3845	-2.1399
Jinan	-0.9440	-1.3724	-1.2364
Zaozhuang	-1.3397	-1.3693	-1.3118
Datong	-1.3776	-1.3352	-0.4627
Weinan	-0.4897	-1.3234	-2.0854

Note: $c = h/\hat{b}$, where $\hat{b} = 2\hat{\sigma}/\sqrt{n}$,
where $\hat{\sigma}$ of the pollutant concentration.
For more details, see Section 4.9.

Table A.16: t-Statistics for McCrary (2008) Test: Cities Not Exhibiting Manipulation of SO2

c	10	15	20	c	10	15	20
Shaoguan	-0.9894	-1.2570	-1.2676	Jiujiang	0.2117	0.3158	0.5675
Lianyungang	-1.4755	-1.2362	-0.9770	Jining	0.0000	0.3327	0.3314
Luoyang	-1.3838	-1.1993	-1.0225	Xuzhou	0.3861	0.4054	0.2574
Jiaozuo	-0.5994	-1.1044	-1.2153	Pingdingshan	1.2480	0.4447	-0.2356
Shanghai	-1.4084	-1.0962	-0.8065	Luzhou	0.7746	0.4726	0.4759
Yichang	-1.3179	-1.0409	-1.0016	Deyang	0.8053	0.4991	0.4018
Changde	-0.8762	-1.0348	-1.4034	Yinchuan	2.1407	0.5198	-0.0946
Nanning	-1.8509	-1.0309	-0.5002	Anyang	0.9737	0.5359	0.4100
Chengdou	-0.9040	-1.0008	-0.6651	Yanan	1.0594	0.7029	0.6957
Nanjing	-0.0143	-0.8299	-1.4825	Xian	1.2896	0.7060	-0.0103
Hangzhou	-0.9132	-0.8118	-0.8383	Wuxi	0.6119	0.7361	0.6726
Sanmenxia	-0.1681	-0.6977	-1.3730	Changsha	0.8909	0.7669	0.5281
Zigong	-1.2222	-0.5473	-0.0462	Huhehaote	1.1563	0.8722	0.3901
Mianyang	-0.2584	-0.5321	-0.2596	Yantai	0.6084	0.9392	1.0544
Wuhan	-0.3176	-0.5151	-0.5834	Yueyang	1.2358	1.0466	0.8286
Guangzhou	-0.0275	-0.5022	-0.5948	Zhenjiang	0.0000	1.0587	0.1110
Panzhihua	-0.7828	-0.4882	-0.7169	Jinchang	1.4615	1.3061	1.5640
Xiangtan	0.1137	-0.4817	-0.7929	Xianyang	1.2651	1.3100	1.7365
Zibo	-0.2277	-0.3441	-0.6712	Lanzhou	1.9271	1.3665	1.1234
Qingdao	-0.2773	-0.2313	-0.2569	Liuzhou	1.9243	1.5305	1.0088
Fushun	-0.6328	-0.2145	-0.5761	Chifeng	1.6708	1.6384	1.7937
Beijing	0.6865	-0.1772	-0.3474	Chongqing	2.3330	1.7463	1.3168
Nanchang	0.7248	-0.1153	-0.5895	Qinhuangdao	1.7523	2.0749	2.6185
Yangzhou	NaN	-0.0623	3.4115	Zhuzhou	2.6425	2.7285	2.7846
Qujing	0.2022	-0.0097	-0.0458	Yibin	2.2339	2.8071	2.8170
Benxi	0.7944	0.0335	-0.3721	Tongchuan	2.8701	2.9114	2.8250
Shaoxing	-0.1967	0.0434	0.1905	Nantong	NaN	5.8736	1.4046
Baoji	NaN	0.1486	-0.3088				

Note: $c = h/\hat{b}$, where $\hat{b} = 2\hat{\sigma}/\sqrt{n}$, where $\hat{\sigma}$ of the pollutant concentration.
For more details, see Section 4.9.

