Essays on Nonparametric Methods in Econometrics

Takahide Yanagi

Graduate School of Economics, Kyoto University

February, 2015

Contents

Chapter 1

Introduction

This dissertation develops three research topics on nonparametric methods in econometrics. First, we present a nonparametric panel data analysis in which dynamic structure of panel data is heterogeneous across individuals. Second, we develop a nonparametric analysis for the sharp regression discontinuity design in which the forcing variable contains measurement error. Third, we study a nonparametric identification analysis in weakly separable models with a binary endogenous explanatory variable.

The analyses based on nonparametric methods are important research consideration in econometrics and statistics. The analyses that involve a known family of probability distributions with unknown finite dimensional parameters (e.g., finite dimensional linear coefficients) are referred as *parametric* analyses. In contrast, the *nonparametric* analyses involve unknown probability distributions with unknown infinite dimensional parameters (e.g., unknown structural functions). Many studies have devoted the literature of nonparametric econometric analyses because of the utility and the flexibility. The recent contributions on the literature are surveyed in Blundell and Powell (2003), Chesher (2007a), and Matzkin (2007) for the nonparametric identification in econometrics and in Pagan and Ullah (1999), Ichimura and Todd (2007), Li and Racine (2007), and Horowitz (2010) for the nonparametric estimation in econometrics.

There are several reasons why the nonparametric analyses are important in econometrics. First, the nonparametric analyses do not generally force the researchers to assume implausible or restrictive assumptions. Economic theory rarely tells the researchers the true probability distribution of the unobservables and/or the true functional form of the structural functions. This issue is emphatically discussed in Reiss and Wolak (2007). For this reason, the researchers who focus on identifying the parameters in an econometric model based on the parametric methods are often forced to introduce somewhat arbitrary and/or restrictive parametric assumptions. In contrast, the nonparametric analyses generally allow the researchers to execute the econometric analyses with unknown probability distributions and unknown structural functions. Indeed, many theoretical studies develop the analyses to identify econometric parameters without introducing any parametric specifications (see, e.g., Matzkin, 2007). Accordingly, the nonparametric analyses are often executed under weaker assumptions than the parametric analyses, so that they are helpful for the researchers who hesitate to introduce restrictive assumptions.

Second, the parametric analyses are sensitive to the problems caused from misspecifications, while the nonparametric analyses are robust to such problems. In a parametric analysis, if the parametric specifications are incorrect, then the parametric analysis generally leads the researchers to imprecise empirical results. For example, consider a researcher who considers the traditional parametric linear specification $Y_i = X_i'\beta + \varepsilon_i$ to study the causal effect of X_i on Y_i, where $Y_i \in \mathbb{R}$ is a dependent variable, $X_i \in \mathbb{R}^k$ is an explanatory vector, $\beta \in \mathbb{R}^k$ is the parameter of interest, and $\varepsilon_i \in \mathbb{R}$ is an additive error term with $E(\varepsilon_i) = 0$ and $E(\varepsilon_i X_i) = 0$. The researcher can then get the ordinary least squares estimator $\hat{\beta}$ for β by regressing Y_i on X_i, i.e., $\hat{\beta} = (\sum_{i=1}^n X_i X_i')^{-1} \sum_{i=1}^n X_i Y_i$. If the linear functional specification is correct, $\hat{\beta}$ is consistent for β and the researcher can execute the statistical inference for the causal effect based on the linear specification. However, it is also well-known in the literature that if the linear specification is incorrect, then $\hat{\beta}$ is generally inconsistent for β and the statistical inference based on the linear specification is also invalid. Contrarily, the nonparametric methods such as the additive nonparametric regression (see, e.g., Blundell and Powell, 2003) allow the researcher to investigate the causal effect of Y_i on X_i without the functional specifications, so that they are robust to the misspecification problems.

Third, the nonparametric methods allow us to examine the correctness of the parametric specifications. Because the correctness of the parametric specifications is essential to appropriately execute the statistical inference based on the parametric models, tests for the parametric specifications are often demanded in empirical situations. For example, suppose that we would like to test the correctness of the parametric functional specification in a parametric regression model. In this case, we can test the correctness of the functional specification by comparing the difference in the fitted values for the dependent variable based on the parametric regression and based on the nonparametric kernel regression (see, e.g., Li and Racine, 2007, Chapter 12). The availability of such specification tests is important because the correctness of the parametric specifications is essential in the parametric models and because economic theory is not generally informative about the parametric specifications in the econometric models.

This dissertation builds on the important and growing literature of the nonparametric methods in econometrics by providing three contributions in the following chapters.

Chapter 2 (joint with Ryo Okui) proposes the nonparametric analysis of panel data whose dynamic structure is heterogeneous across individuals. The aim in the chapter is to estimate the cross-sectional distributions and/or some distributional features of the heterogeneous mean and autocovariances without assuming any specific model for the dynamics. The asymptotic properties of the proposed estimators are investigated using double asymptotics under which both the cross-sectional sample size (N) and the length of the time series (T) tend to infinity. We prove the functional central limit theorem for the empirical process of the proposed distribution estimator. By using the functional delta method, we also derive the asymptotic distributions of the estimators for various parameters of interest. We show that the distribution estimator exhibits a bias whose order is proportional to $1/\sqrt{T}$. Conversely, when the parameter

of interest can be written as the expectation of a smooth function of the heterogeneous mean and/or autocovariances, the bias is of order $1/T$ and can be corrected by the jackknife method. The results of Monte Carlo simulations show that our asymptotic results are informative regarding the finite-sample properties of the estimators. They also demonstrate that the proposed jackknife bias correction is successful.

Chapter 3 develops a nonparametric analysis for the sharp regression discontinuity (RD) design in which the continuous forcing variable may contain measurement error. We show that if the observable forcing variable contains measurement error, the measurement error causes severe identification bias for the average treatment effect given the "true" forcing variable at the discontinuity point. The bias is critical in the sense that even if actually there is significant causal effect, it misleads the researchers into the incorrect conclusion in which there is no causal effect. Furthermore, the measurement error leads the conditional probability of the treatment to be continuous at the threshold. To investigate the average treatment effect using the mismeasured forcing variable, we propose approximating it by the small error variance approximation (SEVA) originally developed by Chesher (1991). Based on the SEVA, the average treatment effect is approximated up to the order of the variance of the measurement error by an identified parameter when the variance is small. We also develop an estimation procedure for the parameter that approximates the average treatment effect based on local polynomial regressions and kernel density estimation. Monte Carlo simulations reveal the severity of the identification bias due to the measurement error and demonstrate that our approximate analysis is successful.

Chapter 4 presents a new identification analysis for the structural function in weakly separable models (Vytlacil and Yildiz, 2007) with a binary endogenous explanatory variable. We focus on identifying the values of the structural function at specified values of the explanatory variables and specific quantiles of the unobservables, in the manner of Chesher (2005) and Jun, Pinkse, and Xu (2011). Our identification analysis is composed of two steps. Firstly, under weak conditions, we provide partial identification results for the values of the structural function based on the modified control variate approach in Chesher (2005). Secondly, by utilizing the results of the first step and the weak separability of the structural function, we show that the values of the structural function are intervally identified by tighter bounds. This second step is based on, using the instrumental variables, identifying the values of the explanatory variables that lead to the values of the structural function being no greater or no less than the value of the structural function we wish to identify. A simple example demonstrates that the identification analysis leads to an informative identification bound for a value of the structural function that is not intervally identified by the existing analysis.

Chapter 2

Panel Data Analysis with Heterogeneous Dynamics (joint with Ryo Okui)

2.1 Introduction

This paper considers the analysis of panel data whose dynamic structure is stationary across time but heterogeneous across individuals. We propose methods for estimating the distributional features of the mean and autocovariances that are heterogeneous across individuals using panel data. Our estimation procedure is simple to implement. We first estimate the mean or autocovariances for each individual. We then estimate the distribution and other distributional quantities using the empirical distribution of the estimated mean or autocovariances. When the parameter of interest can be written as the expected value of a smooth function of the heterogeneous mean or autocovariances, the jackknife method reduces the bias of the estimator.

Understanding the dynamic nature of an economic variable that is potentially heterogeneous when using panel data is an important research consideration in economics. For example, there is considerable study using panel data on income dynamics (see, e.g., Lillard and Willis, 1978, Meghir and Pistaferri, 2004, Guvenen, 2007, and Browning, Ejrnæs, and Alvarez, 2010, among many others). In particular, Browning et al. (2010) show that income dynamics exhibit considerable heterogeneity in that an income shock may have a persistent effect on the future income profiles of some individuals, whereas for others, the effect may disappear quite quickly.

The contribution of this paper is to propose easy-to-implement methods to analyze panel data whose dynamics are heterogeneous without assuming any specific model. To study the heterogeneous dynamic structure, we investigate the cross-sectional distributions of the mean and autocovariances that are heterogeneous across individuals. Investigating these quantities does not depend on a particular model structure. While the literature on dynamic panel data analysis is already voluminous, many studies assume some specific model for the dynamics (such as the autoregressive (AR) model) and the homogeneity in the dynamics, allowing heterogeneity only in the mean of the process.[1] While several analyses consider either heterogeneous dynamics

[1] See, e.g., Baltagi (2008) and Arellano (2003) for excellent reviews of the more important existing contribu-

or model-free analysis (see the section "Related literature" below), we are unaware of any specific study that proposes methods to analyze heterogeneous dynamics using panel data without specifying some particular model. This paper builds on the literature by proposing model-free analysis for a heterogeneous dynamic structure.

The distributions of the heterogeneous mean and autocovariances are informative in various ways. First, the mean and the autocovariances are perhaps the most basic descriptive statistics for dynamics. Indeed, a typical first step in analyzing time-series data is to examine the mean and the autocovariance (or autocorrelation) properties of the data. We believe that the distributions of the heterogeneous mean and autocovariances would also be useful descriptive statistics for understanding the dynamics in panel data analysis. Second, we can use the mean and autocovariances to investigate whether different groups possess dissimilar dynamic structures without relying on some particular model. For example, consider the situation in which we would like to investigate whether males and females face different income dynamics, but we are also aware of the fact that income dynamics are heterogeneous across individuals. In such a case, we can estimate the distributions of the autocovariances for males and females separately and compare them to see if they indeed differ.

Our approach is to estimate the mean and autocovariances for each individual and use the empirical distributions of the estimated mean and autocovariances to estimate the cross-sectional distributions of the heterogeneous mean and autocovariances and other quantities of interest, such as the quantile function. The asymptotic properties of the empirical distributions are derived based on double asymptotics under which both the number of cross-sectional observations, N, and the length of the time series, T, tend to infinity. By using empirical process theory (see, e.g., van der Vaart and Wellner, 1996), we show that the empirical distributions converge weakly to Gaussian processes. However, the condition $N/T \to 0$ is required for this weak convergence because of the bias caused by the estimation error in the estimated mean and autocovariances for each individual. We also derive the asymptotic distributions of the estimators for other distributional characteristics, including quantiles, using the functional delta method.

When the parameter of interest can be written as the expected value of a smooth function of the heterogeneous mean or autocovariances, the condition on the relative magnitudes of N and T can be relaxed. This class of parameters includes the mean, the variance, and other moments of the heterogeneous mean and autocovariances. In this case, the bias becomes of order $O(1/T)$, and the condition $N/T^2 \to 0$ is sufficient for asymptotically unbiased estimation. Moreover, we can analytically evaluate the bias, and jackknife bias correction is available. This bias has two sources. The first is the incidental parameter problem originally discussed in Neyman and Scott (1948) and now well known in the econometrics literature. This type of bias does not affect the estimated mean, but does influence the estimated autocovariances. When we estimate the autocovariance for each individual, we also need to estimate the mean for

tions on dynamic panel data analysis.

each individual. Because there are N individual-specific mean parameters to be estimated, this creates incidental parameter bias. The second source of bias arises when the smooth function is nonlinear. This bias affects both the mean and the autocovariances. However, this source of bias does not appear when the parameter of interest is the mean of the heterogeneous mean or autocovariances because the corresponding function is linear. We propose using the half-panel jackknife in Dhaene and Jochmans (2014) to correct the bias.

We also conduct Monte Carlo simulations to investigate the finite-sample properties of the proposed procedures. The results of the Monte Carlo simulations show that the asymptotic analyses in this paper are informative regarding the finite-sample properties of the proposed estimators. They show that the estimators based on the estimated autocovariances have severe bias when T is small compared with N, but the bias decreases as T increases. They also show that the proposed jackknife bias correction decreases this bias. The half-panel jackknife also reduces the bias allocated with the incidental parameter problem and the nonlinearity of the smooth function, even when T is relatively small.

Related literature: This paper most closely relates to the literature on heterogeneous panel AR models. In these models, we capture the heterogeneity in the dynamics by allowing the AR coefficients to be individual specific. The panel AR(1) models with individual-specific AR coefficients are analyzed by, for example, Pesaran and Smith (1995), Hsiao, Pesaran, and Tahmiscioglu (1999), and Pesaran, Shin, and Smith (1999). These analyses are extended to nonstationary panel data by Phillips and Moon (1999), while Pesaran (2006) considers models with a multifactor error structure. The present analysis differs in two ways from these earlier studies. First, we do not assume any specific model to describe the dynamics, while the abovementioned studies consider an AR or linear-process specification. Second, our aim is to estimate the entire distribution of the mean or autocovariances, which are heterogeneous across individuals. In contrast, Pesaran and Smith (1995) and others focus on the estimation of the means of the AR coefficients.

Elsewhere, Mavroeidis, Sasaki, and Welch (2014) consider the identification and estimation of the distribution of the AR coefficients in heterogeneous panel AR models. The advantage of their approach is that T can be fixed, and thus it is applicable to short panels. While we consider the case in which $T \to \infty$, our method is much simpler to implement. We simply need to estimate the mean and autocovariances for each individual and compute the empirical distributions of the estimated mean and autocovariances. By contrast, the estimation method in Mavroeidis et al. (2014) requires the maximization of a kernel-weighting function that is written as an integration over multiple variables. We also emphasize that our method does not depend on model specification. In addition, we note that identification of the distributions of the heterogeneous mean and autocovariances is trivial in our setting because we consider the setting $T \to \infty$. Alternatively, the identification analysis in Mavroeidis et al. (2014) is mathematically involved because they consider fixed T.

Several studies propose model-free methods to investigate the dynamic structure using panel data. For example, Okui (2010, 2011, 2014) considers the estimation of autocovariances using long panel data and assumes that the autocovariance structure is homogeneous across individuals. By contrast, our paper considers a heterogeneous structure. However, we note that it is easy to show that Okui's autocovariance estimator is equivalent to the estimator of the mean of the heterogeneous autocovariances. In other work, Lee, Okui, and Shintani (2013) consider infinite-order panel AR models. Given we can represent a stationary time series by an infinite-order AR process under mild conditions, their approach is essentially model-free. However, they assume that the dynamics are homogeneous.

A different line of research investigates the properties of the estimators for model-based analysis when the assumed model is possibly misspecified. For instance, Okui (2008) examines the probability limits of various estimators for panel AR(1) models when the true dynamics do not follow an AR(1) process and assumes homogeneity in the dynamics, while the mean is allowed to be heterogeneous. Lee (2012) discusses the fixed effects estimator for panel AR models when the lag order is misspecified and also considers the case where the dynamics are homogeneous. Lastly, Galvao and Kato (2014) investigate the asymptotic properties of the fixed effects estimator in general regression models and allow the data-generating process to be generally heterogeneous. However, the purpose of the current study is to propose new methods to analyze panel data with heterogeneous dynamics, not to examine the properties of existing estimators.

The literature on deconvolution techniques examines the identification and estimation of the distribution of individual effects (see, e.g., Horowitz and Markatou, 1996, Székely and Rao, 2000, and Bonhomme and Robin, 2010). In the context of the present analysis, we may employ these deconvolution techniques to identify and estimate the distribution of the individual-specific mean with fixed T. That T can be fixed is an advantage of these techniques. However, our focus is on the distribution of not only the mean, but also the heterogeneous autocovariance. Moreover, we propose methods that are easily implemented under the requirement that $T \to \infty$. On the other hand, the deconvolution techniques involve the computation of the characteristic function, and the rate of convergence is remarkably slow.

While not directly connected, this paper is also somewhat related to the recent literature on random coefficient models. For example, Arellano and Bonhomme (2012) consider linear regression models with random coefficients in panel data analysis and discuss the identification and estimation of the distribution of random coefficients using deconvolution techniques. Note that Chamberlain (1992) and Graham and Powell (2012) consider a model similar to that of Arellano and Bonhomme (2012), but their focus is on the means of the random coefficients. Fernández-Val and Lee (2013) study moment restriction models with random coefficients and their generalized methods of moment estimation. Their analysis on the smooth function of individual effects is closely related to our analysis on the smooth function of means and autocovariances in terms of technique. Finally, Evdokimov (2009) considers a nonparametric panel

regression model with individual effects entering the unspecified structural function, but also relies on deconvolution techniques.

Organization of the paper: The remainder of the paper is organized as follows. Section 2.2 explains the setting. Section 2.3 introduces the proposed procedures. In Section 2.4, we derive the asymptotic properties of the distribution estimators. Section 2.5 considers the estimation of the expected value of a smooth function of the heterogeneous mean or autocovariances, the inference methods, and the jackknife bias correction. Section 2.6 presents some extensions based on the proposed procedures. Section 2.7 presents the results of the Monte Carlo simulations. Section 2.8 concludes the paper. All technical proofs are presented in the Technical appendix.

2.2 Settings

We observe panel data $\{\{y_{it}\}_{t=1}^{T}\}_{i=1}^{N}$, where y_{it} is a scalar random variable, i represents a cross-sectional unit, and t indicates a time period. The number of cross-sectional observations is N and the length of the time series is T. We consider situations in which both N and T are large. We assume that $\{y_{it}\}_{t=1}^{T}$ is independent across individuals.

The law of $\{y_{it}\}_{t=1}^{T}$ is assumed to be stationary, but its dynamic structure may be heterogeneous. To be specific, we consider the following data-generating process to model the heterogeneous dynamic structure. The unobserved individual effect, α_i, is independently drawn from a distribution common to all individuals. The time series $\{y_{it}\}_{t=1}^{T}$ for individual i is then drawn from some distribution $L(\{y_{it}\}_{t=1}^{T}; \alpha_i)$. The dynamic structure of y_{it} is heterogeneous because α_i varies across individuals. However, note that introducing the parameter α_i is a somewhat abstract way to represent heterogeneity in the dynamics across individuals. We do not directly assume anything about the distribution of α_i, because α_i does not explicitly appear in our analysis. For notational simplicity, we denote "$\cdot|\alpha_i$" by "$\cdot|i$"; that is, "conditional on α_i" becomes "conditional on i" below.

Our aim is to develop statistical tools to analyze the cross-sectional distributions of the heterogeneous mean and autocovariances of y_{it}. The mean for unit i is $\mu_i \coloneqq E(y_{it}|i)$. Note that μ_i is a random variable whose realization differs across individuals. This is because μ_i depends on α_i, which differs among individuals. As we assume stationarity, μ_i is constant over time. The distribution of μ_i represents heterogeneity in the mean of y_{it} across individuals. Let $\gamma_{k,i}$ be the k-th conditional autocovariance of y_{it} given α_i:

$$\gamma_{k,i} \coloneqq E\left((y_{it} - \mu_i)(y_{i,t-k} - \mu_i)|i\right).$$

In other words, $\gamma_{k,i}$ represents the k-th autocovariance of y_{it} for individual i. Note that $\gamma_{0,i}$ is the variance for individual i. Similarly to the case of μ_i, $\gamma_{k,i}$ is a random variable and its realization may be different among individuals. To understand the possibly heterogeneous dynamics of y_{it}, we aim to estimate quantities that characterize the distributions of μ_i and/or $\gamma_{k,i}$, such as the distribution function, the quantile function, and the moments.

Our setting is very general and includes many situations.

Example 2.1. The panel AR(1) model with heterogeneous coefficients considered by Pesaran and Smith (1995) and others is a special case of our setting. This model is

$$y_{it} = c_i + \phi_i y_{i,t-1} + \epsilon_{it},$$

where c_i and ϕ_i are the individual-specific intercept and slope coefficients, respectively, and ϵ_{it} follows a strong white noise process with variance σ^2. In this case, $\alpha_i = (c_i, \phi_i)$, $\mu_i = c_i/(1-\phi_i)$, and $\gamma_{k,i} = \sigma^2\phi_i^k/(1-\phi_i^2)$.

Example 2.2. Another example is the case in which y_{it} is generated by a linear process with heterogeneous coefficients:

$$y_{it} = c_i + \sum_{j=0}^{\infty} \theta_{j,i}\epsilon_{i,t-j},$$

where c_i and $\{\theta_{j,i}\}_{j=0}^{\infty}$ are heterogeneous coefficients and ϵ_{it} follows a strong white noise process with variance σ^2. In this example, $\alpha_i = (c_i, \{\theta_{j,i}\}_{j=0}^{\infty})$, $\mu_i = c_i$, and $\gamma_{k,i} = \sigma^2 \sum_{j=k}^{\infty} \theta_{j,i}\theta_{j-k,i}$.

Example 2.3. Our setting also includes cases in which the true data-generating process follows some nonlinear process. Suppose that y_{it} is generated by

$$y_{it} = m(\alpha_i, \epsilon_{it}),$$

where $m(\cdot,\cdot)$ is some function and ϵ_{it} is stationary over time and independent across individuals. In this case, $\mu_i = E(m(\alpha_i, \epsilon_{it})|\alpha_i)$ and $\gamma_{k,i}$ is the k-th order autocovariance of $w_{it} = y_{it} - \mu_i$ given α_i.

Our focus is on estimating the heterogeneous mean and autocovariance structure; we do not aim to recover the underlying structural form of the data-generating process. For example, even when y_{it} is generated by $y_{it} = m(\alpha_i, \epsilon_{it})$ as in the third example, we estimate not the function $m(\cdot,\cdot)$ but rather the heterogeneous mean and autocovariance structure only. We understand that addressing several important economic questions requires knowledge of the structural function of the dynamics. Nonetheless, we can estimate relatively easily the distribution of the heterogeneous mean and autocovariance without imposing strong assumptions. Moreover, the heterogeneous mean and autocovariance structure can provide valuable information, even if our ultimate goal is to identify the structural function.

2.3 Procedures

In this section, we present the statistical procedures used to estimate the distribution functions and other distributional characteristics of the heterogeneous mean and autocovariances of y_{it}. The proposed procedures are simple: we estimate the mean and autocovariances for each individual and then use their empirical distributions to estimate our parameter of interest. The following sections provide the theoretical justification for the proposed statistical procedures.