Table A.17: Comparison of Results in Applying the McCrary test on Concentrations as opposed to the Air Pollution Index

	Concentration	API
No. of Manipulators ($\alpha = 0.05$)		
PM_{10}	52	50
SO_2	19	4
NO_2	5	7
Ranking (PM_{10})		
Beijing	4	7
Chongqing	13	15
Shanghai	91	95
Tianjin	15	26

This table reports the difference in the results between using pollutant concentration versus API in the McCrary test.

Table A.18: Comparison of the Results of the Panel Matching Approach (PMA) and the McCrary Test

City-pair	PMA	McCrary (2008)	
	Rejections	City 1	City 2
Zhenjiang-Yangzhou	NO	Borderline	NO
Changzhou-Wuxi	NO	YES	NO
Jilin-Changchun	NO	NO	YES
Kaifeng-Zhengzhou	YES	YES	NO
Zhuzhou-Xiangtan	YES	NO	YES
Quanzhou-Xiamen	YES	YES	NO
Hangzhou-Shaoxing	YES	NO	NO
Shenyang-Fushun	YES	YES	NO
Yinchuan-Shizuishan	YES	Borderline	NO
Xian-Xianyang	YES	YES	Borderline
Jinan-Taian	YES	YES	NO
Huhehaote-Baotou	YES	NO	YES

Bibliography

Abadir, Karim M. and Jan R. Magnus. 2005. *Matrix Algebra*. Cambridge: Cambridge UP.

Altonji, Joseph and Rosa Matzkin. 2005. "Cross-section and Panel Data Estimators for Nonseparable Models with Endogenous Regrssors." *Econometrica* 73 (3):1053–1102.

Anderson, T.W. 1962. "On the Distribution of the Two-Sample Cramer von Mises Criterion." *The Annals of Mathematical Statistics* 33 (3):1148–1159.

Andrews, D.W.K. 1997. "A Conditional Kolmogorov Test." *Econometrica* 65 (5):1097–1128.

Andrews, Steven Q. 2008a. "Beijing Plays Air Quality Games." *Far Eastern Economic Review* .

Andrews, Steven Q. 2008b. "Inconsistencies in air quality metrics: 'Blue Sky' days and PM 10 concentrations in Beijing." *Environmental Research Letters* 3 (3):034009.

Angrist, Joshua. 2004. "Treatment heterogeneity in theory and practice." *The Economic Journal* 114 (494):C52–C83.

Angrist, Joshua D. and Whitney K. Newey. 1991. "Over-identification Tests in Earnings Functions with Fixed Effects." *Journal of Business and Economic Statistics* 9 (3):317–323.

Arellano, Manuel. 2003. *Panel Data Econometrics*. Oxford: Oxford UP.

Arellano, Manuel and Stefan Bonhomme. 2009. "Identifying Distributional Characteristics in Random Coefficients Panel Data Models." *CEMFI* .

Arellano, Manuel and Jinyong Hahn. 2005. "Understanding Bias in Nonlinear Panel Models: Some Recent Developments." *CEMFI Working Paper No. 0507* .

Athey, Susan and Guido Imbens. 2006. "Identification and Inference of Nonlinear Difference-in-Difference Models." *Econometrica* 74 (2):431–497.

Bandyopadhyay, Sushenjit and John Horowitz. 2006. "Do Plants Overcomply with Water Pollution Regulations? The Role of Discharge Variability." *The B.E. Journal of Economic Analysis & Policy (Topics)* 6 (1):4.

Bester, C. Alan and Christian Hansen. 2009. "Identification of Marginal Effects in a Nonparametric Correlated Random Effects Model." *Journal of Business and Economic Statistics* 27 (2):235–250.

Brannlund, Runar and Karl-Gustaf Lofgren. 1996. "Emission Standards and Stochastic Waste Load." *Land Economics* 72 (2):218–230.