We first estimate the mean and autocovariances for each individual: μ_i and $\gamma_{k,i}$. We estimate these using the sample average and sample autocovariances:

$$\hat{\mu}_i \coloneqq \bar{y}_i \coloneqq \frac{1}{T}\sum_{t=1}^{T} y_{it},$$

and

$$\hat{\gamma}_{k,i} \coloneqq \frac{1}{T-k}\sum_{t=k+1}^{T}(y_{it}-\bar{y}_i)(y_{i,t-k}-\bar{y}_i).$$

We then compute the empirical distributions of $\{\hat{\mu}_i\}_{i=1}^N$ and $\{\hat{\gamma}_{k,i}\}_{i=1}^N$:

$$\mathbb{F}_N^{\hat{\mu}}(a) := \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}(\hat{\mu}_i \le a), \tag{2.1}$$

and

$$\mathbb{F}_N^{\hat{\gamma}_k}(a) := \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}(\hat{\gamma}_{k,i} \le a), \tag{2.2}$$

where $\mathbf{1}(\cdot)$ is the indicator function and $a \in \mathbb{R}$. These empirical distributions are interesting in their own right because they are estimators of the cross-sectional distribution functions of μ_i and $\gamma_{k,i}$, respectively. Let F_0^{μ} and $F_0^{\gamma_k}$ denote the distribution functions of μ_i and $\gamma_{k,i}$, respectively, so that $F_0^{\mu}(a) := \Pr(\mu_i \le a)$ and $F_0^{\gamma_k}(a) := \Pr(\gamma_{k,i} \le a)$. In Section 2.4, we show the consistency of $\mathbb{F}_N^{\hat{\mu}}$ and $\mathbb{F}_N^{\hat{\gamma}_k}$ for F_0^{μ} and $F_0^{\gamma_k}$, respectively, and derive the asymptotic distributions of $\mathbb{F}_N^{\hat{\mu}}$ and $\mathbb{F}_N^{\hat{\gamma}_k}$ under the condition $N/T \to 0$.

Remark 2.1. The condition $N/T \to 0$ implies that the length of the time series T is large relative to the number of cross-sectional observations N. Accordingly, the analysis based on the distribution function would be more suitable for macroeconomic data than for microeconomic data. Macroeconomic panel data, such as multi-country panels or state-level panels, may include a sufficiently long period compared with the cross-sectional sample size.

We can estimate other distributional quantities based on the empirical distributions of $\hat{\gamma}_{k,i}$ or $\hat{\mu}_i$. For example, consider the estimation of quantiles of $\gamma_{k,i}$. Let q_τ be the τ-th quantile of $\gamma_{k,i}$: $q_\tau := \inf\{a : F_0^{\gamma_k}(a) \ge \tau\}$. This is estimated by the τ-th quantile of $\hat{\gamma}_{k,i}$ so that $\hat{q}_\tau := \inf\{a : \mathbb{F}_N^{\hat{\gamma}_k}(a) \ge \tau\}$. Using the functional delta method, we derive the asymptotic distribution of the quantile estimator when $N, T \to \infty$ with $N/T \to 0$.

We can also test parametric specifications of the distribution of the heterogeneous mean or autocovariances based on the empirical distribution. Moreover, we can examine the difference of the heterogeneous dynamic structures across distinct groups based on the empirical distributions. The tests are based on Kolmogorov–Smirnov statistics based on the empirical distributions. We develop these tests in Section 2.6.

When the parameter of interest is the expectation of a *smooth* function of μ_i or $\gamma_{k,i}$, the condition on the relative magnitudes of N and T can be relaxed. Suppose that we are interested in $G_\mu \coloneqq E(g(\mu_i))$, where $g(\cdot)$ is a known function. We estimate G_μ by

$$\hat{G}_\mu := \frac{1}{N}\sum_{i=1}^{N} g(\hat{\mu}_i). \tag{2.3}$$

When our parameter of interest is $G_{\gamma_k} \coloneqq E(g(\gamma_{k,i}))$, it is estimated by

$$\hat{G}_{\gamma_k} := \frac{1}{N}\sum_{i=1}^{N} g(\hat{\gamma}_{k,i}). \tag{2.4}$$

Suppose that $g(\cdot)$ is twice continuously differentiable with a bounded second derivative. For example, the mean of $\gamma_{k,i}$ satisfies this condition because, for the mean, g is the identity function. The theoretical results in Section 2.5 show that this estimator is consistent as $N, T \to \infty$ and that $\sqrt{N}(\hat{G}_a - G_a)$ for $a = \mu$ or γ_k is asymptotically normal with mean zero when $N/T^2 \to 0$.

Remark 2.2. This result is important because the condition $N/T^2 \to 0$ may be justified, even in the case of microeconomic data, as long as T is moderately large. By contrast, condition $N/T \to 0$ is quite strong in the analysis of microeconomic data because the number of cross-sectional units N is typically larger than the length of the time series T.

The estimation of the variance can also be justified under the weaker condition $N/T^2 \to 0$. Suppose that the parameter of interest is $var(\gamma_{k,i}) = E(\gamma_{k,i}^2) - (E(\gamma_{k,i}))^2$. Thus, it is estimated by

$$\frac{1}{N}\sum_{i=1}^{N} \hat{\gamma}_{k,i}^2 - \left(\frac{1}{N}\sum_{i=1}^{N} \hat{\gamma}_{k,i}\right)^2.$$

Neither of the estimators of $E(\gamma_{k,i}^2)$ and $E(\gamma_{k,i})$ suffer asymptotic bias when $N/T^2 \to 0$. Because the variance is a continuous function of these two moments, it can also be estimated without asymptotic bias when $N/T^2 \to 0$.

We can also estimate the expected value of a smooth function of a vector of the mean and autocovariances. Suppose that we would like to estimate $H := E(h(\theta_i))$, where $h : \mathbb{R}^l \mapsto \mathbb{R}$ is some known smooth function and θ_i is an l-dimensional vector of μ_i and/or $\gamma_{k,i}$s. Let $\hat{\theta}_i$ be the vector of estimators corresponding to the elements of θ_i. This parameter is estimated by

$$\hat{H} := \frac{1}{N}\sum_{i=1}^{N} h(\hat{\theta}_i). \tag{2.5}$$

For example, if we are interested in estimating $H = E(\mu_i\gamma_{0,i})$, it can be estimated by

$$\hat{H} = \frac{1}{N}\sum_{i=1}^{N} \hat{\mu}_i\hat{\gamma}_{0,i}.$$

The covariance between μ_i and $\gamma_{0,i}$ is thus estimated by

$$\frac{1}{N}\sum_{i=1}^{N}\hat{\mu}_i\hat{\gamma}_{0,i} - \left(\frac{1}{N}\sum_{i=1}^{N}\hat{\mu}_i\right)\left(\frac{1}{N}\sum_{i=1}^{N}\hat{\gamma}_{0,i}\right).$$

The half-panel jackknife (HPJ) proposed by Dhaene and Jochmans (2014) can further reduce the bias in $\hat{G}_\mu$ or $\hat{G}_{\gamma_k}$. The estimator exhibits the bias of order $O(1/T)$ and the HPJ bias correction can delete the bias of this order. It thus allows us to relax the condition on the ratio of N and T. The bias correction is easy to implement. Suppose that T is even.[2] We divide the panel data into two subpanels: $\{\{y_{it}\}_{t=1}^{T/2}\}_{i=1}^{N}$ and $\{\{y_{it}\}_{t=T/2+1}^{T}\}_{i=1}^{N}$. The first subpanel, $\{\{y_{it}\}_{t=1}^{T/2}\}_{i=1}^{N}$, consists of observations from the first half of the overall time period, and the second subpanel, $\{\{y_{it}\}_{t=T/2+1}^{T}\}_{i=1}^{N}$, consists of those from the second half. Let $G = G_\mu$ or G_{γ_k} and $\hat{G}$ be the estimator of G. Let $\hat{G}^{(1)}$ and $\hat{G}^{(2)}$ be the estimators of G computed using $\{\{y_{it}\}_{t=1}^{T/2}\}_{i=1}^{N}$ and $\{\{y_{it}\}_{t=T/2+1}^{T}\}_{i=1}^{N}$, respectively. Let $\bar{G} := (\hat{G}^{(1)} + \hat{G}^{(2)})/2$. The HPJ estimator of G is:

$$\hat{G}^H := \hat{G} - (\bar{G} - \hat{G}) = 2\hat{G} - \bar{G}. \tag{2.6}$$

The HPJ estimates the bias in $\hat{G}$ by $\bar{G} - \hat{G}$, and $\hat{G}^H$ corrects the bias in $\hat{G}$ by subtracting the HPJ bias estimate. The bias-corrected estimator $\hat{G}^H$ does not exhibit the bias of order $O(1/T)$ and is asymptotically unbiased even when N/T^2 does not converge to zero. The jackknife bias correction may also be applied to alleviate the bias $\hat{H}$.

When correcting the bias of the variance or covariance estimator, we recommend that the jackknife bias correction is applied for estimation of each expected value, not the variance or covariance estimator itself. For example, to correct the bias for the estimator of $cov(\mu_i, \gamma_{0,i})$, our recommendation is to correct the biases in the estimators of $E(\mu_i\gamma_{0,i})$ and $E(\gamma_{0,i})$ (note that $E(\mu_i)$ can be estimated without bias) and then combine the bias-corrected estimators.

For statistical inferences on parameter G_μ, G_{γ_k}, or H, we suggest the cross-sectional bootstrap. The cross-sectional bootstrap is used to approximate the distribution of the HPJ estimator (or $\hat{G}_\mu$, $\hat{G}_{\gamma_k}$, or $\hat{H}$ when T is sufficiently large). In the cross-sectional bootstrap, we regard the time series from an individual as the unit of observation and approximate the distribution of statistics by that under the empirical distribution of z_i, where $z_i := (y_{i1}, \ldots, y_{iT})$. The algorithm is as follows:

1. Randomly draw $z_1^*, \ldots, z_N^*$ from $\{z_1, \ldots, z_N\}$ with replacement.
2. Compute the statistics of interest, say S, using $z_1^*, \ldots, z_N^*$.
3. Repeat 1 and 2 B times. Let $S^*(b)$ be the statistics computed with the b-th bootstrap sample.

[2] If T is odd, we define $\bar{G} = (\hat{G}^{(1,1)} + \hat{G}^{(2,1)} + \hat{G}^{(1,2)} + \hat{G}^{(2,2)})/4$ as in Dhaene and Jochmans (2014, p. 9), where $\hat{G}^{(1,1)}$, $\hat{G}^{(2,1)}$, $\hat{G}^{(1,2)}$, and $\hat{G}^{(2,2)}$ are the estimators of G computed using $\{\{y_{it}\}_{t=1}^{\lceil T/2\rceil}\}_{i=1}^{N}$, $\{\{y_{it}\}_{t=\lceil T/2\rceil+1}^{T}\}_{i=1}^{N}$, $\{\{y_{it}\}_{t=1}^{\lfloor T/2\rfloor}\}_{i=1}^{N}$, and $\{\{y_{it}\}_{t=\lfloor T/2\rfloor+1}^{T}\}_{i=1}^{N}$, respectively. Here, $\lceil\cdot\rceil$ and $\lfloor\cdot\rfloor$ are the ceiling and floor functions, respectively. We note that the asymptotic properties of the half-panel jackknife estimator for odd T are the same as those for even T. We thus focus on even T in this paper without loss of generality.

4. Compute the distributional quantities of interest for S based on the empirical distribution of $S^*(b)$.

For example, suppose that we are interested in constructing a 95% confidence interval for parameter $G_\mu = E(g(\mu_i))$. We obtain the bootstrap approximation of the distribution of $S = \hat{G}_\mu^H - G_\mu$. Let $\hat{G}_\mu^{H*}(b)$ be the HPJ estimate of G_μ obtained with the b-th bootstrap sample. We then compute the 2.5% and 97.5% quantiles, denoted as $q^*_{0.025}$ and $q^*_{0.975}$, of the empirical distribution of $S^*(b) = \hat{G}_\mu^{H*}(b) - \hat{G}_\mu^H$. The cross-sectional bootstrap 95% confidence interval for G_μ is

$$[\hat{G}_\mu^H - q^*_{0.975}, \hat{G}_\mu^H - q^*_{0.025}].$$

2.4 Asymptotic analysis for the distribution estimators

This section presents the asymptotic properties of the distribution estimators (2.1) and (2.2). We first show the uniform consistency of the empirical distribution of the estimated mean or autocovariance. We then derive the functional central limit theorem for the empirical distributions. We also show that the functional delta method can be applied in this case. All the asymptotic analyses presented in the following sections are under double asymptotics ($N, T \to \infty$). The asymptotic analyses are based on empirical process techniques (see, e.g., van der Vaart and Wellner, 1996).

The following representation is useful for our theoretical analysis. Let $w_{it} \coloneqq y_{it} - E(y_{it}|i) = y_{it} - \mu_i$. By construction, y_{it} is decomposed as

$$y_{it} = \mu_i + w_{it}.$$

The random variable w_{it} is the unobservable idiosyncratic component that varies over both i and t. Note that, by definition, $E(w_{it}|i) = 0$ for any i and t. Note also that $\gamma_{k,i} = E(w_{it}w_{i,t-k}|i)$.

2.4.1 Uniform consistency

In this section, we show that the empirical distributions of $\hat{\mu}_i$ and $\hat{\gamma}_{k,i}$ are uniformly consistent for the true distributions of μ_i and $\gamma_{k,i}$,

Because we use empirical process techniques, it is convenient to rewrite the empirical distributions as empirical processes indexed by a class of indicator functions. Let $\mathbb{P}_N^{\hat{\mu}}$ be the empirical measure of $\hat{\mu}_i$:

$$\mathbb{P}_N^{\hat{\mu}} := \frac{1}{N}\sum_{i=1}^{N} \delta_{\hat{\mu}_i},$$

where $\delta_{\hat{\mu}_i}$ is the probability distribution degenerated at $\hat{\mu}_i$. Let $\mathcal{F}$ be the following class of indicator functions:

$$\mathcal{F} := \{\mathbf{1}_{(-\infty,a]} : a \in \mathbb{R}\},$$

where $\mathbf{1}_{(-\infty,a]}(x) := \mathbf{1}(x \leq a)$. We define the probability measure of μ_i as P_0^{μ}. In this notation, the empirical distribution function, $\mathbb{F}_N^{\hat{\mu}}$, is an empirical process indexed by $\mathcal{F}$. For example, $\mathbb{P}_N^{\hat{\mu}} f = \mathbb{F}_N^{\hat{\mu}}(a)$ for $f = \mathbf{1}_{(-\infty,a]}$. Similarly, for $f = \mathbf{1}_{(-\infty,a]}$, $P_0^{\mu} f = F_0^{\mu}(a) = \Pr(\mu_i \leq a)$. The empirical measure of $\hat{\gamma}_{k,i}$, $\mathbb{P}_N^{\hat{\gamma}_k}$ and the probability measure of $\gamma_{k,i}$, $P_0^{\gamma_k}$ are analogously defined.

Our objective in this section is to show that the class $\mathcal{F}$ is P_0-Glivenko–Cantelli for $P_0 = P_0^{\mu}$ or $P_0^{\gamma_k}$, in the sense that

$$\sup_{f \in \mathcal{F}} |\mathbb{P}_N f - P_0 f| \xrightarrow{as} 0, \tag{2.7}$$

where $\mathbb{P}_N$ is the empirical distribution corresponding to P_0, and $\xrightarrow{as}$ is the almost sure convergence. This is equivalent to the uniform consistency of the empirical distribution function. Note that (2.7) cannot be directly shown by the usual Glivenko–Cantelli theorem, e.g., Theorem 19.1 in van der Vaart (1998). This is because the true distributions of $\hat{\mu}_i$ and $\hat{\gamma}_{k,i}$ change as T increases. Nonetheless, our proof of (2.7) follows similar steps to those of the usual Glivenko–Cantelli theorem.

We use the following assumption throughout the paper, which summarizes the conditions imposed in Section 2.2.

Assumption 2.1. *The sample space of α_i is some Polish space and y_{it} is a scalar real random variable. $\{(\{y_{it}\}_{t=1}^T, \alpha_i)\}_{i=1}^N$ is i.i.d. across i. $\{y_i\}_{t=1}^T$ is strictly stationary given α_i.*

The following conditions are used to show the consistency of $\mathbb{P}_N^{\hat{\mu}}$.

Assumption 2.2. $\sum_{k=-\infty}^{\infty} E|\gamma_{k,i}| < \infty$.

Assumption 2.3. *The random vector $(\mu_i, \bar{y}_i)$ is continuously distributed and its joint density is bounded.*

Assumption 2.2 indicates that the dynamics of w_{it} is a short memory process. We do not here consider the case in which the process has a long memory property. Assumption 2.3 states that μ_i and $\bar{y}_i$ are continuous random variables. This assumption is restrictive in the sense that it does not allow the case in which the distribution of μ_i is discrete or there is no heterogeneity in the mean (i.e., μ_i is homogeneous so that $\mu_i = \mu$ for some constant μ for any i). It should not be very difficult to relax this assumption, but then we would need to employ a different proof technique.

For the consistency of $\mathbb{P}_N^{\hat{\gamma}_k}$, we need a different set of assumptions.

Assumption 2.4. *For each i, $\{y_{it}\}_{t=1}^{\infty}$ is strictly stationary and α-mixing given α_i with mixing coefficients $\{\alpha(m|i)\}_{m=0}^{\infty}$. There exists a sequence $\{\alpha(m)\}_{m=0}^{\infty}$ such that for any i and m, $\alpha(m|i) \leq \alpha(m)$ and $\sum_{m=0}^{\infty}(m+1)^3\alpha(m)^{\delta/(4+\delta)} < \infty$ for some $\delta > 0$.*

Assumption 2.5. $E|w_{it}|^{4+\delta} < \infty$ *for some* $\delta > 0$.

Assumption 2.6. *The random vector $(\gamma_{k,i}, \hat{\gamma}_{k,i})$ is continuously distributed and its joint density is bounded.*

Assumption 2.4 is a mixing condition and restricts the degree of persistency of y_{it}. Assumption 2.5 requires that w_{it} exhibits some moment higher than 4th order. Assumptions 2.4 and 2.5 are satisfied, for example, when y_{it} follows a heterogeneous stationary panel AR(1) model with Gaussian innovations. Assumption 2.6 is similar to Assumption 2.3 and is restrictive in the sense that $\gamma_{k,i}$ must be continuously distributed.

The following theorem establishes the uniform consistency of our distribution estimators.

Theorem 2.1. *Suppose that Assumptions 2.1, 2.2, and 2.3 hold. When $N,T \to \infty$, the class $\mathcal{F}$ is P_0^μ-Glivenko–Cantelli in the sense that*

$$\sup_{f\in\mathcal{F}} \left| \mathbb{P}_N^{\hat{\mu}} f - P_0^\mu f \right| \xrightarrow{as} 0.$$

Suppose that Assumptions 2.1, 2.4, 2.5, and 2.6 hold. When $N,T \to \infty$, the class $\mathcal{F}$ is $P_0^{\gamma_k}$-Glivenko–Cantelli in the sense that

$$\sup_{f\in\mathcal{F}} \left| \mathbb{P}_N^{\hat{\gamma}_k} f - P_0^{\gamma_k} f \right| \xrightarrow{as} 0.$$

2.4.2 Functional central limit theorem

We present the functional central limit theorems for the empirical distributions of $\hat{\mu}_i$ and $\hat{\gamma}_{k,i}$. Our objective is to derive the asymptotic properties of

$$\sqrt{N}(\mathbb{P}_N^{\hat{\mu}} f - P_0^\mu f), \quad \text{and} \quad \sqrt{N}(\mathbb{P}_N^{\hat{\gamma}_k} f - P_0^{\gamma_k} f),$$

where $f \in \mathcal{F}$. This is equivalent to investigating the limiting distributions of $\sqrt{N}(\mathbb{F}_N^{\hat{\mu}}(a) - F_0^\mu(a))$ and $\sqrt{N}(\mathbb{F}_N^{\hat{\gamma}_k}(a) - F_0^{\gamma_k}(a))$ for every $a \in \mathbb{R}$. This result is interesting in its own right because it provides the asymptotic distributions of the empirical distributions. It is also important because the asymptotic distribution of other quantities of interest can be obtained by the functional delta method based on this result.

The functional central limit theorems for $\mathbb{P}_N^{\hat{\mu}}$ and $\mathbb{P}_N^{\hat{\gamma}_k}$ hold under the same set of assumptions for the uniform consistency. However, it requires a stronger condition on the relative magnitude of N and T. Let $\ell^\infty(\mathcal{F})$ be the collection of all bounded real functions on $\mathcal{F}$.

Theorem 2.2. *Suppose that Assumptions 2.1, 2.2, and 2.3 hold. When $N,T \to \infty$ with $N/T \to 0$, we have*

$$\sqrt{N}(\mathbb{P}_N^{\hat{\mu}} - P_0^\mu) \rightsquigarrow \mathbb{G}_{P_0^\mu} \quad \text{in} \quad \ell^\infty(\mathcal{F}),$$

where $\mathbb{G}_{P_0^\mu}$ is a Gaussian process with zero mean and covariance function

$$E(\mathbb{G}_{P_0^\mu}(f_i)\mathbb{G}_{P_0^\mu}(f_j)) = F_0^\mu(a_i \wedge a_j) - F_0^\mu(a_i)F_0^\mu(a_j),$$

for $f_i = 1_{(-\infty,a_i]}$ and $f_j = 1_{(-\infty,a_j]}$.

Suppose that Assumptions 2.1, 2.4, 2.5, and 2.6 hold. When $N, T \to \infty$ with $N/T \to 0$, we have

$$\sqrt{N}(\mathbb{P}_N^{\hat{\gamma}_k} - P_0^{\gamma_k}) \rightsquigarrow \mathbb{G}_{P_0^{\gamma_k}} \quad \text{in} \quad \ell^\infty(\mathcal{F}),$$

where $\mathbb{G}_{P_0}^{\gamma_k}$ is a Gaussian process with zero mean and covariance function

$$E(\mathbb{G}_{P_0^{\gamma_k}}(f_i)\mathbb{G}_{P_0^{\gamma_k}}(f_j)) = F_0^{\gamma_k}(a_i \wedge a_j) - F_0^{\gamma_k}(a_i)F_0^{\gamma_k}(a_j),$$

for $f_i = \mathbf{1}_{(-\infty, a_i]}$ and $f_j = \mathbf{1}_{(-\infty, a_j]}$.

This theorem shows that the asymptotic laws of the empirical processes are Gaussian. This limiting process is then the same as that for the empirical process constructed using the true μ_i or $\gamma_{k,i}$. However, this result requires that $N/T \to 0$. Put differently, the condition $N/T \to 0$ allows us to ignore the estimation error in $\hat{\mu}_i$ or $\hat{\gamma}_{k,i}$ asymptotically.