Browning, Martin and Jesus Carro. 2007. "Heterogeneity in Dynamic Discrete Choice Models." *Oxford Discussion Paper No. 287* .

Card, David. 1999. "The causal effect of education on earnings." *Handbook of Labor Economics* .

Chamberlain, Gary. 1982. "Multivariate regression models for panel data." *Journal of Econometrics* 18 (1):5–46.

———. 1984. "Panel Data." *Handbook of Econometrics, edited by Z. Griliches and M. Intriligator* 2:pp.1247–1318.

———. 2010. "Binary Response Models for Panel Data: Identifcation and Information." *Econometrica* 78 (1):159–168.

Chay, Kenneth Y. and Michael Greenstone. 2005. "Does Air Quality Matter? Evidence from the Housing Market." *Journal of Political Economy* 113 (2):pp. 376–424.

Chen, Ye, Hongbin Li, and Li-An Zhou. 2005. "Relative performance evaluation and the turnover of provincial leaders in China." *Economics Letters* 88 (3):421 – 425.

Chen, Yuyu, Ginger Zhe Jin, Naresh Kumar, and Guang Shi. 2012. "Gaming in Air Pollution Data? Lessons from China." *The B.E. Journal of Economic Analysis & Policy (Advances)* 13 (3):Article 2.

Chernozhukov, Victor, Ivan Fernandez-Val, Jinyong Hahn, and Whitney Newey. 2010. "Average and Quantile Effects in Nonseparable Panel Data Models." *Econometrica, forthcoming* .

———. 2013. "Average and Quantile Effects in Nonseparable Panel Data Models." *Econometrica* 81 (2):pp.535–580.

Christiansen, Charlotte, Juanna Schroeter Joensen, and Jesper Rangvid. 2008. "Are Economists More Likely to Hold Stocks?" *Review of Finance* 12:pp. 465–496.

Congleton, Roger D. 1992. "Political Institutions and Pollution Control." *The Review of Economics and Statistics* 74 (3):412–21.

Davydov, Y.A., M.A. Lifshits, and N.V. Smorodina. 1998. *Local Properties of Distributions of Stochastic Functionals*. providence: American Mathematical Society.

Debaere, Peter and Shalah Mostashari. 2005. "Do Tariffs Matter for the Extensive Margin of International Trade? An Empirical Analysis." *CEPR Discussion Paper No. 5260* .

der Vaart, Aad Van and Jon Wellner. 2000. *Weak Convergence and Empirical Processes*. New York: Springer-Verlag.

der Vaart, A.W. Van. 1998. *Asymptotic Statistics*. Cambridge: Cambridge UP.

Earnhart, Dietrich. 2007. "Effects of permitted effluent limits on environmental compliance levels." *Ecological Economics* 61 (1):178 – 193.

Evdokimov, Kirill. 2010. "Identification and Estimation of a Nonparametric Panel Data Model with Unobserved Heterogeneity." *Department of Economics, Princeton University* .

———. 2011. "Nonparametric Identification of a Nonlinear Panel Model with Application to Duration Analysis with Multiple Spells." *Department of Economics, Princeton University* .

Fernandez-Val, Ivan. 2009. "Fixed Effects Estimation of Structural Parameters and Marginal Effects in Panel Probit Models." *Journal of Econometrics* 150 (1):pp. 71–85.

Gibbons, Charles, Juan Carlos Serrato, and Michael Urbancic. 2011. "Broken or Fixed Effects?"

Gibbons, Jean. 1985. *Nonparametric Statistical Inference*. New York: Marcel Dekker, Inc.

Graff Zivin, Joshua S. and Matthew J. Neidell. 2012. "The Impact of Pollution on Worker Productivity." *American Economic Review* Forthcoming (17004).

Graham, Bryan and James Powell. 2012. "Identification and estimation of average partial effects in 'irregular' correlated random coefficient panel data models." *Econometrica* 80 (5):2105–2152.