Here, we provide a brief summary of the proof and explain the reason why the condition $N/T \to 0$ is required. The same discussion can be applied to both $\mathbb{P}_N^{\hat{\mu}}$ and $\mathbb{P}_N^{\hat{\gamma}_k}$. In the following discussion, we let $\mathbb{P}_N$ be either $\mathbb{P}_N^{\hat{\mu}}$ or $\mathbb{P}_N^{\hat{\gamma}_k}$ and P_0 be the corresponding true distribution.

The key to understanding the mechanism behind the requirement that $N/T \to 0$ is to recognize that $E(\mathbb{P}_N f) \neq P_0 f$. That is, $\mathbb{P}_N f$ is not an unbiased estimator for $P_0 f$. For this reason, the existing results for the empirical process cannot be directly applied to derive the asymptotic distribution. Let P_T be the (true) probability measure of $\hat{\mu}_i$ or $\hat{\gamma}_{k,i}$ so that $P_T f = \Pr(\hat{\mu}_i \leq a)$ or $P_T f = \Pr(\hat{\gamma}_{k,i} \leq a)$, which depends on T. Note that $E(\mathbb{P}_N f) = P_T f$. Let

$$\mathbb{G}_{N,P_T} := \sqrt{N}(\mathbb{P}_N - P_T).$$

We decompose the process in the following way:

$$\sqrt{N}(\mathbb{P}_N f - P_0 f) = \mathbb{G}_{N,P_T} f \tag{2.8}$$

$$+ \sqrt{N}(P_T f - P_0 f). \tag{2.9}$$

We analyze the asymptotic behavior of the terms in (2.8) and (2.9) separately.

For $\mathbb{G}_{N,P_T}$ in (2.8), we can directly apply the uniform central limit theorem for the empirical process based on triangular arrays (van der Vaart and Wellner, 1996, Lemma 2.8.7). Note that $E(\mathbb{G}_{N,P_T} f) = 0$. Using Lemma 2.8.7 in van der Vaart and Wellner (1996), we show that

$$\mathbb{G}_{N,P_T} \rightsquigarrow \mathbb{G}_{P_0} \quad \text{in} \quad \ell^\infty(\mathcal{F}).$$

This part of the proof is standard.

The condition $N/T \to 0$ is needed to eliminate the effect of the bias term in the empirical process: $\sqrt{N}(P_T - P_0)$ in (2.9). In the proof of the theorem, we show that

$$\sup_{f \in \mathcal{F}} \left| \sqrt{N}(P_T f - P_0 f) \right| = O\left(\frac{\sqrt{N}}{\sqrt{T}} \right).$$

This term converges to zero when T is of a higher order than N. This result arises from the fact that the rate of convergence of $\hat{\mu}_i$ to μ_i or $\hat{\gamma}_{k,i}$ to $\gamma_{k,i}$ is $1/\sqrt{T}$. Hence, the difference between the distributions of $\hat{\mu}_i$ and μ_i or $\hat{\gamma}_{k,i}$ and $\gamma_{k,i}$ is of order $1/\sqrt{T}$. This is the reason why the difference between P_T and P_0 is also of order $O(1/\sqrt{T})$.

2.4.3 Functional delta method

The asymptotic distribution of an estimator that is a function of the empirical distribution may be derived using the functional delta method. Suppose that we are interested in the asymptotics of $\phi(\mathbb{P}_N)$ for $\phi : D(\mathcal{F}) \mapsto \mathbb{R}$, where $\mathbb{P}_N = \mathbb{P}_N^{\hat{\mu}}$ or $\mathbb{P}_N^{\hat{\gamma}_k}$ and $D(\mathcal{F})$ is the collection of all cadlag real functions on $\mathcal{F}$. For example, the quantile function of $\gamma_{k,i}$, $\phi(P_0^{\gamma_k}) = q_\tau = \inf\{t : F_0^{\gamma_k}(t) \geq \tau\}$ for $\tau \in (0,1)$, may be estimated by:

$$\phi(\mathbb{P}_N^{\hat{\gamma}_k}) = \hat{q}_\tau = (\mathbb{F}_N^{\hat{\gamma}_k})^{-1}(\tau) = \inf\{t : \mathbb{F}_N^{\hat{\gamma}_k}(t) \geq \tau\}.$$

The derivation of the asymptotic distribution of $\phi(\mathbb{P}_N)$ is an application of the functional delta method (see, e.g., van der Vaart and Wellner, 1996, Theorem 3.9.4) and Theorem 2.2. We summarize this result in the following corollary.

Corollary 2.1. *Let $\mathbb{E}$ be a normed linear space. Let $\phi : D(\mathcal{F}) \subset \ell^\infty(\mathcal{F}) \mapsto \mathbb{E}$ be Hadamard differentiable at P_0^μ. Denote its derivative by $\phi'_{P_0^\mu}$. Under Assumptions 2.1, 2.2, and 2.3, when $N, T \to \infty$ with $N/T \to 0$, we have*

$$\sqrt{N}(\phi(\mathbb{P}_N^{\hat{\mu}}) - \phi(P_0^\mu)) \rightsquigarrow \phi'_{P_0^\mu}(\mathbb{G}_{P_0^\mu}).$$

Similarly, suppose that ϕ has the Hadamard derivative, $\phi'_{P_0^{\gamma_k}}$, at $P_0^{\gamma_k}$. Under Assumptions 2.1, 2.4, 2.5, and 2.6, when $N, T \to \infty$ and $N/T \to 0$, we have

$$\sqrt{N}(\phi(\mathbb{P}_N^{\hat{\gamma}_k}) - \phi(P_0^{\gamma_k})) \rightsquigarrow \phi'_{P_0^{\gamma_k}}(\mathbb{G}_{P_0^{\gamma_k}}).$$

Proof. This is immediate by the functional delta method and Theorem 2.2. □

This result can be used, for example, to derive the asymptotic distribution of $\hat{q}_\tau$. The form of $\phi'_{P_0^{\gamma_k}}$ for $\hat{q}_\tau$ is available in Example 20.5 in van der Vaart (1998) and indicates that as $N, T \to \infty$ with $N/T \to 0$,

$$\sqrt{N}(\hat{q}_\tau - q_\tau) \rightsquigarrow N\left(0, \frac{\tau(1-\tau)}{(f_0^{\gamma_k}(q_\tau))^2}\right),$$

where $f_0^{\gamma_k}$ is the density function of $\gamma_{k,i}$.

2.5 Expected value of a smooth function of the heterogeneous mean and/or autocovariances

In this section, we consider the estimation of the expected value of a smooth function of the heterogeneous mean and/or autocovariances. A close inspection of the asymptotic expansion of the estimator reveals that a milder condition on the relative magnitude of N and T is sufficient for the asymptotically unbiased estimation in this case. Furthermore, half-panel jackknife bias correction can reduce the asymptotic bias in the estimator and further relax the condition on the ratio of N to T.

2.5.1 Function of the mean

We first analyze the asymptotic property of $\hat{G}_\mu = N^{-1}\sum_{i=1}^N g(\hat{\mu}_i)$ in (2.3). Recall that the parameter of interest is $G_\mu = E(g(\mu_i))$. We consider the case in which $g(\cdot)$ is sufficiently smooth. We derive the asymptotic distribution of $\hat{G}_\mu$ under the condition $N/T^2 \to 0$.

We make the following assumption on $g(\cdot)$.

Assumption 2.7. *The function $g(\cdot)$ is twice differentiable. $E(g(\mu_i)^2) < \infty$ and $E(g'(\mu_i)^4) < \infty$. $\sup_a |g''(a)| < M$ for some $M < \infty$.*

Assumption 2.7 states that the function $g(\cdot)$ is sufficiently smooth. This assumption is satisfied, for example, when the parameter of interest is the mean (i.e., $g(a) = a$) or when it is the p-th order moment (i.e., $g(a) = a^p$). However, this assumption is not satisfied when the distribution function is estimated (i.e., $g(a) = \mathbf{1}(a \leq c)$ for some $c \in \mathbb{R}$) or when a quantile is estimated. The existence of the first derivative is crucial for relaxing the condition $N/T \to 0$. The existence of the second derivative is useful for evaluating the order of the asymptotic bias.

The asymptotic property of $\hat{G}_\mu$ is given in the following theorem.

Theorem 2.3. *Suppose that Assumptions 2.1, 2.4, 2.5, and 2.7 hold. When $N, T \to \infty$, we have*

$$\hat{G}_\mu - G_\mu \xrightarrow{p} 0,$$

and

$$\sqrt{N}\left(\hat{G}_\mu - G_\mu - \frac{1}{2}E\left(\bar{w}_i^2 g''(\tilde{\mu}_i)\right)\right) \rightsquigarrow N(0, var(g(\mu_i))),$$

where $\tilde{\mu}_i$ is between μ_i and $\bar{y}_i$. In addition, when $N/T^2 \to 0$, we have

$$\sqrt{N}\left(\hat{G}_\mu - G_\mu\right) \rightsquigarrow N(0, var(g(\mu_i))).$$

This theorem states that $\hat{G}_\mu$ is consistent for G_μ and that the asymptotic distribution of $\hat{G}_\mu$ is normal and centered at zero when $N/T^2 \to 0$. Note that we use the mixing and moment conditions that have been used for $\mathbb{P}_N^{\hat{\gamma}_k}$ here. The remarkable result is that the asymptotically unbiased estimation holds under $N/T^2 \to 0$, which is a markedly weaker condition than $N/T \to 0$. This result is because of the smoothness of $g(\cdot)$ and the fact that $\hat{\mu}_i$ is unbiased for μ_i. In fact, when we are interested in $E(\mu_i)$ (i.e., when $g(a) = a$), no conditions on the relative magnitude of N and T are needed to achieve an asymptotically unbiased estimation (indeed T can be fixed for the estimation of $E(\mu_i)$). However, if $g(\cdot)$ is nonlinear, $N/T^2 \to 0$ is needed to remove the asymptotic bias.

In order to obtain a better understanding of the results in the theorem, we observe the

following expansion:

$$\begin{aligned}
&\sqrt{N}\left(\hat{G}_\mu - G_\mu\right) \\
&= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(g(\mu_i) - E(g(\mu_i))\right) + \frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\hat{\mu}_i - \mu_i)g'(\mu_i) + \frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\hat{\mu}_i - \mu_i)^2 g''(\tilde{\mu}_i) \\
&= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(g(\mu_i) - E(g(\mu_i))\right) + \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\bar{w}_i g'(\mu_i) + \frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\bar{w}_i)^2 g''(\tilde{\mu}_i). \qquad (2.10)
\end{aligned}$$

The second term in (2.10) has a mean of zero and is of order $O_p(1/\sqrt{T})$. The fact that the second term has a mean of zero is the key reason that a milder condition, $N/T^2 \to 0$, is sufficient for the asymptotically unbiased estimation of G_μ. This result relies on the assumption that $g(\cdot)$ is smooth. When $g(\cdot)$ is not smooth, this expansion cannot be executed and we cannot exploit the fact that $\hat{\mu}_i$ is unbiased for μ_i. The third term corresponds to the bias caused by the nonlinearity of $g(\cdot)$. When $g(\cdot)$ is linear, this term does not appear and the parameter can be estimated without any restriction on the relative magnitude between N and T. The nonlinearity bias is of order $O_p(\sqrt{N}/T)$. The condition $N/T^2 \to 0$ is used to eliminate the effect of this bias.

We also note that the joint asymptotic distribution of the estimators of $E(g_1(\mu_i))$ and $E(g_2(\mu_i))$ for different $g_1(\cdot)$ and $g_2(\cdot)$ can be easily derived. This is because the estimator is asymptotically linear. This observation is important, for example, when we are interested in the variance of μ_i. The variance of μ_i is a function of $E(\mu_i^2)$ and $E(\mu_i)$. To derive the asymptotic distribution of the variance estimator, we need the joint asymptotic distribution of $N^{-1}\sum_{i=1}^{N}\hat{\mu}_i^2$ and $N^{-1}\sum_{i=1}^{N}\hat{\mu}_i$. The fact that the estimator is asymptotically linear enables us to derive it easily.

2.5.2 Function of the autocovariances

We next consider the asymptotic properties of $\hat{G}_{\gamma_k} = N^{-1}\sum_{i=1}^{N} g(\hat{\gamma}_{k,i})$ in (2.4), which is the estimator for $G_{\gamma_k} = E(g(\gamma_{k,i}))$, and obtain results similar to those for $\hat{G}_\mu$. However, $\hat{G}_{\gamma_k}$ suffers from an additional source of bias, namely incidental parameter bias.

We make the following additional assumptions to study the asymptotic properties of $\hat{G}_{\gamma_k}$.

Assumption 2.8. *For each i, $\{y_{it}\}_{t=1}^{\infty}$ is strictly stationary and α-mixing given α_i with mixing coefficients $\{\alpha(m|i)\}_{m=0}^{\infty}$. There exists a sequence $\{\alpha(m)\}_{m=0}^{\infty}$ such that for any i and m, $\alpha(m|i) \leq \alpha(m)$ and $\sum_{m=0}^{\infty}(m+1)^3\alpha(m)^{\delta/(8+\delta)} < \infty$ for some $\delta > 0$.*

Assumption 2.9. *$E|w_{it}|^{8+\delta} < \infty$ for some $\delta > 0$.*

Assumption 2.10. *The function $g(\cdot)$ is twice differentiable. $E(g(\gamma_{k,i})^2) < \infty$, $E((g'(\gamma_{k,i}))^4) < \infty$. $\sup_a |g''(a)| < M$ for some $M < \infty$.*

Assumption 2.8 is a stronger version of Assumption 2.4 and imposes restrictions on the persistency of w_{it}. Assumption 2.9 is a stronger version of Assumption 2.5 and states that w_{it}

has some moment of higher order than 8. Assumption 2.10 is similar to Assumption 2.7 and states that function $g(\cdot)$ is sufficiently smooth.

The asymptotic property of $\hat{G}_{\gamma_k}$ is given in the following theorem.

Theorem 2.4. *Suppose that Assumptions 2.1, 2.8, 2.9, and 2.10 are satisfied. When $N, T \to \infty$, it holds that*

$$\hat{G}_{\gamma_k} - G_{\gamma_k} \xrightarrow{p} 0.$$

Moreover, when additionally $N/T^2 \to 0$ holds, we have

$$\sqrt{N}(\hat{G}_{\gamma_k} - G_{\gamma_k}) \rightsquigarrow N(0, var(g(\gamma_{k,i})).$$

This theorem states that $\hat{G}_{\gamma_k}$ is consistent for G_{γ_k} and that the asymptotic distribution of $\hat{G}_{\gamma_k}$ is normal and is centered at zero when $N/T^2 \to 0$. Similarly to Theorem 2.3, this theorem merely requires that $N/T^2 \to 0$ because of the smoothness of $g(\cdot)$ and the fact that the leading term in the expansion of $\hat{\gamma}_{k,i}$ has a mean of zero. However, contrary to Theorem 2.3, even if our parameter of interest is $E(\gamma_{k,i})$ so that $g(\cdot)$ is linear, we cannot relax the condition $N/T^2 \to 0$. This is because $\hat{\gamma}_{k,i}$ is not unbiased for $\gamma_{k,i}$.

The results of the theorem can be better understood by examining the asymptotic expansion of $\hat{\gamma}_{k,i}$ and $\hat{G}_{\gamma_k}$. The autocovariance estimator, $\hat{\gamma}_{k,i}$, is expanded as follows:

$$\begin{aligned}
\hat{\gamma}_{k,i} &= \frac{1}{T-k}\sum_{t=k+1}^{T}(y_{it} - \bar{y}_i)(y_{i,t-k} - \bar{y}_i) \\
&= \gamma_{k,i} + \frac{1}{T-k}\sum_{t=k+1}^{T}(w_{it}w_{i,t-k} - \gamma_{k,i}) - (\bar{w}_i)^2 + O_p\left(\frac{1}{T^2}\right).
\end{aligned}$$

It is important to observe that the second term in the second line has a mean of zero although it is of order $T^{-1/2}$. The third term, $(\bar{w}_i)^2$, is the estimation error in $\bar{y}_i$ $(= \hat{\mu}_i)$. This term is of order $O(1/T)$ and is the cause of the incidental parameter bias (Neyman and Scott, 1948; Nickell, 1981). By the Taylor expansion of $\hat{G}_{\gamma_k}$, we have

$$\begin{aligned}
\sqrt{N}(\hat{G}_{\gamma_k} - G_{\gamma_k}) &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}(g(\gamma_{k,i}) - E(g(\gamma_{k,i}))) \\
&\quad + \frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\hat{\gamma}_{k,i} - \gamma_{k,i})g'(\gamma_{k,i}) + \frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\hat{\gamma}_{k,i} - \gamma_{k,i})^2 g''(\tilde{\gamma}_{k,i}) \\
&= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}(g(\gamma_{k,i}) - E(g(\gamma_{k,i}))) \qquad (2.11) \\
&\quad + \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\frac{1}{T-k}\sum_{t=k+1}^{T}w_{it}w_{i,t-k} - \gamma_{k,i}\right)g'(\gamma_{k,i}) \qquad (2.12) \\
&\quad - \frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\bar{w}_i)^2 g'(\gamma_{k,i}) + \frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\hat{\gamma}_{k,i} - \gamma_{k,i})^2 g''(\tilde{\gamma}_{k,i}) + O_p\left(\frac{\sqrt{N}}{T^2}\right), \qquad (2.13)
\end{aligned}$$

where the second equality is obtained by plugging the expansion for $\hat{\gamma}_{k,i}$, and $\tilde{\gamma}_{k,i}$ is between $\hat{\gamma}_{k,i}$ and $\gamma_{k,i}$.

Contrary to $\hat{G}_\mu$, $\hat{G}_{\gamma_k}$ exhibits the incidental parameter bias that corresponds to the first term in (2.13). This bias is caused by the estimation of μ_i by $\bar{y}_i$ and is of order $O_p(\sqrt{N}/T)$, but does not appear in the expansion of $\hat{G}_\mu$. Because of this term, the condition $N/T^2 \to 0$ is needed even when $g(\cdot)$ is linear. The other terms are similar to those in the expansion of $\hat{G}_\mu$. The term on the right-hand side of (2.11) yields the asymptotic normality of $\hat{G}$. The term in (2.12) has a mean of zero and is of order $O_p(1/\sqrt{T})$. That this term has a mean of zero is crucial for the condition $N/T^2 \to 0$ to be sufficient for the asymptotic unbiasedness of $\hat{G}_{\gamma_k}$. The second term in (2.13) is the nonlinearity bias term that also appears in $\hat{G}_\mu$. This is also of order $O_p(\sqrt{N}/T)$.

As in the case of $\hat{G}_\mu$, as the estimator is asymptotically linear, it is easy to derive the joint asymptotic distribution of the estimators of, say, $E(g_1(\gamma_{k,i}))$ and $E(g_2(\gamma_{k,i}))$ for different $g_1(\cdot)$ and $g_2(\cdot)$. Similarly, it is also easy to derive the joint asymptotic distribution of $\hat{G}_{\gamma_k}$ and $\hat{G}_\mu$.

2.5.3 Function of a vector of the mean and autocovariances

We now discuss the asymptotic properties of $\hat{H} = N^{-1}\sum_{i=1}^N h(\hat{\theta}_i)$ in (2.5), which is the estimator of $H = E(h(\theta_i))$. Recall that $h : \mathbb{R}^l \mapsto \mathbb{R}$ is some known smooth function, θ_i is an l-dimensional random vector of μ_i and/or $\gamma_{k,i}$s, and $\hat{\theta}_i$ is the estimator of θ_i with $\hat{\mu}_i$ for μ_i and $\hat{\gamma}_{k,i}$ for $\gamma_{k,i}$. The asymptotic results and the mechanism behind them are similar to those of $\hat{G}_{\gamma_k}$, and the asymptotically unbiased estimation is achieved when $N/T^2 \to 0$.

We make the following assumptions to develop the asymptotic properties of $\hat{H}$. They impose conditions on the smoothness of $h(\cdot)$ and the existence of the moments, and are similar to Assumptions 2.7 and 2.10.

Assumption 2.11. *The function $h(\cdot)$ is twice differentiable and $\sup_a \left|\frac{\partial^2}{\partial z_{j_1}\partial z_{j_2}} h(z)|_{z=a}\right| < M$ for some $M < \infty$ and any $j_1, j_2 = 1, \ldots, l$.*

Assumption 2.12. *$E(h^2(\theta_i)) < \infty$ and $E((\frac{\partial}{\partial z_j} h(z)|_{z=\theta_i})^4) < \infty$ for any $j = 1, \ldots, l$.*

The following theorem demonstrates the asymptotic properties of $\hat{H}$.

Theorem 2.5. *Suppose that Assumptions 2.1, 2.8, 2.9, 2.11, and 2.12 hold. When $N, T \to \infty$, it holds that*

$$\hat{H} - H \xrightarrow{p} 0.$$

Moreover, when $N/T^2 \to 0$ holds additionally, we have

$$\sqrt{N}(\hat{H} - H) \rightsquigarrow N(0, var(h(\theta_i))).$$

The theorem states that $\hat{H}$ is consistent when both N and T tend to infinity, and is asymptotically normal with mean zero when $N/T^2 \to 0$. The condition $N/T^2 \to 0$ is needed because

of the incidental parameter bias in $\hat{\gamma}_{k,i}$ and the nonlinearity bias. The proof is very similar to that of Theorem 2.4.

We remark that it is easy to derive the joint asymptotic distribution for the case in which $h(\cdot)$ is multivalued because the estimator is asymptotically linear. Similarly, deriving the joint asymptotic distribution of $\hat{H}$, $\hat{G}_{\gamma_k}$, and $\hat{G}_\mu$ is also possible. For example, when we are interested in the asymptotic distribution of the estimator of $cov(\mu_i, \gamma_{0,i})$, we need to derive the joint asymptotic distribution of $\hat{H} = N^{-1}\sum_{i=1}^N \hat{\mu}_i\hat{\gamma}_{0,i}$, $\hat{G}_\mu = N^{-1}\sum_{i=1}^N \hat{\mu}_i$, and $\hat{G}_{\gamma_k} = N^{-1}\sum_{i=1}^N \hat{\gamma}_{0,i}$. This is possible because of the asymptotic linearity.

2.5.4 Jackknife bias correction

Here, we provide a theoretical justification of the half-panel jackknife (HPJ) bias-corrected estimator (2.6), which is based on the bias-correction method proposed by Dhaene and Jochmans (2014). It results that the bias of order $O(1/T)$ in $\hat{G}_\mu$ and $\hat{G}_{\gamma_k}$ is eliminated by the HPJ procedure. Recall the definitions: $G = G_\mu$ or G_{γ_k} and $\hat{G}$ is the corresponding estimator of G; $\hat{G}^{(1)}$ and $\hat{G}^{(2)}$ are the estimators of G computed using $\{\{y_{it}\}_{t=1}^{T/2}\}_{i=1}^N$ and $\{\{y_{it}\}_{t=T/2+1}^{T}\}_{i=1}^N$, respectively, with even T; the HPJ estimator of G is $\hat{G}^H = 2\hat{G} - \bar{G}$, where $\bar{G} = (\hat{G}^{(1)} + \hat{G}^{(2)})/2$.

We make the following additional assumptions to study the asymptotic property of the HPJ estimator of G_μ.

Assumption 2.13. *The function $g(\cdot)$ is thrice differentiable. $E(g(\mu_i)^2) < \infty$, $E((g'(\mu_i))^4) < \infty$, $E((g''(\mu_i))^4) < \infty$. $\sup_a |g'''(a)| < M$ for some $M < \infty$.*

For the HPJ estimator of G_{γ_k}, we use the following assumptions.

Assumption 2.14. *$E|w_{it}|^{16+\delta} < \infty$ for some $\delta > 0$.*

Assumption 2.15. *The function $g(\cdot)$ is thrice differentiable. $E(g(\gamma_{k,i})^2) < \infty$, $E((g'(\gamma_{k,i}))^4) < \infty$, $E((g''(\gamma_{k,i}))^4) < \infty$. $\sup_a |g'''(a)| < M$ for some $M < \infty$. $\lim_{T\to\infty} T^{-1}(\sum_{t=k+1}^T (w_{it}w_{i,t-k} - \gamma_{k,i}))^2 g''(\gamma_{k,i})$ exists almost surely.*

Assumption 2.14 provides a condition on the existence of moments of w_{it}. It is stronger than Assumption 2.9. A stronger moment condition is called for because the asymptotic expansion needs to be executed for a higher order to derive the asymptotic properties of $\hat{G}^H$. Assumptions 2.13 and 2.15 require that $g(\cdot)$ is thrice differentiable, contrary to Assumptions 2.7 and 2.10. This condition is also needed to conduct a higher-order asymptotic expansion.

The following theorem shows the asymptotic normality of the HPJ estimator.

Theorem 2.6. *Suppose that Assumptions 2.1, 2.8, 2.9, and 2.13 are satisfied. Then, as $N, T \to \infty$ with $N/T^2 \to r$ for some $r \in [0, \infty)$, it holds that*

$$\sqrt{N}(\hat{G}_\mu^H - G_\mu) \rightsquigarrow N(0, var(g(\mu_i))).$$

Suppose that Assumptions 2.1, 2.8, 2.14, and 2.15 are satisfied. Then, as $N, T \to \infty$ with $N/T^2 \to r$ for some $r \in [0, \infty)$, it holds that

$$\sqrt{N}(\hat{G}^H_{\gamma_k} - G_{\gamma_k}) \rightsquigarrow N(0, var(g(\gamma_{k_i}))).$$

The remarkable result is that the HPJ estimator is asymptotically unbiased even when $N/T^2 \to 0$ is violated. Moreover, this bias correction does not inflate the asymptotic variance. To see how the HPJ works, we observe that

$$\hat{G} = G + O_p\left(\frac{1}{\sqrt{N}}\right) + O_p\left(\frac{1}{\sqrt{NT}}\right) + \frac{B}{T} + O_p\left(\frac{1}{T^2}\right),$$

where B is a constant. Similarly, we have

$$\hat{G}^{(j)} = G + O_p\left(\frac{1}{\sqrt{N}}\right) + O_p\left(\frac{1}{\sqrt{NT}}\right) + \frac{2B}{T} + O_p\left(\frac{1}{T^2}\right),$$

for $j = 1, 2$. Therefore, it holds that

$$\hat{G}^H = G + O_p\left(\frac{1}{\sqrt{N}}\right) + O_p\left(\frac{1}{\sqrt{NT}}\right) + O_p\left(\frac{1}{T^2}\right).$$

Thus, the HPJ reduces the order of the bias from $O(1/T)$ to $O(1/T^2)$.

The HPJ bias correction for $\hat{H}$ can also be similarly developed and reduces the bias of order $O(1/T)$. The theoretical justification of the HPJ estimator can be done along the same lines as the proof of Theorem 2.6 under a similar set of assumptions.

Remark 2.3. One may also consider applying the half-series jackknife (Quenouille, 1949, 1956) to each $\hat{\gamma}_{k,i}$, but we argue that this jackknife is not suitable in the current context. In the method based on the half-series jackknife, we bias-correct each $\hat{\gamma}_{k,i}$ using the jackknife. Suppose T is even. Let $\hat{\gamma}^{(1)}_{k,i}$ be the estimator of $\gamma_{k,i}$ using $\{y_{it}\}_{t=1}^{T/2}$, and $\hat{\gamma}^{(2)}_{k,i}$ be that based on $\{y_{it}\}_{t=T/2+1}^{T}$. Let $\bar{\gamma}_{k,i} = (\hat{\gamma}^{(1)}_{k,i} + \hat{\gamma}^{(2)}_{k,i})/2$. The half-series jackknife bias-corrected estimator of $\gamma_{k,i}$ is $\hat{\gamma}^H_{k,i} = \hat{\gamma}_{k,i} - (\bar{\gamma}_{k,i} - \hat{\gamma}_{k,i}) = 2\hat{\gamma}_{k,i} - \bar{\gamma}_{k,i}$. We then estimate G using $\{\hat{\gamma}^H_{k,i}\}_{i=1}^N$. The half-series jackknife method can reduce the bias of order $1/T$ in $\hat{\gamma}_{k,i}$ and, therefore, the incidental parameter bias in $\hat{G}$. However, it does not reduce the bias caused by the nonlinearity of $g(\cdot)$. Indeed, our Monte Carlo simulations show that the half-series jackknife may not work as well as the half-panel jackknife. Therefore, we do not pursue theoretical investigation of the half-series jackknife in this paper. We note that when $g(\cdot)$ is linear, the half-panel jackknife and the half-series jackknife are numerically equivalent.

Remark 2.4. We may also consider a higher-order jackknife bias correction. This is discussed in Dhaene and Jochmans (2014). The HPJ bias correction can eliminate bias up to the order of $O(1/T)$. The higher-order jackknife bias correction is expected to eliminate bias of higher order. Here, we consider the third-order jackknife. Suppose that T is a multiple of six.[3] The panel

[3] See Dhaene and Jochmans (2014) for the treatment of the case in which T is not a multiple of 6. Note that the asymptotic properties of the third-order jackknife estimator do not depend on whether or not T is a multiple of 6.

data are divided into three subpanels: $\{\{y_{it}\}_{i=1}^N\}_{t=1}^{T/3}$, $\{\{y_{it}\}_{i=1}^N\}_{t=T/3+1}^{2T/3}$, and $\{\{y_{it}\}_{i=1}^N\}_{t=2T/3+1}^{T}$. Let $\hat{G}^{(3,1)}$, $\hat{G}^{(3,2)}$, and $\hat{G}^{(3,3)}$ be the estimates of G computed from each of these three subpanels. The third-order jackknife estimator is

$$\hat{G}^{J3} = 3\hat{G} - \frac{3}{2}\left(\hat{G}^{(1)} + \hat{G}^{(2)}\right) + \frac{1}{3}\left(\hat{G}^{(3,1)} + \hat{G}^{(3,2)} + \hat{G}^{(3,3)}\right).$$

However, we do not examine its theoretical property in this paper. Our Monte Carlo results indicate that the higher-order jackknife can eliminate the bias effectively in some cases, but in other cases, we observe that it inflates the bias substantially. This result may be related to the caution noted by Dhaene and Jochmans (2014): a higher-order jackknife may inflate the bias by an order higher than that to be corrected. We also find that the variance inflation may be substantial in certain cases.

2.5.5 Cross-sectional bootstrap

In this section, we present the theorems that justify the use of the cross-sectional bootstrap. The first theorem is concerned with $\hat{G}_\mu$ and $\hat{G}_{\gamma_k}$, and the second theorem discusses the case of $\hat{G}_\mu^H$ and $\hat{G}_{\gamma_k}^H$.

We require the following additional assumptions. The following assumption is used to satisfy Lyapunov's conditions for $\hat{G}_\mu^*$ that is the estimator of G_μ obtained with the bootstrap sample.

Assumption 2.16. *$E(g^6(\mu_i)) < \infty$ and $E((g^2(\mu_i)g'(\mu_i))^4) < \infty$.*

The following assumption is for $\hat{G}_{\gamma_k}^*$

Assumption 2.17. *$E(g^6(\gamma_{k,i})) < \infty$ and $E((g^2(\gamma_{k,i})g'(\gamma_{k,i}))^4) < \infty$.*

The following theorem states that the bootstrap distribution converges to the asymptotic distribution of $\hat{G}_\mu$ or $\hat{G}_{\gamma_k}$ when T is sufficiently large, but it fails to capture the bias term.

Theorem 2.7. *Suppose that Assumptions 2.1, 2.4, 2.5, 2.7, and 2.16 are satisfied. As $N, T \to \infty$, we have*

$$\sup_{x\in\mathbb{R}}\left|\Pr\left(\sqrt{N}(\hat{G}_\mu^* - \hat{G}_\mu) \le x \,\middle|\, \{\{y_{it}\}_{t=1}^T\}_{i=1}^N\right) - \Pr\left(Z_\mu \le x\right)\right| \xrightarrow{p} 0, \tag{2.14}$$

where $Z_\mu \sim N(0, var(g(\mu_i)))$.

Suppose that Assumptions 2.1, 2.8, 2.9, 2.10, and 2.17 are satisfied. As $N, T \to \infty$, we have

$$\sup_{x\in\mathbb{R}}\left|\Pr\left(\sqrt{N}(\hat{G}_{\gamma_k}^* - \hat{G}_{\gamma_k}) \le x \,\middle|\, \{\{y_{it}\}_{t=1}^T\}_{i=1}^N\right) - \Pr\left(Z_{\gamma_k} \le x\right)\right| \xrightarrow{p} 0, \tag{2.15}$$

where $Z_{\gamma_k} \sim N(0, var(g(\gamma_{k,i})))$.

It is important to note that the bootstrap does not capture the bias properties of $\hat{G}_\mu$ and $\hat{G}_{\gamma_k}$. The bootstrap distribution is asymptotically centered at zero. Thus, even if $\hat{G}_\mu$ or $\hat{G}_{\gamma_k}$ suffers from the bias as seen in Section 2.5.2, the bootstrap distribution cannot capture the bias.

This implies that when T is small, we must be cautious about the use of the bootstrap to make statistical inferences. Galvao and Kato (2014), Gonçalves and Kaffo (2014), and Kaffo (2014) also observe that the bootstrap fails to approximate the bias in dynamic panel data settings for different estimators.

We can also show that the bootstrap can approximate the asymptotic distribution of the HPJ estimator.

Theorem 2.8. *Suppose that Assumptions 2.1, 2.8, 2.9, 2.13, and 2.16 are satisfied. As $N,T \to \infty$, we have*

$$\sup_{x\in\mathbb{R}}\left|\Pr\left(\sqrt{N}(\hat{G}_{\mu}^{H*}-\hat{G}_{\mu}^{H})\leq x\,\middle|\,\{\{y_{it}\}_{t=1}^{T}\}_{i=1}^{N}\right)-\Pr\left(Z_{\mu}\leq x\right)\right|\xrightarrow{p}0.$$

Suppose that Assumptions 2.1, 2.8, 2.14, 2.15, and 2.17 are satisfied. As $N,T \to \infty$, we have

$$\sup_{x\in\mathbb{R}}\left|\Pr\left(\sqrt{N}(\hat{G}_{\gamma_k}^{H*}-\hat{G}_{\gamma_k}^{H})\leq x\,\middle|\,\{\{y_{it}\}_{t=1}^{T}\}_{i=1}^{N}\right)-\Pr\left(Z_{\gamma_k}\leq x\right)\right|\xrightarrow{p}0.$$

The proof is analogous to the proof of Theorem 2.7, and is thus omitted.

The theorem indicates that the cross-sectional bootstrap can approximate the asymptotic distribution of the HPJ estimator correctly under the condition that N/T^2 does not diverge. Because the HPJ estimator does not suffer from bias as long as N/T^2 does not diverge, the bootstrap approximation would be more comfortably used for the HPJ estimator.

2.6 Extensions

In this section, we present two extensions based on the proposed procedure. The first is a test for parametric specifications on the distribution of the heterogeneous mean or autocovariance. The second is a test for whether the distributions of the mean or autocovariance are the same across different groups.

2.6.1 Testing parametric specifications

This subsection develops a testing procedure for hypotheses on parametric specifications of the distribution of the heterogeneous mean or autocovariance. The test is based on one-sample Kolmogorov–Smirnov (KS) statistics based on the empirical distributions of $\hat{\mu}_i$ and $\hat{\gamma}_{k,i}$. We derive their asymptotic null distributions. The results indicate that they are equivalent to those of the usual one-sample KS statistics and thus the critical values can be computed easily.

It is not uncommon to impose a parametric specification to model heterogeneous dynamics, and it is important to have a test for such a parametric specification. For example, Browning et al. (2010) develops a parametric model of heterogeneous income dynamics. Hsiao et al. (1999) consider random coefficients panel AR(1) models and impose parametric assumptions to implement a Bayesian procedure. Our test may be used to examine the validity of these parametric specifications.

We consider the following hypotheses:

$$H_0^\mu : P_0^\mu = Q^\mu \quad \text{v.s.} \quad H_1^\mu : P_0^\mu \neq Q^\mu,$$

and

$$H_0^{\gamma_k} : P_0^{\gamma_k} = Q^{\gamma_k} \quad \text{v.s.} \quad H_1^{\gamma_k} : P_0^{\gamma_k} \neq Q^{\gamma_k},$$

where Q^μ and Q^{γ_k} are known continuous probability distributions. The hypotheses are concerned with whether the distributions P_0^μ or $P_0^{\gamma_k}$ are the same as Q^μ or Q^{γ_k}, respectively. We note that Q^μ and Q^{γ_k} cannot be discrete probability distributions. This is because our asymptotic analyses are based on Assumptions 2.3 and 2.6.

We consider tests based on one-sample KS statistics (Kolmogorov, 1933; Smirnov, 1944):

$$KS_1^\mu := \sqrt{N} \left\| \mathbb{P}_N^{\hat{\mu}} - Q^\mu \right\|_\infty = \sqrt{N} \sup_{f \in \mathcal{F}} \left| \mathbb{P}_N^{\hat{\mu}} f - Q^\mu f \right|,$$
$$KS_1^{\gamma_k} := \sqrt{N} \left\| \mathbb{P}_N^{\hat{\gamma}_k} - Q^{\gamma_k} \right\|_\infty = \sqrt{N} \sup_{f \in \mathcal{F}} \left| \mathbb{P}_N^{\hat{\gamma}_k} f - Q^{\gamma_k} f \right|,$$

where $\|\cdot\|_\infty$ is the uniform norm. The test statistics measure the distances between the empirical distributions and the null distributions. We note that KS_1^μ and $KS_1^{\gamma_k}$ are different from the usual one-sample KS statistics in the sense that they are based on the empirical distributions of the estimates $\hat{\mu}_i$ and $\hat{\gamma}_{k,i}$, respectively.

We derive the asymptotic distributions of KS_1^μ and $KS_1^{\gamma_k}$ under H_0^μ and $H_0^{\gamma_k}$, respectively, utilizing Theorem 2.2. The following theorem presents the asymptotic null distributions.

Theorem 2.9. *Suppose that Assumptions 2.1, 2.2, and 2.3 hold for the case of KS_1^μ, and Assumptions 2.1, 2.4, 2.5, and 2.6 hold for the case of $KS_1^{\gamma_k}$. When $N, T \to \infty$ with $N/T \to 0$, it holds that KS_1^μ converges in distribution to $\|\mathbb{G}_{Q^\mu}\|_\infty$ under H_0^μ. Similarly, when $N, T \to \infty$ with $N/T \to 0$, it holds that $KS_1^{\gamma_k}$ converges in distribution to $\|\mathbb{G}_{Q^{\gamma_k}}\|_\infty$ under $H_0^{\gamma_k}$.*

This theorem shows that the asymptotic null distributions of KS_1^μ and $KS_1^{\gamma_k}$ are the uniform norms of the Gaussian processes. The asymptotic null distributions in the theorem are identical to those of the usual one-sample KS statistics developed in Kolmogorov (1933) and Smirnov (1944) so that they are equivalent to those of the one-sample KS statistics based on the true μ_i and $\gamma_{k,i}$. This is because the estimation errors in $\hat{\mu}_i$ or $\hat{\gamma}_{k,i}$ can be ignored asymptotically under the condition $N/T \to 0$.

Note that the asymptotic distributions do not depend on Q^μ or Q^{γ_k}, and critical values can be computed readily. As shown by Kolmogorov (1933) and Smirnov (1944) (for easy reference, see, e.g., Theorem 6.10 in Shao, 2003 or Section 2.1.5 in Serfling, 2002),

$$\Pr(\|\mathbb{G}_{Q^\mu}\|_\infty \leq a) = \Pr(\|\mathbb{G}_{Q^{\gamma_k}}\|_\infty \leq a) = 1 - 2\sum_{j=1}^{\infty} (-1)^{j-1} \exp\left(-2j^2a^2\right), \tag{2.16}$$

for any continuous distributions Q^μ and Q^{γ_k}, with $a > 0$. The far right-hand side of (2.16) does not depend on Q^μ or Q^{γ_k}. Moreover, the critical values are readily available in many statistical software packages and the implementation of our tests is easy.

2.6.2 Testing the difference in degrees of heterogeneity

Next, we develop tests to examine whether the distributions of the heterogeneous mean or autocovariances differ across distinct groups. The test statistics are two-sample KS statistics based on our empirical distribution estimators. We develop the asymptotic null distributions of the test statistics.

In many applications, it would be interesting to see whether distinct groups possess different heterogeneous dynamic structures. For example, when studying income dynamics, one would be interested in whether the distribution of individual average incomes differs between males and females. One may also be interested in whether the degrees of heterogeneity of income dynamics depend on racial group. We develop test procedures for such hypotheses without any parametric specification. Suppose that we have two panel data sets for two different groups: $\{\{y_{it,(1)}\}_{t=1}^{T_1}\}_{i=1}^{N_1}$ and $\{\{y_{it,(2)}\}_{t=1}^{T_2}\}_{i=1}^{N_2}$. We allow the situation in which $T_1 \neq T_2$ and/or $N_1 \neq N_2$. We define $y_{i,(1)} := \{y_{it,(1)}\}_{t=1}^{T_1}$ for $i = 1, \ldots, N_1$ and $y_{i,(2)} := \{y_{it,(2)}\}_{t=1}^{T_2}$ for $i = 1, \ldots, N_2$.

We introduce the following assumption on the data sets.

Assumption 2.18. *Each of* $\{\{y_{it,(1)}\}_{t=1}^{T_1}\}_{i=1}^{N_1}$ *and* $\{\{y_{it,(2)}\}_{t=1}^{T_2}\}_{i=1}^{N_2}$ *satisfies Assumptions 2.1, 2.2, and 2.3 for the case of the mean, and Assumptions 2.1, 2.4, 2.5, and 2.6 for the case of the autocovariances.* $(y_{1,(1)}, \ldots, y_{N_1,(1)})$ *and* $(y_{1,(2)}, \ldots, y_{N_2,(2)})$ *are independent.*

We need the assumptions introduced in the previous sections and require the independence assumption. This assumption implies that our test cannot be used to test the equivalence of the distributions of two variables from the same individuals. Our test is intended to be used to compare the distributions of the same variable from different groups.

We estimate the distribution of the mean or autocovariances for each group. Let $\mu_{i,(a)}$ be the heterogeneous mean of $y_{i,(a)}$ for group $a = 1, 2$. We estimate $\mu_{i,(a)}$ by the sample mean $\hat{\mu}_{i,(a)} := \bar{y}_{i,(a)} := T_a^{-1} \sum_{t=1}^{T_a} y_{it,(a)}$ for $a = 1, 2$. We denote the probability distribution of $\mu_{i,(a)}$ by $P_{0,(a)}^{\mu}$ and the empirical distribution of $\hat{\mu}_{i,(a)}$ by $\mathbb{P}_{N_a,(a)}^{\hat{\mu}}$ for $a = 1, 2$. Similarly, let $\gamma_{k,i,(a)}$ be the k-th individual autocovariance of $y_{i,(a)}$ for $a = 1, 2$. We estimate $\gamma_{k,i,(a)}$ by the sample k-th autocovariance $\hat{\gamma}_{k,i,(a)} := T_a^{-1} \sum_{t=k+1}^{T_a} (y_{it,(a)} - \bar{y}_{i,(a)})(y_{i,t-k,(a)} - \bar{y}_{i,(a)})$ for $a = 1, 2$. We write the distribution of $\gamma_{k,i,(a)}$ by $P_{0,(a)}^{\gamma_k}$ and the empirical distribution of $\hat{\gamma}_{k,i,(a)}$ by $\mathbb{P}_{N_a,(a)}^{\hat{\gamma}_k}$ for $a = 1, 2$.

We focus on the following hypotheses to examine the difference in the degrees of heterogeneity between the two groups:

$$H_0^{\mu} : P_{0,(1)}^{\mu} = P_{0,(2)}^{\mu} \quad \text{v.s.} \quad H_1^{\mu} : P_{0,(1)}^{\mu} \neq P_{0,(2)}^{\mu},$$

and

$$H_0^{\gamma_k} : P_{0,(1)}^{\gamma_k} = P_{0,(2)}^{\gamma_k} \quad \text{v.s.} \quad H_1^{\gamma_k} : P_{0,(1)}^{\gamma_k} \neq P_{0,(2)}^{\gamma_k}.$$

Under the null hypothesis H_0^{μ} ($H_0^{\gamma_k}$), the distribution of the heterogeneous mean (autocovariances) is identical across the two groups.

We investigate the hypotheses using the following two-sample KS statistics based on our empirical distribution estimators:

$$KS_2^{\mu} := \sqrt{\frac{N_1 N_2}{N_1+N_2}} \left\| \mathbb{P}^{\hat{\mu}}_{N_1,(1)} - \mathbb{P}^{\hat{\mu}}_{N_2,(2)} \right\|_{\infty} = \sqrt{\frac{N_1 N_2}{N_1+N_2}} \sup_{f \in \mathcal{F}} \left| \mathbb{P}^{\hat{\mu}}_{N_1,(1)} f - \mathbb{P}^{\hat{\mu}}_{N_2,(2)} f \right|,$$

$$KS_2^{\gamma_k} := \sqrt{\frac{N_1 N_2}{N_1+N_2}} \left\| \mathbb{P}^{\hat{\gamma}_k}_{N_1,(1)} - \mathbb{P}^{\hat{\gamma}_k}_{N_2,(2)} \right\|_{\infty} = \sqrt{\frac{N_1 N_2}{N_1+N_2}} \sup_{f \in \mathcal{F}} \left| \mathbb{P}^{\hat{\gamma}_k}_{N_1,(1)} f - \mathbb{P}^{\hat{\gamma}_k}_{N_2,(2)} f \right|.$$

The test statistics measure the distances between the empirical distributions of the two groups. KS_2^{μ} and $KS_2^{\gamma_k}$ are different from the usual two-sample KS statistics in the sense that KS_2^{μ} and $KS_2^{\gamma_k}$ are based on the empirical distributions of the estimates $\hat{\mu}_{i,(a)}$ and $\hat{\gamma}_{k,i,(a)}$, respectively, for $a = 1, 2$.