Guo, Jian-Ping, Xiao-Ye Zhang, Hui-Zheng Che, Sun-Ling Gong, Xingqin An, Chun-Xiang Cao, Jie Guang, Hao Zhang, Ya-Qiang Wang, Xiao-Chun Zhang, Min Xue, and Xiao-Wen Li. 2009. "Correlation between PM concentrations and

aerosol optical depth in eastern China." *Atmospheric Environment* 43 (37):5876 – 5886.

Hahn, Jinyong and Guido Kuersteiner. 2011. "Bias Reduction for Dynamic Nonlinear Panel Data with Fixed Effects." *Econometric Theory* 27 (6):pp.1152–1191.

Hahn, Jinyong and Whitney Newey. 2004. "Jackknife and Analytical Bias Reduction for Nonlinear Panel Models." *Econometrica* 72 (4):pp. 1295–1319.

Heckman, James. 1981. "The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating a Discrete Time-Discrete Data Stochastic Process and Some Monte Carlo Evidence." *Structural Analysis of Discrete Data, edited by D.l McFadden and C. Manski* .

Heckman, James and T.E. MaCurdy. 1980. "A life cycle model of female labor supply." *Review of Economic Studies* 47:47–74.

———. 1982. "Corrigendum on: A life cycle model of female labor supply." *Review of Economic Studies* 49:659–660.

Heckman, James and Richard Robb. 1985. "Alternative Methods for Evaluating the Impact of Interventions: An Overview." *Journal of Econometrics* :239–267.

Hoderlein, Stefan and Halbert White. 2012. "Nonparametric Identification of Nonseparable Panel Data Models with Generalized Fixed Effects." *Journal of Econometrics* 168 (2):300–314.

Honore, B. and E. Kyriazydou. 2000. "Panel Discrete Choice Models with Lagged Dependent Variables." *Econometrica* 68 (4):839–874.

Hsiao, Cheng. 2003. *Analysis of Panel Data*. Cambridge: Cambridge UP.

Hyslop, D.R. 1999. "State dependence, serial correlation and heterogeneity in intertemporal labor force participation of married women." *Econometrica* 67 (6).

James, W. and C. Stein. 1961. "Estimation with Quadratic Loss." *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* 1:pp. 361–379.

Keys, Benjamin J., Tanmoy Mukherjee, Amit Seru, and Vikrant Vig. 2010. "Did Securitization Lead to Lax Screening? Evidence from Subprime Loans." *The Quarterly Journal of Economics* 125 (1):307–362.

Kim, Tae-Hwan and Halbert White. 2000. "James-Stein Type Estimators in Large Samples with Application to the Least Absolute Deviations Estimator." *UC San Diego Discusssion Paper No. 99-04R* .

Koech, Janet and Jian Wang. 2012. "Chinas Slowdown May Be Worse Than Official Data Suggest." *Federal Reserve Bank of Dallas Economic Letter* 7 (8):1–4.

Koop, Gary and Lise Tole. 2004. "Measuring the health effects of air pollution: to what extent can we really say that people are dying from bad air?" *Journal of Environmental Economics and Management* 47 (1):30 – 54.

Lancaster, Tony. 2002. "Orthogonal Parameters and Panel Data." *The Review of Economic Studies* 69 (3):pp. 647–666.

Li, Hongbin and Li-An Zhou. 2005. "Political turnover and economic performance: the incentive role of personnel control in China." *Journal of Public Economics* 89 (910):1743 – 1762.

Magnac, Thierry. 2004. "Panel Binary variables and sufficiency: Generalizing conditional logit." *Econometrica* 76 (6):1859–1876.

Malm, William C. 1999. *Introduction to Visibility*. Fort Collins, CO: Cooperative Institute for Research in the Atmosphere (CIRA), NPS Visibility Program, Colorado State University.