The asymptotic null distributions of KS_2^{μ} and $KS_2^{\gamma_k}$ are derived using Theorem 2.2.

Theorem 2.10. *Suppose that Assumption 2.18 is satisfied. When $N_1, T_1 \to \infty$ with $N_1/T_1 \to 0$ and $N_2, T_2 \to \infty$ with $N_2/T_2 \to 0$, and $N_1/(N_1+N_2) \to \lambda$ for some $\lambda \in (0,1)$, it holds that KS_2^{μ} converges in distribution to $\|\mathbb{G}_{P^{\mu}_{0,(1)}}\|_{\infty}$ under H_0^{μ} and that $KS_2^{\gamma_k}$ converges in distribution to $\|\mathbb{G}_{P^{\gamma_k}_{0,(1)}}\|_{\infty}$ under $H_0^{\gamma_k}$.*

This theorem shows that the asymptotic null distributions of KS_2^{μ} and $KS_2^{\gamma_k}$ are the uniform norms of the Gaussian processes. The conditions $N_1/T_1 \to 0$ and $N_2/T_2 \to 0$ are required in order to use the result of Theorem 2.2. The condition $N_1/(N_1+N_2) \to \lambda$ implies that N_1 is not much greater or less than N_2 and guarantees the existence of the asymptotic null distributions.

The asymptotic null distributions in the theorem are the same as those in Theorem 2.9 when we set $Q^{\mu} = P^{\mu}_{0,(1)}$ or $Q^{\gamma_k} = P^{\gamma_k}_{0,(1)}$. Hence, the asymptotic null distributions can be evaluated easily by (2.16) and the critical values of our test are readily available.

Remark 2.5. When the true distributions of $\hat{\mu}_{i,(1)}$ and $\hat{\mu}_{i,(2)}$ (or $\hat{\gamma}_{k,i,(1)}$ and $\hat{\gamma}_{k,i,(2)}$) are the same, i.e., when $P^{\hat{\mu}}_{T_1,(1)} = P^{\hat{\mu}}_{T_2,(2)}$ (or $P^{\hat{\gamma}_k}_{T_1,(1)} = P^{\hat{\gamma}_k}_{T_2,(2)}$), neither the condition $N_1/T_1 \to 0$ nor the condition $N_2/T_2 \to 0$ is needed to establish Theorem 2.10. This is clear from the proof of Theorem 2.10. In particular, when $T_1 = T_2$ and the mean and dynamic structures across the two groups are completely identical under the null hypothesis, we can test the null hypotheses H_0^{μ} or $H_0^{\gamma_k}$ without restricting the relative order of N_a and T_a for $a = 1, 2$. Note that we still need the condition $N_1/(N_1+N_2) \to \lambda \in (0,1)$.

2.7 Monte Carlo simulations

This section presents the results of the Monte Carlo simulations. We investigate the finite-sample performance of the proposed methods in the simulations. We also evaluate the performance of the proposed bias-correction method. The simulations are conducted with R 3.1.1 for Mac OS X 10.9.5. The number of replications in the simulations is 5000.

2.7.1 Designs

The data-generating process is the following random coefficients panel ARMA(1,1) process:

$$y_{it} = \eta_i + \phi_i y_{i,t-1} + \epsilon_{it} + \theta_i \epsilon_{i,t-1},$$

for $i = 1, \ldots, N$ and $t = 1, \ldots, T$, where $\epsilon_{it} \sim i.i.d.N(0,1)$. We consider two specifications of the distribution of the random coefficients $(\eta_i, \phi_i, \theta_i)$. In the first specification (design A), $\eta_i \sim i.i.d.N(0,1)$, $\phi_i \sim i.i.d.U[-0.9, 0.9]$, and $\theta_i = 0$ and η_i, ϕ_i, and θ_i are independent. In the second specification (design B), $\eta_i = \phi_i + \xi_i$ where $\phi_i \sim 0.4 + 0.5\ i.i.d.Beta(5,2)$, $\xi_i \sim i.i.d.N(0, 0.25)$, and $\theta_i \sim i.i.d.U[-0.2, 0.3]$ and ϕ_i, ξ_i, and θ_i are independent. The second specification is motivated by the empirical results of Browning et al. (2010). However, our specification is simpler and the process here is less persistent. Moreover, the mean part is different. We note that the specification for individual-specific means (η_i) does not affect the estimation of the autocovariances because individual-specific means are eliminated when we estimate these autocovariances. We generate the initial observations from the stationary distribution given $(\eta_i, \phi_i, \theta_i)$:

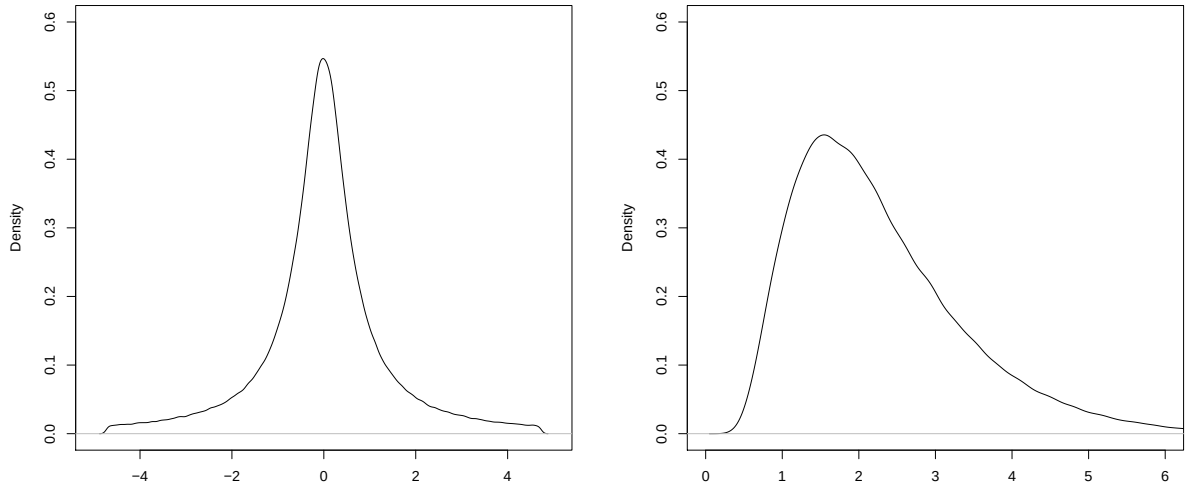
$$\begin{pmatrix} y_{i0} \\ \epsilon_{i0} \end{pmatrix} \sim N\left(\frac{\eta_i}{1-\phi_i}, \begin{pmatrix} \frac{1+\theta_i^2+\phi_i\theta_i}{1-\phi_i^2} & 1 \\ 1 & 1 \end{pmatrix} \right).$$

We set $N = 100$ and 1000, and $T = 24$ and 48.

We estimate the distributions of the mean $(\mu_i = \eta_i/(1-\phi_i))$, the variance $(\gamma_{0,i})$, and the first-order autocovariance $(\gamma_{1,i})$. In particular, we consider the estimation of the mean (Mean), variance (Var), and 25%, 50%, and 75% quantiles (25%Q, 50%Q, 75%Q) of the distributions of these quantities. We also estimate the covariances between these quantities. By way of illustration, the densities of $\gamma_{1,i}$ for designs A and B are plotted in Figure 2.1, parts (a) and (b), respectively.

We consider the following four estimators. The first estimator is a naive estimator based on the empirical distribution of the estimated means or autocovariances. We denote this "NE". The second estimator is the half-panel jackknife estimator (HPJ). Note that for the quantiles, the HPJ estimator is not theoretically justified because they are not expected values of smooth functions. The third estimator is the third-order jackknife estimator (TOJ) discussed in Remark 2.4. The fourth estimator is based on the half-series jackknife autocovariance estimator (HSJ) considered in Remark 2.3.

Importantly, as noted in Section 2.3, the estimation of the variance is done by separately estimating the uncentered second and first moments. Likewise, for the HPJ and TOJ estimators, we do not bias-correct the variance estimator directly. Rather, we separately bias-correct the estimators of the uncentered second and first moments and then compute the variance estimate by combining these bias-corrected estimates. Similarly, when we estimate the covariances with the split-panel bias correction, we bias-correct the cross-moment estimate and the mean



(a) Design A (b) Design B

Figure 2.1: The density function of $\gamma_{1,i}$

estimates separately. We then combine these bias-corrected estimates to form the covariance estimates.

2.7.2 Results

Tables 2.1–2.4 and 2.5–2.8 summarize the results of the Monte Carlo simulations with designs A and B, respectively. They present the bias and the standard deviation (std) of each estimator. In the column labeled "true," the true value of the corresponding quantity is presented. We note that all estimates of the mean of μ_i are numerically equivalent by construction, and that the estimates of the mean of each quantity by HPJ and HSJ are numerically equivalent.

The simulation results demonstrate that our asymptotic analyses are informative regarding the finite-sample behavior of the estimators and the importance of bias correction. When both N and T are small, NE exhibits severe biases for many parameters of interest. In particular, large biases are observed in the estimation of $var(\mu_i)$, all quantities of $\gamma_{0,i}$ and $\gamma_{1,i}$, $cov(\mu_i, \gamma_{0,i})$, and $cov(\mu_i, \gamma_{1,i})$ with design B. With design A, the magnitudes of the biases are relatively moderate. However, the estimation of $var(\mu_i)$, $var(\gamma_{0,i})$, and $cov(\gamma_{0,i}, \gamma_{1,i})$ with design A involves large biases. As T increases while holding N fixed, the biases of NE decrease, which is expected from our asymptotic analyses. Nonetheless, a significant portion of the bias remains even with large T with design B. Worse, the biases are often large compared with the standard deviations. This result suggests the importance of developing the bias-correction method. The standard deviations of NE do not decrease as T becomes large with N fixed. However, the standard deviations decrease as N becomes large. This result can also be expected, as our asymptotic results show that the variances are of order $O(1/N)$.

HPJ successfully reduces the bias in most cases, and works especially well when the biases

μ_i				NE		HPJ		TOJ		HSJ	
	N	T	true	bias	std	bias	std	bias	std	bias	std
Mean	100	24	-0.001	0.002	0.232	0.002	0.232	0.002	0.232	0.002	0.232
	100	48	-0.001	-0.003	0.234	-0.003	0.234	-0.003	0.234	-0.003	0.234
	1000	24	-0.001	0.002	0.074	0.002	0.074	0.002	0.074	0.002	0.074
	1000	48	-0.001	0.001	0.073	0.001	0.073	0.001	0.073	0.001	0.073
Var	100	24	5.280	0.230	2.433	0.056	2.400	0.025	2.397	0.230	2.433
	100	48	5.280	0.116	2.343	-0.017	2.313	-0.033	2.311	0.116	2.343
	1000	24	5.280	0.154	0.750	0.031	0.747	-0.000	0.746	0.154	0.750
	1000	48	5.280	0.087	0.738	0.002	0.735	-0.013	0.734	0.087	0.738
25%Q	100	24	-0.723	-0.009	0.168	0.006	0.191	0.007	0.231	-0.009	0.168
	100	48	-0.723	-0.006	0.170	0.001	0.187	-0.001	0.218	-0.006	0.170
	1000	24	-0.723	-0.014	0.054	0.002	0.061	0.001	0.076	-0.014	0.054
	1000	48	-0.723	-0.005	0.053	0.003	0.059	0.003	0.070	-0.005	0.053
50%Q	100	24	0.000	0.002	0.128	0.001	0.146	0.000	0.178	0.002	0.128
	100	48	0.000	0.001	0.125	0.002	0.137	0.003	0.160	0.001	0.125
	1000	24	0.000	0.001	0.040	0.001	0.046	0.001	0.057	0.001	0.040
	1000	48	0.000	0.001	0.040	0.001	0.044	0.001	0.053	0.001	0.040
75%Q	100	24	0.722	0.013	0.174	-0.003	0.197	-0.003	0.239	0.013	0.174
	100	48	0.722	0.007	0.171	-0.001	0.187	-0.001	0.216	0.007	0.171
	1000	24	0.722	0.016	0.054	0.001	0.062	0.001	0.076	0.016	0.054
	1000	48	0.722	0.008	0.054	0.000	0.061	0.001	0.072	0.008	0.054

Table 2.1: Monte Carlo simulation results: distribution of μ_i with design A

$\gamma_{0,i}$				NE		HPJ		TOJ		HSJ	
	N	T	true	bias	std	bias	std	bias	std	bias	std
Mean	100	24	1.636	-0.180	0.116	-0.062	0.131	-0.030	0.149	-0.062	0.131
	100	48	1.636	-0.102	0.102	-0.022	0.116	-0.007	0.130	-0.022	0.116
	1000	24	1.636	-0.179	0.036	-0.061	0.041	-0.030	0.047	-0.061	0.041
	1000	48	1.636	-0.100	0.033	-0.021	0.037	-0.006	0.042	-0.021	0.037
Var	100	24	0.775	0.516	0.972	0.075	0.793	0.001	0.944	0.882	1.099
	100	48	0.775	0.289	0.570	0.039	0.542	0.024	0.731	0.598	0.811
	1000	24	0.775	0.517	0.306	0.096	0.261	0.033	0.314	0.893	0.355
	1000	48	0.775	0.288	0.182	0.047	0.175	0.029	0.237	0.591	0.256
25%Q	100	24	1.053	-0.170	0.047	-0.002	0.075	1.770	0.155	-0.129	0.050
	100	48	1.053	-0.070	0.040	0.029	0.060	1.936	0.128	-0.048	0.041
	1000	24	1.053	-0.173	0.015	-0.004	0.024	1.762	0.049	-0.133	0.016
	1000	48	1.053	-0.073	0.012	0.027	0.019	1.932	0.040	-0.050	0.013
50%Q	100	24	1.254	-0.094	0.063	0.034	0.097	1.562	0.166	-0.032	0.069
	100	48	1.254	-0.030	0.058	0.034	0.085	1.460	0.145	0.004	0.062
	1000	24	1.254	-0.096	0.020	0.032	0.030	1.557	0.052	-0.034	0.021
	1000	48	1.254	-0.032	0.018	0.032	0.026	1.457	0.047	0.002	0.019
75%Q	100	24	1.838	-0.224	0.128	-0.126	0.194	1.009	0.313	-0.104	0.148
	100	48	1.838	-0.153	0.137	-0.077	0.201	0.928	0.323	-0.080	0.151
	1000	24	1.838	-0.226	0.041	-0.132	0.062	1.002	0.100	-0.109	0.047
	1000	48	1.838	-0.151	0.044	-0.075	0.065	0.932	0.104	-0.079	0.048

Table 2.2: Monte Carlo simulation results: distribution of $\gamma_{0,i}$ with design A

$\gamma_{1,i}$				NE		HPJ		TOJ		HSJ	
	N	T	true	bias	std	bias	std	bias	std	bias	std
Mean	100	24	-0.001	-0.183	0.138	-0.062	0.158	-0.029	0.177	-0.062	0.158
	100	48	-0.001	-0.098	0.134	-0.016	0.150	-0.001	0.164	-0.016	0.150
	1000	24	-0.001	-0.181	0.044	-0.060	0.050	-0.027	0.056	-0.060	0.050
	1000	48	-0.001	-0.100	0.043	-0.019	0.048	-0.003	0.052	-0.019	0.048
Var	100	24	1.816	0.114	1.130	-0.094	1.007	-0.081	1.248	0.726	1.347
	100	48	1.816	0.061	0.726	-0.024	0.732	0.001	0.956	0.539	1.018
	1000	24	1.816	0.115	0.354	-0.065	0.326	-0.043	0.410	0.737	0.430
	1000	48	1.816	0.061	0.233	-0.006	0.239	0.016	0.314	0.534	0.325
25%Q	100	24	-0.565	0.053	0.132	0.033	0.178	-0.797	0.382	0.071	0.133
	100	48	-0.565	0.035	0.137	0.014	0.173	-0.985	0.399	0.044	0.137
	1000	24	-0.565	0.054	0.041	0.031	0.055	-0.794	0.118	0.071	0.041
	1000	48	-0.565	0.031	0.043	0.009	0.055	-0.992	0.124	0.041	0.043
50%Q	100	24	0.000	-0.041	0.069	-0.005	0.099	-0.787	0.214	-0.009	0.074
	100	48	0.000	-0.020	0.080	0.000	0.103	-1.018	0.248	-0.002	0.082
	1000	24	0.000	-0.040	0.021	-0.005	0.031	-0.783	0.067	-0.009	0.023
	1000	48	0.000	-0.020	0.025	-0.001	0.033	-1.016	0.079	-0.003	0.026
75%Q	100	24	0.564	-0.209	0.096	-0.070	0.149	-1.493	0.287	-0.120	0.117
	100	48	0.564	-0.111	0.116	-0.015	0.161	-1.494	0.322	-0.056	0.130
	1000	24	0.564	-0.207	0.030	-0.068	0.047	-1.488	0.090	-0.118	0.037
	1000	48	0.564	-0.110	0.037	-0.013	0.053	-1.491	0.103	-0.056	0.042

Table 2.3: Monte Carlo simulation results: distribution of $\gamma_{1,i}$ with design A

Cov				NE		HPJ		TOJ		HSJ	
	N	T	true	bias	std	bias	std	bias	std	bias	std
μ_i&$\gamma_{0,i}$	100	24	-0.001	0.003	0.311	0.003	0.505	0.002	0.685	0.003	0.508
	100	48	-0.001	0.004	0.377	0.005	0.540	0.007	0.686	0.005	0.545
	1000	24	-0.001	0.004	0.101	0.005	0.164	0.006	0.222	0.005	0.164
	1000	48	-0.001	0.000	0.118	-0.000	0.172	-0.000	0.218	0.000	0.172
μ_i&$\gamma_{1,i}$	100	24	-0.002	0.003	0.445	0.003	0.638	0.003	0.806	0.003	0.643
	100	48	-0.002	-0.000	0.540	0.002	0.697	0.004	0.828	0.001	0.704
	1000	24	-0.002	0.005	0.141	0.007	0.205	0.008	0.260	0.007	0.205
	1000	48	-0.002	0.000	0.169	-0.000	0.221	-0.000	0.263	0.000	0.222
$\gamma_{0,i}$&$\gamma_{1,i}$	100	24	-0.001	-0.814	1.051	-0.277	0.894	-0.104	1.072	-0.350	1.227
	100	48	-0.001	-0.416	0.652	-0.043	0.641	0.033	0.842	-0.029	0.925
	1000	24	-0.001	-0.802	0.331	-0.263	0.295	-0.082	0.361	-0.327	0.397
	1000	48	-0.001	-0.418	0.204	-0.047	0.200	0.030	0.266	-0.037	0.284

Table 2.4: Monte Carlo simulation results: covariances with design A

μ_i				NE		HPJ		TOJ		HSJ	
	N	T	true	bias	std	bias	std	bias	std	bias	std
Mean	100	24	3.592	0.006	0.300	0.006	0.300	0.006	0.300	0.006	0.300
	100	48	3.592	0.007	0.295	0.007	0.295	0.007	0.295	0.007	0.295
	1000	24	3.592	0.014	0.095	0.014	0.095	0.014	0.095	0.014	0.095
	1000	48	3.592	0.006	0.094	0.006	0.094	0.006	0.094	0.006	0.094
Var	100	24	8.275	0.948	1.709	0.337	1.685	0.139	1.687	0.948	1.709
	100	48	8.275	0.544	1.657	0.075	1.633	-0.016	1.637	0.544	1.657
	1000	24	8.275	0.965	0.542	0.438	0.540	0.240	0.540	0.965	0.542
	1000	48	8.275	0.516	0.527	0.125	0.524	0.034	0.525	0.516	0.527
25%Q	100	24	1.620	-0.064	0.302	-0.013	0.361	0.005	0.464	-0.064	0.302
	100	48	1.620	-0.027	0.287	0.010	0.335	0.018	0.421	-0.027	0.287
	1000	24	1.620	-0.083	0.096	-0.031	0.117	-0.011	0.153	-0.083	0.096
	1000	48	1.620	-0.045	0.093	-0.007	0.111	0.001	0.142	-0.045	0.093
50%Q	100	24	3.165	0.007	0.333	0.003	0.391	0.008	0.490	0.007	0.333
	100	48	3.165	0.010	0.325	0.009	0.375	0.012	0.462	0.010	0.325
	1000	24	3.165	0.010	0.106	0.003	0.126	0.004	0.159	0.010	0.106
	1000	48	3.165	0.005	0.105	0.004	0.123	0.005	0.153	0.005	0.105
75%Q	100	24	5.130	0.077	0.467	0.018	0.547	-0.003	0.679	0.077	0.467
	100	48	5.130	0.037	0.461	-0.003	0.525	-0.014	0.636	0.037	0.461
	1000	24	5.130	0.106	0.149	0.047	0.176	0.028	0.222	0.106	0.149
	1000	48	5.130	0.053	0.145	0.013	0.168	0.006	0.208	0.053	0.145

Table 2.5: Monte Carlo simulation results: distribution of μ_i with design B

$\gamma_{0,i}$				NE		HPJ		TOJ		HSJ	
	N	T	true	bias	std	bias	std	bias	std	bias	std
Mean	100	24	2.843	-0.852	0.130	-0.334	0.204	-0.135	0.278	-0.334	0.204
	100	48	2.843	-0.480	0.134	-0.100	0.186	-0.009	0.240	-0.100	0.186
	1000	24	2.843	-0.852	0.042	-0.335	0.065	-0.137	0.088	-0.335	0.065
	1000	48	2.843	-0.480	0.042	-0.098	0.058	-0.007	0.075	-0.098	0.058
Var	100	24	1.179	0.549	0.675	0.449	1.076	0.464	1.496	3.042	2.006
	100	48	1.179	0.602	0.643	0.368	0.938	0.281	1.326	2.270	1.619
	1000	24	1.179	0.549	0.214	0.486	0.346	0.534	0.488	3.034	0.642
	1000	48	1.179	0.607	0.204	0.410	0.309	0.353	0.442	2.287	0.521
25%Q	100	24	2.033	-0.885	0.081	-0.493	0.143	0.898	0.216	-0.752	0.101
	100	48	2.033	-0.543	0.088	-0.200	0.147	1.101	0.213	-0.430	0.102
	1000	24	2.033	-0.893	0.026	-0.503	0.046	0.880	0.070	-0.764	0.032
	1000	48	2.033	-0.552	0.028	-0.210	0.047	1.083	0.067	-0.440	0.032
50%Q	100	24	2.608	-0.963	0.115	-0.487	0.199	0.798	0.290	-0.693	0.152
	100	48	2.608	-0.585	0.122	-0.201	0.197	0.954	0.281	-0.371	0.148
	1000	24	2.608	-0.968	0.036	-0.493	0.062	0.788	0.092	-0.699	0.048
	1000	48	2.608	-0.589	0.039	-0.206	0.064	0.946	0.092	-0.376	0.047
75%Q	100	24	3.411	-0.985	0.198	-0.391	0.342	0.937	0.501	-0.396	0.290
	100	48	3.411	-0.570	0.204	-0.144	0.331	0.976	0.473	-0.141	0.262
	1000	24	3.411	-0.983	0.063	-0.390	0.110	0.946	0.159	-0.396	0.093
	1000	48	3.411	-0.567	0.065	-0.141	0.105	0.988	0.150	-0.138	0.084

Table 2.6: Monte Carlo simulation results: distribution of $\gamma_{0,i}$ with design B

$\gamma_{1,i}$				NE		HPJ		TOJ		HSJ	
	N	T	true	bias	std	bias	std	bias	std	bias	std
Mean	100	24	2.266	-0.879	0.125	-0.333	0.203	-0.124	0.280	-0.333	0.203
	100	48	2.266	-0.491	0.133	-0.095	0.188	-0.003	0.242	-0.095	0.188
	1000	24	2.266	-0.879	0.040	-0.335	0.065	-0.126	0.089	-0.335	0.065
	1000	48	2.266	-0.490	0.041	-0.094	0.059	-0.003	0.076	-0.094	0.059
Var	100	24	1.229	0.369	0.640	0.353	1.031	0.425	1.433	2.965	2.001
	100	48	1.229	0.515	0.633	0.345	0.937	0.282	1.326	2.265	1.646
	1000	24	1.229	0.368	0.204	0.388	0.334	0.494	0.470	2.954	0.641
	1000	48	1.229	0.520	0.201	0.388	0.310	0.356	0.444	2.282	0.531
25%Q	100	24	1.444	-0.864	0.074	-0.504	0.133	-0.216	0.193	-0.735	0.094
	100	48	1.444	-0.533	0.087	-0.200	0.146	-0.046	0.207	-0.419	0.102
	1000	24	1.444	-0.872	0.023	-0.516	0.041	-0.236	0.060	-0.746	0.029
	1000	48	1.444	-0.542	0.027	-0.210	0.046	-0.064	0.066	-0.430	0.032
50%Q	100	24	2.040	-1.000	0.109	-0.509	0.191	-0.449	0.272	-0.719	0.150
	100	48	2.040	-0.603	0.119	-0.203	0.195	-0.239	0.276	-0.383	0.146
	1000	24	2.040	-1.004	0.034	-0.514	0.060	-0.458	0.086	-0.724	0.048
	1000	48	2.040	-0.607	0.039	-0.207	0.064	-0.246	0.090	-0.388	0.047
75%Q	100	24	2.857	-1.068	0.191	-0.404	0.337	-0.422	0.487	-0.440	0.290
	100	48	2.857	-0.611	0.201	-0.144	0.328	-0.258	0.464	-0.162	0.267
	1000	24	2.857	-1.067	0.061	-0.403	0.107	-0.414	0.152	-0.437	0.093
	1000	48	2.857	-0.608	0.064	-0.142	0.105	-0.245	0.149	-0.159	0.085

Table 2.7: Monte Carlo simulation results: distribution of $\gamma_{1,i}$ with design B

Cov				NE		HPJ		TOJ		HSJ	
	N	T	true	bias	std	bias	std	bias	std	bias	std
μ_i&$\gamma_{0,i}$	100	24	1.331	-0.731	0.489	-0.403	0.787	-0.205	1.093	-0.389	0.790
	100	48	1.331	-0.446	0.509	-0.135	0.727	-0.008	0.956	-0.121	0.732
	1000	24	1.331	-0.731	0.156	-0.397	0.251	-0.201	0.348	-0.392	0.251
	1000	48	1.331	-0.440	0.162	-0.118	0.236	0.007	0.310	-0.116	0.235
μ_i&$\gamma_{1,i}$	100	24	1.365	-0.753	0.472	-0.418	0.782	-0.211	1.094	-0.404	0.786
	100	48	1.365	-0.456	0.504	-0.136	0.731	-0.004	0.963	-0.122	0.736
	1000	24	1.365	-0.754	0.149	-0.413	0.249	-0.208	0.347	-0.409	0.248
	1000	48	1.365	-0.451	0.161	-0.119	0.237	0.010	0.313	-0.117	0.237
$\gamma_{0,i}$&$\gamma_{1,i}$	100	24	1.203	0.439	0.656	0.405	1.052	0.448	1.462	2.974	2.000
	100	48	1.203	0.549	0.637	0.358	0.937	0.283	1.324	2.254	1.631
	1000	24	1.203	0.438	0.208	0.441	0.340	0.517	0.478	2.964	0.640
	1000	48	1.203	0.555	0.202	0.401	0.310	0.355	0.442	2.272	0.525

Table 2.8: Monte Carlo simulation results: covariances with design B

of NE are large. In particular, HPJ succeeds markedly in correcting the biases of the estimation of $var(\gamma_{0,i})$ and $cov(\gamma_{0,i}, \gamma_{1,i})$ with design A and those of $var(\mu_i)$ and $E(\gamma_{0,i})$ with design B, even when both N and T are small. Interestingly, HPJ also eliminates the biases in the quantile estimates in many cases, despite our theoretical justification that HPJ does not apply to the estimation of the distribution function or quantiles. This result indicates that HPJ may in fact be useful, even when the parameter of interest is not the expected value of a smooth function. However, we need to develop alternative asymptotic analyses to show this formally. As expected by our asymptotic results, the biases in HPJ tend to decrease as T increases. When both N and T are large, the biases in HPJ are satisfactorily small in many cases.

While HPJ slightly increases the finite-sample standard deviations in some cases, the inflation of the standard deviations would be acceptable. The biases are more serious than the standard deviations in many cases. Although expected, when NE is almost unbiased (e.g., in the estimation of $cov(\mu_i, \gamma_{0,i})$ in design A), HPJ slightly increases the mean squared errors of estimates in some cases. Otherwise, the bias reduction of HPJ sufficiently compensates for the inflation of the standard deviations except for $var(\gamma_{0,i})$ and $var(\gamma_{1,i})$ in design B. When both N and T are large, the standard deviations of NE and HPJ are similar. This is expected given our asymptotic result that NE and HPJ possess the same asymptotic variance. For these reasons, we stress that HPJ is more reliable than NE.

In some cases, TOJ reduces the bias more successfully than HPJ, but in other cases, TOJ inflates the bias substantially. We observe the large bias of TOJ in estimating the quantiles of $\gamma_{0,i}$ and $\gamma_{1,i}$ with design A and $var(\gamma_{0,i})$, the quantiles of $\gamma_{0,i}$, $var(\gamma_{1,i})$, and $cov(\gamma_{0,i}, \gamma_{1,i})$ with design B. In addition, TOJ often increases the standard deviation considerably. The examples include the estimation of $var(\gamma_{0,i})$, $var(\gamma_{1,i})$, and $cov(\gamma_{0,i}, \gamma_{1,i})$ with design A and $var(\gamma_{0,i})$ with design B. This result corresponds with the note in Remark 2.4: the higher-order jackknife may inflate the higher-order bias and the small-sample standard deviation. The inflation of the bias or the standard deviation is critical, especially when the biases of NE and HPJ are not large. We thus recommend HPJ rather than the higher-order jackknife as a precaution.

HSJ does not reduce the bias except in the case of the mean. This is because HSJ fails to eliminate the bias caused by the nonlinearity of smooth functions, as discussed in Remark 2.3. Worse, HSJ substantially increases the biases in some cases. For example, they are observed for the estimation of $var(\gamma_{1,i})$ in design A and for the estimation of $var(\gamma_{0,i})$, $var(\gamma_{1,i})$, and $cov(\gamma_{0,i}, \gamma_{1,i})$ in design B. Because of these Monte Carlo results, we do not pursue a theoretical investigation of HSJ in this paper.

Our preferred procedure is HPJ, given the results of these Monte Carlo experiments. NE is often considerably biased, whereas HPJ can alleviate the bias problem without significant variance inflation. TOJ may be used for the estimation of the mean of a quantity, but in other cases it may inflate both the bias and the variance. The performance of TOJ appears to be highly situation dependent, and we hesitate to recommend its use for the moment. HSJ is not recommended.

2.8 Conclusion

This paper proposes methods to analyze the heterogeneous dynamic structure using panel data. Our proposed methods do not require model specification and are easy to implement. We first compute the sample mean or the sample autocovariances of each individual. We then use these to estimate the parameters of interest, such as the distribution function, the quantile function, and the other moments of the heterogeneous mean and/or autocovariances. We show that the estimator for the distribution function of the heterogeneous mean or autocovariances exhibits a bias of order $O(1/\sqrt{T})$. When the parameter of interest can be written as the expected value of a smooth function of the heterogeneous mean or autocovariances, the bias of the estimator becomes of order $O(1/T)$ and can be reduced by the half-panel jackknife bias-correction method. We also present extensions based on the proposed procedures to test parametric specifications on the distribution of the heterogeneous mean or autocovariances and to test the difference of the heterogeneous dynamic structures across distinct groups. The results of Monte Carlo simulations show that our asymptotic analyses are informative regarding the finite-sample properties of the proposed estimators. Based on the simulation results, we recommend the half-panel jackknife estimator. We believe that our proposed methods can be used to address several important questions regarding the dynamics of economic variables.

Future work: Several future research topics are possible. First, methods for prediction could be considered based on the proposed analysis. Given that our proposed analysis estimates the distributions of the heterogeneous mean and autocovariances we could, in principle, use them to construct a best linear predictor of future values of y_{it}.

Second, while this paper develops the analysis for stationary panel data, it could be used to extend our analysis to cover nonstationary panel data. Two types of nonstationarity are relevant to our analysis. The first is the effect of initial distributions. In this paper, we assume that the initial values are drawn from the stationary distributions for simplicity. As we consider the case in which T is large, the effect of initial values would be negligible in large samples. However, the effect in a finite sample remains untested. The second type of nonstationarity is a stochastic trend. For example, in the literature on income dynamics, there is debate over whether the income process exhibits a unit root (see, e.g., Browning et al., 2010). As autocovariances are not well defined in the presence of a unit root, we require a different procedure to handle unit root cases.

Third, whereas this paper focuses only on balanced panel data, an analysis based on unbalanced panel data would be useful. We believe that, at least in terms of implementation, this extension is not too difficult. This is because there is no difficulty in estimating the mean and autocovariances for each individual, even when the panel is unbalanced, and there is no change in the procedure after obtaining the individual mean and autocovariance estimates. However, theoretical investigation of the properties of the procedure using an unbalanced panel may not

be straightforward.

2.9 Technical appendix

This appendix presents the proofs of the theorems and technical lemmas used to prove the theorems. Section 2.9.1 contains the proofs of the theorems in the main text. The technical lemmas are given in Section 2.9.2.

2.9.1 Proofs of the theorems

This section contains the proofs of the theorems in the main text. We repeatedly cite the results in van der Vaart and Wellner (1996), subsequently abbreviated as VW.

Proof of Theorem 2.1

The proof for $\mathbb{P}_N^{\hat{\mu}}$ and that for $\mathbb{P}_N^{\hat{\gamma}_k}$ are basically the same, so we present that for $\mathbb{P}_N^{\hat{\gamma}_k}$ only. Let $\mathbb{P}_N = \mathbb{P}_N^{\hat{\gamma}_k}$, $P_T = P_T^{\hat{\gamma}_k}$, and $P_0 = P_0^{\gamma_k}$.

By the triangle inequality,

$$\sup_{f\in\mathcal{F}} |\mathbb{P}_N f - P_0 f| \leq \sup_{f\in\mathcal{F}} |\mathbb{P}_N f - P_T f| + \sup_{f\in\mathcal{F}} |P_T f - P_0 f| .$$

For the second term on the right-hand side, Lemma 2.7 (for the case of $\mathbb{P}_N^{\hat{\mu}}$, Lemma 2.6) implies that $\hat{\gamma}_{k,i}$ converges to $\gamma_{k,i}$ in mean square convergence and thus also implies that $\hat{\gamma}_{k,i}$ converges to $\gamma_{k,i}$ in distribution. By Lemma 2.11 in van der Vaart (1998), it holds that

$$\sup_{f\in\mathcal{F}} |P_T f - P_0 f| \to 0,$$

since $\gamma_{k,i}$ is continuously distributed by Assumption 2.6 (for the case of $\mathbb{P}_N^{\hat{\mu}}$, Assumption 2.3).

We then show that the first term converges to 0 almost surely. Note that, for $f = \mathbf{1}_{(-\infty,a]}$, $\mathbb{P}_N f = \mathbb{F}_N^{\hat{\gamma}_{k,i}}(a)$ and $E(\mathbb{F}_N^{\hat{\gamma}_{k,i}}(a)) = \Pr(\hat{\gamma}_{k,i} \leq a) = P_T f$. We first fix a monotone sequence $T = T(N)$ such that $T \to \infty$ as $N \to \infty$, which makes our sample triangular arrays. We use the strong law of large numbers for triangular arrays (see, e.g., Hu, Móricz, and Taylor, 1989, Theorem 2). This is possible because under Assumption 2.1, $\mathbf{1}(\hat{\gamma}_{k,i} \leq a)$ for any a is i.i.d. across individuals, the condition (1.5) in Hu et al. (1989) is clearly satisfied, and the condition (1.6) in Hu et al. (1989) is also satisfied when we set $X = 2$ in the condition (1.6). Thus, we have $\mathbb{F}_N^{\hat{\gamma}_{k,i}}(a) - \Pr(\hat{\gamma}_{k,i} \leq a) \xrightarrow{as} 0$ and $\mathbb{F}_N^{\hat{\gamma}_{k,i}}(a-) - \Pr(\hat{\gamma}_{k,i} < a) \xrightarrow{as} 0$ for every $a \in \mathbb{R}$, as $N, T \to \infty$. Given a fixed $\varepsilon > 0$, there exists a partition $-\infty = a_0 < a_1 < \cdots < a_k = \infty$ such that $\Pr(\gamma_{k,i} < a_i) - \Pr(\gamma_{k,i} \leq a_{i-1}) < \varepsilon/3$ for every i. We have shown that $\sup_{f\in\mathcal{F}} |P_T f - P_0 f| \to 0$, and this implies that for sufficiently large N, T, $\sup_{f\in\mathcal{F}} |P_T f - P_0 f| < \varepsilon/3$. Therefore, we have $\Pr(\hat{\gamma}_{k,i} < a_i) - \Pr(\hat{\gamma}_{k,i} \leq a_{i-1}) < \varepsilon$ for every i. The rest of the proof is the same as the proof of Theorem 19.1 in van der Vaart (1998). For $a_{i-1} \leq a < a_i$,

$$\mathbb{F}_N^{\hat{\gamma}_{k,i}}(a) - \Pr(\hat{\gamma}_{k,i} \leq a) \leq \mathbb{F}_N^{\hat{\gamma}_{k,i}}(a_i-) - \Pr(\hat{\gamma}_{k,i} < a_i) + \varepsilon,$$
$$\mathbb{F}_N^{\hat{\gamma}_{k,i}}(a) - \Pr(\hat{\gamma}_{k,i} \leq a) \geq \mathbb{F}_N^{\hat{\gamma}_{k,i}}(a_{i-1}-) - \Pr(\hat{\gamma}_{k,i} < a_{i-1}) - \varepsilon.$$

Accordingly, we have $\limsup_{N,T\to\infty}(\sup_{f\in\mathcal{F}}|\mathbb{P}_N f - P_T f|) \leq \varepsilon$ almost surely. This is true for every $\varepsilon > 0$, and thus we get

$$\sup_{f\in\mathcal{F}}|\mathbb{P}_N f - P_T f| \xrightarrow{as} 0. \tag{2.17}$$

We note that (2.17) holds for all monotonic diagonal paths $N \to \infty, T(N) \to \infty$. As stated in REMARKS (a) in Phillips and Moon (1999), (2.17) thus holds under doubly asymptotics $N, T \to \infty$. Consequently, we obtain the desired result by the continuous mapping theorem.

□

Proof of Theorem 2.2

The proof for $\mathbb{P}_N^{\hat{\mu}}$ and that for $\mathbb{P}_N^{\hat{\gamma}_k}$ are basically the same, so we present that for $\mathbb{P}_N^{\hat{\gamma}_k}$ only. Let $\mathbb{P}_N = \mathbb{P}_N^{\hat{\gamma}_k}$, $P_T = P_T^{\hat{\gamma}_k}$, and $P_0 = P_0^{\gamma_k}$. The proof is based on the decomposition in (2.8) and (2.9). To study the asymptotic behavior of (2.8), we use Lemma 2.8.7 in VW. We first fix a monotone sequence $T = T(N)$ such that $T \to \infty$ as $N \to \infty$, which makes our sample triangular arrays. By Theorem 2.8.3, Example 2.5.4, and Example 2.3.4 in VW, the class $\mathcal{F}$ is Donsker and pre-Gaussian uniformly in $\{P_T\}$. Thus, we need to check the conditions (2.8.5) and (2.8.6) in VW. The condition (2.8.6) in VW is immediately satisfied by setting the envelope function $F = 1$ (constant).

We check the condition (2.8.5) in VW. Let ρ_{P_T} and ρ_{P_0} be the variance semimetrics with respect to P_T and P_0, respectively. Then,

$$\begin{aligned}
&\sup_{f,g\in\mathcal{F}}|\rho_{P_T}(f,g) - \rho_{P_0}(f,g)| \\
=\ &\sup_{f,g\in\mathcal{F}}|\sqrt{P_T((f-g)-P_T(f-g))^2} - \sqrt{P_0((f-g)-P_0(f-g))^2}| \\
=\ &\sup_{a,a'\in\mathbb{R}}|\sqrt{P_T(\mathbf{1}_{(-\infty,a]}-\mathbf{1}_{(-\infty,a']}-P_T(\mathbf{1}_{(-\infty,a]}-\mathbf{1}_{(-\infty,a']}))^2} \\
&-\sqrt{P_0(\mathbf{1}_{(-\infty,a]}-\mathbf{1}_{(-\infty,a']}-P_0(\mathbf{1}_{(-\infty,a]}-\mathbf{1}_{(-\infty,a']}))^2}| \\
\leq\ &\sup_{a,a'\in\mathbb{R}}|P_T(\mathbf{1}_{(-\infty,a]}-\mathbf{1}_{(-\infty,a']}-P_T(\mathbf{1}_{(-\infty,a]}-\mathbf{1}_{(-\infty,a']}))^2 \\
&-P_0(\mathbf{1}_{(-\infty,a]}-\mathbf{1}_{(-\infty,a']}-P_0(\mathbf{1}_{(-\infty,a]}-\mathbf{1}_{(-\infty,a']}))^2|^{1/2},
\end{aligned}$$

where the first inequality follows from the triangle inequality. Without loss of generality, we assume $a > a'$. Then, by simple algebra,

$$\begin{aligned}
&\sup_{f,g\in\mathcal{F}}|\rho_{P_T}(f,g) - \rho_{P_0}(f,g)| \\
\leq &\sup_{a,a'\in\mathbb{R}}\big|(P_T\mathbf{1}_{(-\infty,a]}-P_0\mathbf{1}_{(-\infty,a]})-(P_T\mathbf{1}_{(-\infty,a]}\mathbf{1}_{(-\infty,a']}-P_0\mathbf{1}_{(-\infty,a]}\mathbf{1}_{(-\infty,a']}) \\
&+(P_T\mathbf{1}_{(-\infty,a']}-P_0\mathbf{1}_{(-\infty,a']})-((P_T\mathbf{1}_{(-\infty,a]})^2-(P_0\mathbf{1}_{(-\infty,a]})^2) \\
&-((P_T\mathbf{1}_{(-\infty,a']})^2-(P_0\mathbf{1}_{(-\infty,a]'})^2)+2(P_T\mathbf{1}_{(-\infty,a]}P_T\mathbf{1}_{(-\infty,a']}-P_T\mathbf{1}_{(-\infty,a]}P_0\mathbf{1}_{(-\infty,a']}) \\
&+2(P_T\mathbf{1}_{(-\infty,a]}P_0\mathbf{1}_{(-\infty,a']}-P_0\mathbf{1}_{(-\infty,a]}P_0\mathbf{1}_{(-\infty,a']})\big|^{1/2}
\end{aligned}$$

$$
\begin{aligned}
&\leq 11 \sup_{a\in\mathbb{R}} \left|P_T \mathbf{1}_{(-\infty,a]} - P_0 \mathbf{1}_{(-\infty,a]}\right|^{1/2} \\
&\to 0,
\end{aligned}
$$

where the last conclusion follows from Lemma 2.11 in van der Vaart (1998), and $\hat{\gamma}_{k,i} \xrightarrow{p} \gamma_{k,i}$, which follows from Lemma 2.7 (for the case of $\hat{\mu}_i$, Lemma 2.6). Therefore, condition (2.8.5) in VW is satisfied.

Therefore, by Lemma 2.8.7 in VW, we have shown that

$$
\mathbb{G}_{N,P_T} \rightsquigarrow \mathbb{G}_{P_0} \quad \text{in} \quad \ell^\infty(\mathcal{F}). \tag{2.18}
$$

We note that (2.18) holds for all monotonic diagonal paths $N \to \infty, T(N) \to \infty$. As stated in REMARKS (a) in Phillips and Moon (1999), (2.18) thus holds under doubly asymptotic $N, T \to \infty$.

Next, we study the asymptotic behavior of (2.9):

$$
\sqrt{N}(P_T f - P_0 f).
$$

Because the nonstochastic function sequence $P_T f - P_0 f$ is uniformly bounded in $f \in \mathcal{F}$, we should consider the convergence rate of

$$
\sup_{f\in\mathcal{F}} \left|P_T f - P_0 f\right|.
$$

Lemmas 2.7 and 2.8 (for the case of $\mathbb{P}_N^{\hat{\mu}}$, Lemmas 2.6 and 2.8) imply that

$$
\sup_{f\in\mathcal{F}} \left|P_T f - P_0 f\right| = O\left(\frac{1}{\sqrt{T}}\right).
$$

Therefore, given $N/T \to 0$, the desired result holds by Slutsky's theorem.