Matus, Kira, Kyung-Min Nam, Noelle E. Selin, Lok N. Lamsal, John M. Reilly, and Sergey Paltsev. 2012. "Health damages from air pollution in China." *Global Environmental Change* 22 (1):55 – 66.

McCrary, Justin. 2008. "Manipulation of the running variable in the regression discontinuity design: A density test." *Journal of Econometrics* 142 (2):698–714.

Mundlak, Yair. 1961. "Empirical Production Function Free of Management Bias." *Journal of Farm Economics* 43 (1):pp.44–56.

———. 1978. "On the Pooling of Time Series and Cross Section Data." *Econometrica* .

Natural Resources Defense Council. 2009. "Amending China's Air Pollution Prevention And Control Law." Available at http://www.nrdc.com/.

Newey, Whitney. 1991. "Uniform Convergence in Probability and Stochastic Equicontinuity." *Econometrica* 59 (4):pp. 1161–1167.

Neyman, J. and E. Scott. 1948. "Consistent estimates based on partially consistent observation." *Econometrica* 16:pp. 132.

Papke, Leslie and Jeffrey Wooldridge. 2008. "Panel data methods for fractional response variables with an application to test pass rates." *Journal of Econometrics* 145:pp.121–133.

Quessy, Jean-Francois and Fancois Ethier. 2012. "Cramer-von Mises and characteristic fuunction tests for the two and k-samlpe problems with dependent data." *Computational Statistics and Data Analysis* 56 (6):2097–2111.

Rawski, Thomas G. 2001. "What is happening to China's GDP statistics?" *China Economic Review* 12 (4):347 – 354.

Rilstone, P., V.K. Srivastava, and Aman Ullah. 1996. "The Second-Order Bias and Mean Square Error of Nonlinear Estimators." *Journal of Econometrics* 75:pp. 369–395.

Romano, Joseph P. and Azeem M. Shaikh. 2006. "Stepup procedures for control of generalizations of the familywise error rate." *Annals of Statistics* 34 (4):1850–1873.

Shih, Victor, Christopher Adolph, and Mingxing Liu. 2012. "Getting Ahead in the Communist Party: Explaining the Advancement of Central Committee Members in China." *American Political Science Review* 106:166–187.

Shimshack, Jay P. and Michael B. Ward. 2008. "Enforcement and overcompliance." *Journal of Environmental Economics and Management* 55 (1):90 – 105.

Sloane, Christine S. and Warren H. White. 1986. "Visibility: An evolving number." *Environmental Science & Technology* 20 (8):760–766.

Stein, C. 1955. "Inadmissibility of the Usual Estimator for the Mean of a Multivariate Distribution." *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* 1:pp. 197–206.

Wooldridge, Jeffrey. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge: MIT Press.

World Bank. 2007. "Cost of Pollution in China." World Bank, East Asia and Pacific Region, Available at http://www.worldbank.org/eapenvironment.

Wouterson, Tiemen. 2002. "Robustness Against Incidental Parameters."

Zhang, Junjie. 2008. "Production from Natural Resources with Latent Abundance Information: The Case of the Fishery."

———. 2012. "Delivering Environmentally Sustainable Economic Growth: The Case of China." Asia Society Policy Report, Available at http://asiasociety.org/policy/.

Zhang, Junjie and Can Wang. 2011. "Co-benefits and additionality of the clean development mechanism: An empirical analysis." *Journal of Environmental Economics and Management* 62 (2):140 – 154.

Zhang, Qingfeng and Robert Crooks. 2012. *Toward an Environmentally Sustainable Future: Country Environmental Analysis of the Peoples Republic of China.* Mandaluyong City, Philippines: Asian Development Bank.

Zheng, Siqi, Matthew E. Kahn, Weizeng Sun, and Danglun Luo. 2013. "Incentivizing China, s Urban Mayors to Mitigate Pollution Externalities: The Role of the Central Government and Public Environmentalism." NBER Working Papers 18872, National Bureau of Economic Research, Inc.