□

Proof of Theorem 2.3

By Taylor's theorem, we have the decomposition:

$$
\begin{aligned}
\sqrt{N}(\hat{G}_\mu - G_\mu) \quad = \quad & \frac{1}{\sqrt{N}} \sum_{i=1}^{N} (g(\mu_i) - E(g(\mu_i))) + \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left((\bar{y}_i - \mu_i) g'(\mu_i)\right) \\
& + \frac{1}{2\sqrt{N}} \sum_{i=1}^{N} \left((\bar{y}_i - \mu_i)^2 g''(\tilde{\mu}_i)\right).
\end{aligned}
$$

The first term on the right-hand side converges in distribution to $N(0, var(g(\mu_i)))$ by Assumptions 2.1 and 2.7.

The second term on the right-hand side is

$$
\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left((\bar{y}_i - \mu_i) g'(\mu_i)\right) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \bar{w}_i g'(\mu_i),
$$

and the expectation is

$$E\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\bar{w}_i g'(\mu_i)\right)=0,$$

by the law of iterated expectations. The variance is

$$var\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\bar{w}_i g'(\mu_i)\right)=E\left(\left(\bar{w}_i g'(\mu_i)\right)^2\right)\leq\sqrt{E(\bar{w}_i^4)}\sqrt{E(g'(\mu_i)^4)}=O(T^{-1}),$$

where the first inequality follows from the Cauchy–Schwarz inequality and the last equality follows from Lemma 2.4 and Assumption 2.7.

For the third term on the right-hand side,

$$E\left(\frac{1}{2\sqrt{N}}\sum_{i=1}^{N}\left((\bar{y}_i-\mu_i)^2 g''(\tilde{\mu}_i)-E(\bar{w}_i^2 g''(\tilde{\mu}_i))\right)\right)=0,$$

and

$$\begin{aligned}&var\left(\frac{1}{2\sqrt{N}}\sum_{i=1}^{N}\left((\bar{y}_i-\mu_i)^2 g''(\tilde{\mu}_i)-E(\bar{w}_i^2 g''(\tilde{\mu}_i))\right)\right)\\ \leq\ & var(\bar{w}_i^2 g''(\tilde{\mu}_i))\leq M\cdot var(\bar{w}_i^2)=O(T^{-2}),\end{aligned}$$

where the first inequality follows from the i.i.d. assumption, the second inequality follows from Assumption 2.7, and the last equality follows from Lemma 2.4.

Therefore, the first claim of the theorem is obtained by Markov's inequality and Slutsky's theorem. We also have the second claim of the theorem because $|E(\bar{w}_i^2 g''(\tilde{\mu}_i))|=O(T^{-1})$, which follows from Lemma 2.1 and Assumption 2.7.

□

Proof of Theorem 2.4

We concentrate on proving the asymptotic normality of $\sqrt{N}(\hat{G}_{\gamma_k}-G_{\gamma_k})$, because it is clear that $\hat{G}_{\gamma_k}$ is consistent for G_{γ_k} by the following proof and the law of large numbers. By Taylor's theorem, we have

$$\begin{aligned}\sqrt{N}(\hat{G}_{\gamma_k}-G_{\gamma_k}) &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}(g(\hat{\gamma}_{k,i})-E(g(\gamma_{k,i})))\\ &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}(g(\gamma_{k,i})-E(g(\gamma_{k,i}))) &(2.19)\\ &\quad+\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\hat{\gamma}_{k,i}-\gamma_{k,i})g'(\gamma_{k,i}) &(2.20)\\ &\quad+\frac{1}{2\sqrt{N}}\sum_{i=1}^{N}(\hat{\gamma}_{k,i}-\gamma_{k,i})^2 g''(\tilde{\gamma}_{k,i}), &(2.21)\end{aligned}$$

where $\tilde{\gamma}_{k,i}$ is located between $\gamma_{k,i}$ and $\hat{\gamma}_{k,i}$. We examine each term in this expansion.

For (2.19), under Assumptions 2.1 and 2.10,

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(g(\gamma_{k,i})-E(g(\gamma_{k,i})))\rightsquigarrow N(0,var(g(\gamma_{k,i}))),$$

by the central limit theorem.

For (2.20), we use the expansion for $\hat{\gamma}_{k,i}$. We have the following:

$$\begin{aligned}
\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\hat{\gamma}_{k,i}-\gamma_{k,i})g'(\gamma_{k,i}) &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\frac{1}{T-k}\sum_{t=k+1}^{T}w_{it}w_{i,t-k}-\gamma_{k,i}\right)g'(\gamma_{k,i}) && (2.22)\\
&\quad -\frac{1}{\sqrt{N}}\frac{T+k}{T-k}\sum_{i=1}^{N}(\bar{w}_i)^2g'(\gamma_{k,i}) && (2.23)\\
&\quad +\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{1}{T-k}\sum_{t=1}^{k}w_{it}\bar{w}_ig'(\gamma_{k,i}) && (2.24)\\
&\quad +\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{1}{T-k}\sum_{t=T-k+1}^{T}w_{it}\bar{w}_ig'(\gamma_{k,i}). && (2.25)
\end{aligned}$$

For (2.22), its expectation is

$$E\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\frac{1}{T-k}\sum_{t=k+1}^{T}w_{it}w_{i,t-k}-\gamma_{k,i}\right)g'(\gamma_{k,i})\right)=0,$$

by the law of iterated expectations. The variance of (2.22) is

$$\begin{aligned}
&var\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\frac{1}{T-k}\sum_{t=k+1}^{T}w_{it}w_{i,t-k}-\gamma_{k,i}\right)g'(\gamma_{k,i})\right)\\
=\;& E\left(\left(\left(\frac{1}{T-k}\sum_{t=k+1}^{T}w_{it}w_{i,t-k}-\gamma_{k,i}\right)g'(\gamma_{k,i})\right)^2\right)\\
\leq\;& \sqrt{E\left(\left(\frac{1}{T-k}\sum_{t=k+1}^{T}w_{it}w_{i,t-k}-\gamma_{k,i}\right)^4\right)}\sqrt{E(g'(\gamma_{k,i})^4)},
\end{aligned}$$

where the first equality follows from the i.i.d. assumption and the first inequality follows from the Cauchy–Schwarz inequality. We have

$$E\left(\left(\frac{1}{T-k}\sum_{t=k+1}^{T}(w_{it}w_{i,t-k}-\gamma_{k,i})\right)^4\right)=O\left(\frac{1}{T^2}\right),$$

by Lemma 2.5. This result and Assumption 2.10 imply that

$$var\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\frac{1}{T-k}\sum_{t=k+1}^{T}w_{it}w_{i,t-k}-\gamma_{k,i}\right)g'(\gamma_{k,i})\right)=O(T^{-1}).$$

Therefore, by Markov's inequality, term (2.22) is of order $O_p(T^{-1/2})$.

We next examine (2.23). We observe that:

$$\begin{aligned} E\left|\frac{1}{\sqrt{N}}\frac{T+k}{T-k}\sum_{i=1}^{N}(\bar{w}_i)^2 g'(\gamma_{k,i})\right| &\leq \sqrt{N}\frac{T+k}{T-k}E\left|(\bar{w}_i)^2 g'(\gamma_{k,i})\right| \\ &\leq \sqrt{N}\frac{T+k}{T-k}\sqrt{E((\bar{w}_i)^4)}\sqrt{E((g'(\gamma_{k,i}))^2)} \\ &= O\left(\frac{\sqrt{N}}{T}\right), \end{aligned}$$

where the second inequality is the Cauchy–Schwarz inequality and the equality follows from Lemma 2.4 and Assumption 2.10. Thus, the term in (2.23) is $O_p(\sqrt{N}/T)$ by Markov's inequality.

For (2.24), we have by the Cauchy–Schwarz inequality that

$$\begin{aligned} E\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{1}{T-k}\sum_{t=1}^{k}w_{it}\bar{w}_i g'(\gamma_{k,i})\right| &\leq \frac{\sqrt{N}}{T-k}E\left|\sum_{t=1}^{k}w_{it}\bar{w}_i g'(\gamma_{k,i})\right| \\ &\leq \frac{\sqrt{N}}{T-k}\sqrt{E\left[\left(\sum_{t=1}^{k}w_{it}\bar{w}_i\right)^2\right]}\sqrt{E[(g'(\gamma_{k,i}))^2]}. \end{aligned}$$

It holds that

$$E\left(\left(\sum_{t=1}^{k}w_{it}\bar{w}_i\right)^2\right) \leq \left(E\left(\left(\sum_{t=1}^{k}w_{it}\right)^4\right)\right)^{1/2}\left(E\left((\bar{w}_i)^4\right)\right)^{1/2},$$

by the Cauchy–Schwarz inequality. Lemma 2.4 implies that $E\left((\bar{w}_i)^4\right) = O(T^{-2})$ under Assumptions 2.8 and 2.9. It is easy to see that $E((\sum_{t=1}^{k}w_{it})^4) = O(k^4)$ by Assumption 2.9. Thus, it follows that

$$E\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{1}{T-k}\sum_{t=1}^{k}w_{it}\bar{w}_i g'(\gamma_{k,i})\right| = O\left(\frac{\sqrt{kN}}{\sqrt{T}(T-k)}\right),$$

and the fourth term (2.24) is of order $O_p\left(\sqrt{kN}/(\sqrt{T}(T-k))\right)$ by Markov's inequality. The same argument can be applied to (2.25), and gives

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{1}{T-k}\sum_{t=T-k+1}^{T}w_{it}\bar{w}_i g'(\gamma_{k,i}) = O_p\left(\frac{\sqrt{kN}}{\sqrt{T}(T-k)}\right).$$

For (2.21),

$$\begin{aligned} E\left|\frac{1}{2\sqrt{N}}\sum_{i=1}^{N}(\hat{\gamma}_{k,i}-\gamma_{k,i})^2 g''(\tilde{\gamma}_{k,i})\right| &\leq \frac{\sqrt{N}}{2}E\left|(\hat{\gamma}_{k,i}-\gamma_{k,i})^2 g''(\tilde{\gamma}_{k,i})\right| \\ &\leq \frac{\sqrt{N}}{2}M\cdot E\left((\hat{\gamma}_{k,i}-\gamma_{k,i})^2\right) \\ &= O\left(\frac{\sqrt{N}}{T}\right), \end{aligned}$$

where the second inequality follows from Assumption 2.10 and the last equality follows from Lemma 2.7. By Markov's inequality, (2.21) is of order $O_p(\sqrt{N}/T)$.

Consequently, we obtain the desired result using Slutsky's theorem.

□

Proof of Theorem 2.5

We show only the asymptotic normality of $\sqrt{N}(\hat{H}-H)$, because the consistency of $\hat{H}$ is clear by the following proof and the law of large numbers. Let $\hat{\theta}_i = (\hat{\theta}_{i,1}, \ldots, \hat{\theta}_{i,l})$ and $\theta_i = (\theta_{i,1}, \ldots, \theta_{i,l})$. By Taylor's theorem, we have

$$\begin{aligned}
&\sqrt{N}(\hat{H} - H) \\
=&\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left(h(\hat{\theta}_i) - E\left(h(\theta_i)\right) \right) \\
=&\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left(h(\theta_i) - E\left(h(\theta_i)\right) \right) \qquad (2.26) \\
&+ \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \sum_{j=1}^{l} (\hat{\theta}_{i,j} - \theta_{i,j}) \frac{\partial}{\partial z_j} h(z) \Big|_{z=\theta_i} \qquad (2.27) \\
&+ \frac{1}{2\sqrt{N}} \sum_{i=1}^{N} \sum_{\sum_{s=1}^{l} j_s = 2} (\hat{\theta}_{i,1} - \theta_{i,1})^{j_1} \cdots (\hat{\theta}_{i,l} - \theta_{i,l})^{j_l} \frac{\partial^2}{\partial z_1^{j_1} \cdots \partial z_l^{j_l}} h(z) \Big|_{z=\tilde{\theta}_i}, \qquad (2.28)
\end{aligned}$$

where $\tilde{\theta}_i$ is located between θ_i and $\hat{\theta}_i$.

For (2.26), under Assumption 2.12, it holds that

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left(h(\theta_i) - E\left(h(\theta_i)\right) \right) \rightsquigarrow N(0, var(h(\theta_i))),$$

by the central limit theorem.

For (2.27), we have that for any $j = 1, \ldots, l$

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} (\hat{\theta}_{i,j} - \theta_{i,j}) \frac{\partial}{\partial z_j} h(z) \Big|_{z=\theta_i} = O_p\left(\frac{\sqrt{N}}{T} \right),$$

which follows from the proof similar to that for Theorems 2.3 and 2.4 under Assumptions 2.1, 2.8, 2.9, 2.11, and 2.12.

For (2.28), we observe that

$$\begin{aligned}
&E \left| \frac{1}{2\sqrt{N}} \sum_{i=1}^{N} \sum_{\sum_{s=1}^{l} j_s = 2} (\hat{\theta}_{i,1} - \theta_{i,1})^{j_1} \cdots (\hat{\theta}_{i,l} - \theta_{i,l})^{j_l} \frac{\partial^2}{\partial z_1^{j_1} \cdots \partial z_l^{j_l}} h(z) \Big|_{z=\tilde{\theta}_i} \right| \\
&\leq \frac{\sqrt{N}}{2} M \sum_{\sum_{s=1}^{l} j_s = 2} E \left| (\hat{\theta}_{i,1} - \theta_{i,1})^{j_1} \cdots (\hat{\theta}_{i,l} - \theta_{i,l})^{j_l} \right|,
\end{aligned}$$

by Assumption 2.11 and the triangular inequality. Note that for any $k_1, k_2 = 1, \ldots, l$,

$$\begin{aligned}
E|(\hat{\theta}_{i,k_1} - \theta_{i,k_1})(\hat{\theta}_{i,k_2} - \theta_{i,k_2})| &\leq \sqrt{E\left((\hat{\theta}_{i,k_1} - \theta_{i,k_1})^2 \right)} \sqrt{E\left((\hat{\theta}_{i,k_2} - \theta_{i,k_2})^2 \right)} \\
&= O(T^{-1}),
\end{aligned}$$

where the inequality follows from the Cauchy–Schwarz inequality and the equality follows from Lemmas 2.6 and 2.7 under Assumptions 2.1, 2.8, and 2.9. Hence, it holds that

$$E\left|\frac{1}{2\sqrt{N}}\sum_{i=1}^{N}\sum_{\sum_{s=1}^{l}j_s=2}(\hat{\theta}_{i,1}-\theta_{i,1})^{j_1}\cdots(\hat{\theta}_{i,l}-\theta_{i,l})^{j_l}\frac{\partial^2}{\partial z_1^{j_1}\cdots\partial z_l^{j_l}}h(z)\Big|_{z=\tilde{\theta}_i}\right|=O\left(\frac{\sqrt{N}}{T}\right).$$

Therefore, (2.28) is $O_p(\sqrt{N}/T)$ by Markov's inequality.

Consequently, we obtain the desired result using Slutsky's theorem.

□

Proof of Theorem 2.6

We first consider $\hat{G}_\mu$. The Taylor expansion gives

$$\begin{aligned}\sqrt{N}(\hat{G}_\mu-G_\mu)=&\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(g(\mu_i)-E(g(\mu_i)))+\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\bar{w}_i g'(\mu_i)\\&+\frac{1}{2\sqrt{N}}\sum_{i=1}^{N}(\bar{w}_i)^2 g''(\mu_i)+\frac{1}{6\sqrt{N}}\sum_{i=1}^{N}(\bar{w}_i)^3 g'''(\tilde{\mu}_i),\end{aligned}$$

where $\tilde{\mu}_i$ is between μ_i and $\bar{y}_i$. As in the proof of Theorem 2.3, we have $\sum_{i=1}^{N}\bar{w}_i g'(\mu_i)/\sqrt{N}$ is $O_p(1/\sqrt{T})$. We show that

$$\frac{1}{2\sqrt{N}}\sum_{i=1}^{N}(\bar{w}_i)^2 g''(\mu_i)=\frac{\sqrt{N}}{T}B+o_p\left(\frac{\sqrt{N}}{T}\right),\tag{2.29}$$

$$\frac{1}{6\sqrt{N}}\sum_{i=1}^{N}(\bar{w}_i)^3 g'''(\tilde{\mu}_i)=o_p\left(\frac{\sqrt{N}}{T}\right),\tag{2.30}$$

for some constant B. When (2.29) and (2.30) hold, the asymptotic normality of the HPJ estimator is established following the argument in Dhaene and Jochmans (2014) when $N/T^2\to r$, where $r\in[0,\infty)$ is some constant. By (2.29), (2.30), and simple algebra, we have

$$\begin{aligned}\sqrt{N}(\hat{G}_\mu-G_\mu)=&\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(g(\mu_i)-E(g(\mu_i)))+\frac{\sqrt{N}}{T}B+o_p\left(\frac{\sqrt{N}}{T}\right),\\\sqrt{N}(\bar{G}_\mu-G_\mu)=&\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(g(\mu_i)-E(g(\mu_i)))-2\frac{\sqrt{N}}{T}B+o_p\left(\frac{\sqrt{N}}{T}\right).\end{aligned}$$

That is, the bias of order $1/T$ of $\bar{G}_\mu$ is twice as large as that of $\hat{G}_\mu$. By the central limit theorem, we have

$$\sqrt{N}\begin{pmatrix}\hat{G}_\mu-G_\mu-T^{-1}B\\ \bar{G}_\mu-G_\mu-2T^{-1}B_1\end{pmatrix}\rightsquigarrow N\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}var(g(\mu_i))&var(g(\mu_i))\\var(g(\mu_i))&var(g(\mu_i))\end{pmatrix}\right),$$

as $N,T\to\infty$ and $N/T^2\to r$. Consequently, when $N,T\to\infty$ and $N/T^2\to r$, we have

$$\sqrt{N}(\hat{G}_\mu^H-G_\mu)\rightsquigarrow N(0,var(g(\mu_i))),$$

by the continuous mapping theorem.

We first prove (2.29). We note that

$$E\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left((\bar{w}_i)^2 g''(\mu_i)\right)\right) = \frac{\sqrt{N}}{T}E(V_{T,i}g''(\mu_i)),$$

where $V_{T,i} := TE((\bar{w}_i)^2|i) = \sum_{j=-T}^{T}\gamma_{j,i}(T-|j|)/T$. The variance is

$$var\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left((\bar{w}_i)^2 g''(\mu_i)\right)\right) = var((\bar{w}_i)^2 g''(\mu_i)).$$

We have

$$\begin{aligned} var((\bar{w}_i)^2 g''(\mu_i)) &\leq E((\bar{w}_i)^4 (g''(\mu_i))^2) \\ &\leq \sqrt{E((\bar{w}_i)^8)}\sqrt{E((g''(\mu_i))^4)} \\ &\leq \frac{1}{T^2}C\sqrt{E((g''(\mu_i))^4)} = O\left(\frac{1}{T^2}\right), \end{aligned}$$

where the second inequality is the Cauchy–Schwarz inequality, the third inequality follows from Assumptions 2.8 and 2.9 and Lemma 2.4, and the last equality follows from Assumption 2.13. Therefore, (2.29) holds with $B = \lim_{T\to\infty} E(V_{T,i}g''(\mu_i))/2$.

Next, we prove (2.30). We have

$$\begin{aligned} E\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\bar{w}_i)^3 g'''(\tilde{\mu}_i)\right| \leq & ME\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left|(\bar{w}_i)^3\right|\right) \\ \leq & M\sqrt{N}E(|\bar{w}_i|^3) \\ \leq & M\sqrt{N}\left(E\left((\bar{w}_i)^4\right)\right)^{\frac{3}{4}}, \end{aligned}$$

where the first inequality follows from Assumption 2.13 and the triangle inequality, the second inequality follows from the i.i.d. assumption, and the third inequality is the Lyapunov's inequality. Thus, we only need to evaluate the order of $E((\bar{w}_i)^4)$. Lemma 2.4 implies $E((\bar{w}_i)^4) = O(1/T^2)$. (2.30) therefore holds because

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\bar{w}_i)^3 g'''(\tilde{\mu}_i) = O_p\left(\frac{\sqrt{N}}{T\sqrt{T}}\right) = o_p\left(\frac{\sqrt{N}}{T}\right),$$

by the Markov inequality. Thus, the asymptotic normality of $\hat{G}_\mu^H$ is proved.

Next, we consider $\hat{G}_{\gamma_k}$. The Taylor expansion gives

$$\begin{aligned} \sqrt{N}(\hat{G}_{\gamma_k} - G_{\gamma_k}) = & \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(g(\gamma_{k,i}) - E(g(\gamma_{k,i}))\right) \\ & - \frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\bar{w}_i)^2 g'(\gamma_{k,i}) + \frac{1}{2\sqrt{N}}\sum_{i=1}^{N}(\hat{\gamma}_{k,i} - \gamma_{k,i})^2 g''(\gamma_{k,i}) \\ & + \frac{1}{3!\sqrt{N}}\sum_{i=1}^{N}(\hat{\gamma}_{k,i} - \gamma_{k,i})^3 g'''(\tilde{\gamma}_{k,i}) + O_p\left(\frac{\sqrt{N}}{T^2}\right). \end{aligned} \tag{2.31}$$

The second and third terms on the right-hand side of (2.31) are of order $O_p(\sqrt{N}/T)$. To establish the asymptotic normality of the HPJ estimator, we focus on showing the following results:

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\bar{w}_i)^2 g'(\gamma_{k,i}) = \frac{\sqrt{N}}{T}B_1 + o_p\left(\frac{\sqrt{N}}{T}\right), \tag{2.32}$$

$$\frac{1}{2\sqrt{N}}\sum_{i=1}^{N}(\hat{\gamma}_{k,i}-\gamma_{k,i})^2 g''(\gamma_{k,i}) = \frac{\sqrt{N}}{T}B_2 + o_p\left(\frac{\sqrt{N}}{T}\right), \tag{2.33}$$

$$\frac{1}{3!\sqrt{N}}\sum_{i=1}^{N}(\hat{\gamma}_{k,i}-\gamma_{k,i})^3 g'''(\tilde{\gamma}_{k,i}) = o_p\left(\frac{\sqrt{N}}{T}\right), \tag{2.34}$$

where B_1 and B_2 are constants. When (2.32), (2.33), and (2.34) hold, we can show the asymptotic normality of the HPJ estimator following an argument similar to that for G_μ that is based on Dhaene and Jochmans (2014).

Therefore, we focus on showing (2.32), (2.33), and (2.34) separately.

The incidental parameter bias; Term (2.32) We first note by the proof of Theorem 2.4 that

$$E\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left((\bar{w}_i)^2 g'(\gamma_{k,i})\right)\right) = \frac{\sqrt{N}}{T}E(V_{T,i}g'(\gamma_{k,i})) = O\left(\frac{\sqrt{N}}{T}\right),$$

where $V_{T,i} := TE((\bar{w}_i)^2|i) = \sum_{j=-T}^{T}\gamma_{j,i}(T-|j|)/T$. The variance is

$$var\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left((\bar{w}_i)^2 g'(\gamma_{k,i})\right)\right) = var((\bar{w}_i)^2 g'(\gamma_{k,i})),$$

by the i.i.d. assumption. We have

$$\begin{aligned} var((\bar{w}_i)^2 g'(\gamma_{k,i})) &\le E((\bar{w}_i)^4 (g'(\gamma_{k,i}))^2) \\ &\le \sqrt{E((\bar{w}_i)^8)}\sqrt{E((g'(\gamma_{k,i}))^4)} \\ &\le \frac{1}{T^2}C\sqrt{E((g'(\gamma_{k,i}))^4)} = O\left(\frac{1}{T^2}\right), \end{aligned}$$

where the second inequality is the Cauchy–Schwarz inequality, the third inequality follows from Assumptions 2.8 and 2.14 and Lemma 2.4, and the last equality follows from Assumption 2.15. We have thus shown that (2.32) holds with $B_1 = \lim_{T\to\infty} E(V_{T,i}g'(\gamma_{k,i}))$.

The bias caused by the nonlinearity of g; Term (2.33) We shall compute the expectation of

$$A := \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\frac{1}{T-k}\sum_{t=k+1}^{T} w_{it}w_{i,t-k} - \gamma_{k,i}\right)^2 g''(\gamma_{k,i}),$$

and will show that

$$E\left((A-E(A))^2\right) = var(A) = o(N/T^2).$$

Under this condition, (2.33) is established by Loéve's c_r inequality and the proof of Theorem 2.4.

We examine the order of $E(A)$. We observe that

$$E(A) = \frac{\sqrt{N}}{(T-k)^2} E\left(\left(\sum_{t=k+1}^{T} (w_{it} w_{i,t-k} - \gamma_{k,i}) \right)^2 g''(\gamma_{k,i}) \right).$$

By the Cauchy–Schwarz inequality, we have

$$\begin{aligned} & E\left(\left(\sum_{t=k+1}^{T} (w_{it} w_{i,t-k} - \gamma_{k,i}) \right)^2 g''(\gamma_{k,i}) \right) \\ & \leq \sqrt{E\left(\left(\sum_{t=k+1}^{T} (w_{it} w_{i,t-k} - \gamma_{k,i}) \right)^4 \right)} \sqrt{E((g''(\gamma_{k,i}))^2)}. \end{aligned}$$

Thus, Assumption 2.15 and Lemma 2.5 imply that

$$|E(A)| = O\left(\frac{\sqrt{N}}{T} \right).$$

Further, $(T/\sqrt{N})E(A)$ converges by the dominated convergence theorem under Assumption 2.15. We set $B_2 = \lim_{T\to\infty}(T/\sqrt{N})E(A)$ (note that $(T/\sqrt{N})E(A)$ does not depend on N).

We next examine $E((A - E(A))^2) = var(A)$ and show that it is of order $O(1/T^2)$. We first note that

$$var(A) = var\left(\left(\frac{1}{T-k} \sum_{t=k+1}^{T} w_{it} w_{i,t-k} - \gamma_{k,i} \right)^2 g''(\gamma_{k,i}) \right),$$

by the i.i.d. assumption. It then holds that

$$\begin{aligned} & var\left(\left(\frac{1}{T-k} \sum_{t=k+1}^{T} w_{it} w_{i,t-k} - \gamma_{k,i} \right)^2 g''(\gamma_{k,i}) \right) \\ & \leq E\left(\left(\frac{1}{T-k} \sum_{t=k+1}^{T} w_{it} w_{i,t-k} - \gamma_{k,i} \right)^4 (g''(\gamma_{k,i}))^2 \right) \\ & \leq \sqrt{E\left(\left(\frac{1}{T-k} \sum_{t=k+1}^{T} w_{it} w_{i,t-k} - \gamma_{k,i} \right)^8 \right)} \sqrt{E((g''(\gamma_{k,i}))^4)} \\ & = O\left(\frac{1}{T^2} \right) = o\left(\frac{N}{T^2} \right), \end{aligned}$$

where the second inequality is the Cauchy–Schwarz inequality and the third equality follows from Assumption 2.15 and Lemma 2.5.

It is therefore shown that (2.33) holds with $B_2 = \lim_{T\to\infty}(T/\sqrt{N})E(A)$ by Markov's inequality.

The third-order term; Term (2.34) We have

$$\begin{aligned} E\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\hat{\gamma}_{k,i}-\gamma_{k,i})^3 g'''(\tilde{\gamma}_{k,i})\right| \leq & ME\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left|(\hat{\gamma}_{k,i}-\gamma_{k,i})^3\right|\right) \\ \leq & M\sqrt{N}E(|\hat{\gamma}_{k,i}-\gamma_{k,i}|^3) \\ \leq & M\sqrt{N}\left(E\left((\hat{\gamma}_{k,i}-\gamma_{k,i})^4\right)\right)^{\frac{3}{4}}, \end{aligned} \tag{2.35}$$

where the first inequality follows from Assumption 2.15 and the triangle inequality, the second inequality follows from the i.i.d. assumption, and the third inequality is the Lyapunov's inequality. Thus, we only need to evaluate the order of $E((\hat{\gamma}_{k,i}-\gamma_{k,i})^4)$. We can write:

$$\begin{aligned} E((\hat{\gamma}_{k,i}-\gamma_{k,i})^4) = & E\Bigg(\Bigg(\frac{1}{T-k}\sum_{t=k+1}^{T}(w_{it}w_{i,t-k}-\gamma_{k,i}) - \frac{T+k}{T-k}(\bar{w}_i)^2 \\ & + \frac{1}{T-k}\sum_{t=1}^{k}w_{it}\bar{w}_i + \frac{1}{T-k}\sum_{t=T-k+1}^{T}w_{it}\bar{w}_i\Bigg)^4\Bigg). \end{aligned} \tag{2.36}$$

Thanks to Loéve's c_r inequality, we only need to examine the fourth-order moment of each term in parentheses on the right-hand side of (2.36).

For the first term, we have

$$E\left(\left(\frac{1}{T-k}\sum_{t=k+1}^{T}(w_{it}w_{i,t-k}-\gamma_{k,i})\right)^4\right) = O\left(\frac{1}{T^2}\right),$$

by Lemma 2.5.

For the second term in (2.36), we first note that $(T+k)/(T-k) = O(1)$. We observe that

$$E(((\bar{w}_i)^2)^4) = E((\bar{w}_i)^8) = O\left(\frac{1}{T^4}\right),$$

by Lemma 2.4. We thus have that

$$E\left(\left(\frac{T+k}{T-k}(\bar{w}_i)^2\right)^4\right) = O\left(\frac{1}{T^4}\right).$$

For the third term, we first observe that by the Cauchy–Schwarz inequality,

$$\begin{aligned} E\left(\left(\frac{1}{T-k}\sum_{t=1}^{k}w_{it}\bar{w}_i\right)^4\right) &= E\left((\bar{w}_i)^4\left(\frac{1}{T-k}\sum_{t=1}^{k}w_{it}\right)^4\right) \\ &\leq \left(E\left((\bar{w}_i)^8\right)\right)^{1/2}\left(E\left(\left(\frac{1}{T-k}\sum_{t=1}^{k}w_{it}\right)^8\right)\right)^{1/2}. \end{aligned}$$

It is shown in the discussion on the second term that $E\left((\bar{w}_i)^8\right)$ is of order $1/T^4$. Moreover, because k is fixed, it is easy to see that

$$E\left(\left(\frac{1}{T-k}\sum_{t=1}^{k}w_{it}\right)^8\right) = O\left(\frac{1}{T^8}\right).$$

Therefore, it holds that

$$E\left(\left(\frac{1}{T-k}\sum_{t=1}^{k} w_{it}\bar{w}_i\right)^4\right) = O\left(\frac{1}{T^6}\right).$$

The same argument can be used to show that

$$E\left(\left(\frac{1}{T-k}\sum_{t=T-k+1}^{T} w_{it}\bar{w}_i\right)^4\right) = O\left(\frac{1}{T^6}\right).$$

Thus, we have shown that

$$E((\hat{\gamma}_{k,i} - \gamma_{k,i})^4) = O\left(\frac{1}{T^2}\right). \tag{2.37}$$

Consequently, we have

$$E\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\hat{\gamma}_{k,i} - \gamma_{k,i})^3 g'''(\tilde{\gamma}_{k,i})\right| = O\left(\frac{\sqrt{N}}{T\sqrt{T}}\right) = o\left(\frac{\sqrt{N}}{T}\right). \tag{2.38}$$

Therefore, we get the desired result (2.34) by Markov's inequality.

□

Proof of Theorem 2.7

As the proofs for $\hat{G}^*_\mu$ and $\hat{G}^*_{\gamma_k}$ are identical, we discuss the case for $\hat{G}^*_{\gamma_k}$ only. We first show that the moments of $\hat{G}^*_{\gamma_k}$ under the bootstrap distribution satisfy Lyapunov's conditions. It is then proved that $\sqrt{N}(\hat{G}^*_{\gamma_k} - \hat{G}_{\gamma_k})$ converges in distribution to Z_{γ_k} almost surely under a subsequence of any subsequences of the original sequence. This implies that the bootstrap distribution of $\sqrt{N}(\hat{G}^*_{\gamma_k} - \hat{G}_{\gamma_k})$ converges almost surely under a subsequence of any subsequences. It then implies that it converges in probability in the original sequence.

We first examine the moments of $\hat{G}^*_{\gamma_k}$. By definition, $\hat{G}^*_{\gamma_k}$ is the sample average of $g(\hat{\gamma}^*_{k,i})$, where $\hat{\gamma}^*_{k,i}$ is the sample autocovariance of the time series z^*_i. It is easy to see that conditionally on the data, the mean and the variance of $g(\hat{\gamma}^*_{k,i})$ are

$$E(g(\hat{\gamma}^*_{k,i})|\{\{y_{it}\}_{t=1}^T\}_{i=1}^N) = \frac{1}{N}\sum_{i=1}^{N} g(\hat{\gamma}_{k,i}) = \hat{G}_{\gamma_k},$$

$$var(g(\hat{\gamma}^*_{k,i})|\{\{y_{it}\}_{t=1}^T\}_{i=1}^N) = \frac{1}{N}\sum_{i=1}^{N}(g(\hat{\gamma}_{k,i}) - \hat{G}_{\gamma_k})^2 = \frac{1}{N}\sum_{i=1}^{N} g(\hat{\gamma}_{k,i})^2 - (\hat{G}_{\gamma_k})^2.$$

The conditional variance converges to

$$E(g(\gamma_{k,i})^2) - (G_{\gamma_k})^2 = var(g(\gamma_{k,i})),$$

in probability by Theorem 2.4 and the continuous mapping theorem. We then consider the third-order moment. We note that

$$E\left(\frac{1}{\sqrt{N}}|g(\hat{\gamma}^*_{k,i}) - \hat{G}_{\gamma_k}|^3|\{\{y_{it}\}_{t=1}^T\}_{i=1}^N\right) = \frac{1}{N^{3/2}}\sum_{i=1}^{N}|g(\hat{\gamma}_{k,i}) - \hat{G}_{\gamma_k}|^3.$$

We note that

$$|g(\hat{\gamma}_{k,i}) - \hat{G}_{\gamma_k}|^3 \leq 4|g(\hat{\gamma}_{k,i})|^3 + 4|\hat{G}_{\gamma_k}|^3.$$

As $\hat{G}_{\gamma_k}$ converges, we have

$$\frac{1}{N^{1/2}}|\hat{G}_{\gamma_k}|^3 \xrightarrow{p} 0.$$

Let $h(\cdot)$ be a twice-differentiable function such that $h(a) \geq 0$, $h(a) = |a|^3$ for $|a| \geq 1$, and $|h(a) - |a|^3| < 1$ for $|a| < 1$. It follows that

$$\frac{1}{N^{3/2}}\sum_{i=1}^{N}|g(\hat{\gamma}_{k,i})|^3 \leq \frac{1}{N^{3/2}}\sum_{i=1}^{N}h(g(\hat{\gamma}_{k,i})) + \frac{1}{N^{3/2}}\sum_{i=1}^{N}|h(g(\hat{\gamma}_{k,i})) - |g(\hat{\gamma}_{k,i})|^3| \xrightarrow{p} 0,$$

as $N^{-1}\sum_{i=1}^{N}h(g(\hat{\gamma}_{k,i})) = O_p(1)$ by Theorem 2.4 under the condition of this theorem and the definition of $h(\cdot)$ implies that $\sum_{i=1}^{N}|h(g(\hat{\gamma}_{k,i})) - |g(\hat{\gamma}_{k,i})|^3|/N^{3/2} \leq 1/N^{1/2}$. Thus, we have

$$\frac{1}{N^{3/2}}\sum_{i=1}^{N}|g(\hat{\gamma}_{k,i}) - \hat{G}_{\gamma_k}|^3 \xrightarrow{p} 0.$$

We argue that for any subsequence of the original sequence, there exists a further subsequence under which $\sqrt{N}(\hat{G}^*_{\gamma_k} - \hat{G}_{\gamma_k})$ converges in distribution conditionally on $\{\{y_{it}\}_{t=1}^{T}\}_{i=1}^{N}$ almost surely. We have shown that the first, second, and third moments of $g(\hat{\gamma}^*_{k,i})$ satisfy Lyapunov's conditions in probability. For any subsequence of the original sequence, there thus exists a further subsequence under which these moment conditions are satisfied almost surely. Thus, under a subsequence of any subsequences, $\sqrt{N}(\hat{G}^*_{\gamma_k} - \hat{G}_{\gamma_k})$ converges in distribution to Z conditionally almost surely. This implies that for any subsequence, there exists a further subsequence under which

$$\sup_{x\in\mathbb{R}}\left|\Pr\left(\sqrt{N}(\hat{G}^*_{\gamma_k} - \hat{G}_{\gamma_k}) \leq x \,\middle|\, \{\{y_{it}\}_{t=1}^{T}\}_{i=1}^{N}\right) - \Pr\left(Z \leq x\right)\right|,$$

converges to 0 almost surely. It thus holds that, for the original sequence,

$$\sup_{x\in\mathbb{R}}\left|\Pr\left(\sqrt{N}(\hat{G}^*_{\gamma_k} - \hat{G}_{\gamma_k}) \leq x \,\middle|\, \{\{y_{it}\}_{t=1}^{T}\}_{i=1}^{N}\right) - \Pr\left(Z \leq x\right)\right| \xrightarrow{p} 0.$$

□

Proof of Theorem 2.9

We give only the proof for $KS_1^{\gamma_k}$ because that for KS_1^{μ} is the same. The proof is almost identical to the proof of Corollary 19.21 in van der Vaart (1998). We first note that, under $H_0^{\gamma_k}$, $\sqrt{N}(\mathbb{P}_N^{\hat{\gamma}_k} - Q^{\gamma_k}) \rightsquigarrow \mathbb{G}_{Q^{\gamma_k}}$ in $\ell^\infty(\mathcal{F})$ given $N, T \to \infty$ with $N/T \to 0$ by Theorem 2.2. Therefore, because the norm $\|\cdot\|_\infty$ on $D[-\infty,\infty]$, where $D[-\infty,\infty]$ is the class of all cadlag functions from $[-\infty,\infty]$ into $\mathbb{R}$, is continuous with respect to the uniform norm, we have $KS_1^{\gamma_k} \rightsquigarrow \|\mathbb{G}_{Q^{\gamma_k}}\|_\infty$ under $H_0^{\gamma_k}$ by the continuous mapping theorem.

□

Proof of Theorem 2.10

We present only the proof for $KS_2^{\gamma_k}$ because that for KS_2^{μ} is the same. We first observe that

$$KS_2^{\gamma_k} = \left\| \sqrt{\frac{N_1 N_2}{N_1+N_2}}(\mathbb{P}^{\hat{\gamma}_k}_{N_1,(1)} - P^{\gamma_k}_{0,(1)}) - \sqrt{\frac{N_1 N_2}{N_1+N_2}}(\mathbb{P}^{\hat{\gamma}_k}_{N_2,(2)} - P^{\gamma_k}_{0,(2)}) + \sqrt{\frac{N_1 N_2}{N_1+N_2}}(P^{\gamma_k}_{0,(1)} - P^{\gamma_k}_{0,(2)}) \right\|_\infty .$$

We note that, under Assumption 2.18, $\sqrt{N_1}(\mathbb{P}^{\hat{\gamma}_k}_{N_1,(1)} - P^{\gamma_k}_{0,(1)})$ and $\sqrt{N_2}(\mathbb{P}^{\hat{\gamma}_k}_{N_2,(2)} - P^{\gamma_k}_{0,(2)})$ jointly converge in distribution to independent Brownian processes $\mathbb{G}_{P^{\gamma_k}_{0,(1)}}$ and $\mathbb{G}_{P^{\gamma_k}_{0,(2)}}$ given $N_1, T_1 \to \infty$ with $N_1/T_1 \to 0$ and $N_2, T_2 \to \infty$ with $N_2/T_2 \to 0$ by Theorem 2.2. Therefore, under $H_0^{\gamma_k} : P^{\gamma_k}_{0,(1)} = P^{\gamma_k}_{0,(2)}$, $KS_2^{\gamma_k}$ converges in distribution to

$$\left\| \sqrt{1-\lambda}\mathbb{G}_{P^{\gamma_k}_{0,(1)}} - \sqrt{\lambda}\mathbb{G}_{P^{\gamma_k}_{0,(2)}} \right\|_\infty ,$$

by the continuous mapping theorem given $N_1/(N_1+N_2) \to \lambda \in (0,1)$. It is easy to see that the distribution of the limit random variable $\sqrt{1-\lambda}\mathbb{G}_{P^{\gamma_k}_{0,(1)}} - \sqrt{\lambda}\mathbb{G}_{P^{\gamma_k}_{0,(2)}}$ is identical to that of $\mathbb{G}_{P^{\gamma_k}_{0,(1)}}$ under $H_0^{\gamma_k}$. Thus, we have the desired result.

□

2.9.2 Technical lemmas

Lemma 2.1 (Galvao and Kato (2014) based on Davydov (1968)). *Let $\{\xi_t\}_{t=1}^\infty$ denote a stationary process taking values in $\mathbb{R}$ and let $\alpha(m)$ denote its α-mixing coefficients. Suppose that $E(|\xi_1|^q) < \infty$ and $\sum_{m=1}^\infty \alpha(m)^{1-2/q} < \infty$ for some $q > 2$. Then, we have*

$$var\left(\sum_{t=1}^T \xi_t\right) \le CT$$

with $C = 12(E(|\xi_1|^q))^{2/q} \sum_{m=0}^\infty \alpha(m)^{1-2/q}$.

Proof. The proof is available in Galvao and Kato (2014) (the discussion after Theorem C.1). □

Lemma 2.2 (Yokoyama (1980)). *Let $\{\xi_t\}_{t=1}^\infty$ denote a strictly stationary α-mixing process taking values in $\mathbb{R}$, and let $\alpha(m)$ denote its α-mixing coefficients. Suppose that $E(\xi_t) = 0$ and for some constants $\delta > 0$ and $r > 2$, $E(|\xi_1|^{r+\delta}) < \infty$. If $\sum_{m=0}^\infty (m+1)^{r/2-1}\alpha(m)^{\delta/(r+\delta)} < \infty$, then there exists a constant C independent of T such that*

$$E\left(\left|\sum_{t=1}^T \xi_t\right|^r\right) \le CT^{r/2}.$$

Lemma 2.3. *Suppose that Assumptions 2.1 and 2.4 hold. Then, $\{w_{it}w_{i,t-k}\}_{t=k+1}^\infty$ for a fixed k given α_i is strictly stationary and α-mixing and its mixing coefficients $\{\alpha_k(m|i)\}_{m=0}^\infty$ possess the following properties: there exists a sequence $\{\alpha_k(m)\}_{m=0}^\infty$ such that for any i and m, $\alpha_k(m|i) \le \alpha_k(m)$ and $\sum_{m=0}^\infty (m+1)^3\alpha_k(m)^{\delta/(r+\delta)} < \infty$ for some $\delta > 0$ and $r = 4$. The result holds with $r = 8$ if Assumption 2.4 is replaced by Assumption 2.8.*

Proof. The proof is similar to the proof of Theorem 14.1 in Davidson (1994). It is easy to see that for any i and any $0 \leq m < k$, $\alpha_k(m|i) \leq 1$, and that for any i and any $m \geq k$, $\alpha_k(m|i) \leq \alpha(m-k|i) \leq \alpha(m-k)$ by the definition of α-mixing coefficients and Assumption 2.4 or 2.8. Thus, we have $\sum_{m=0}^{\infty}(m+1)^3\alpha_k(m)^{\delta/(r+\delta)} \leq \sum_{m=0}^{k-1}(m+1)^3 + \sum_{m=k}^{\infty}(m+1)^3\alpha(m-k)^{\delta/(r+\delta)} < \infty$ for $r = 4$ and 8 under Assumptions 2.4 and 2.8, respectively. Thus, the lemma holds. □

Lemma 2.4. *Suppose that Assumptions 2.1, 2.4, and 2.5 hold. Then, it holds that* $E((\bar{w}_i)^r) \leq CT^{-r/2}$ *for* $r = 2, 4$ *and some constant* $C < \infty$. *If Assumptions 2.8 and 2.9 hold additionally,* $E((\bar{w}_i)^8) \leq CT^{-4}$ *holds as well.*

Proof. We first consider the case with $r = 2$. Given $E(\bar{w}_i|i) = 0$, Lemma 2.1 states that

$$E\left((\bar{w}_i)^2|i\right) \leq C_i/T,$$

where $C_i = 12(E(|w_{it}|^{(4+\delta)/2}|i))^{4/(4+\delta)} \sum_{m=0}^{\infty} \alpha(m|i)^{\delta/(4+\delta)}$. Assumption 2.4 implies that

$$C_i \leq 12(E(|w_{it}|^{(4+\delta)/2}|i))^{4/(4+\delta)} \sum_{m=0}^{\infty} \alpha(m)^{\delta/(4+\delta)}.$$

Thus, we have

$$\begin{aligned}
E\left((\bar{w}_i)^2\right) &= E\left(E\left((\bar{w}_i)^2|i\right)\right) \\
&\leq 12E\left((E(|w_{it}|^{(4+\delta)/2}|i))^{4/(4+\delta)}\right) \sum_{m=0}^{\infty} \alpha(m)^{\delta/(4+\delta)}/T \\
&\leq 12\left(E\left(E(|w_{it}|^{(4+\delta)/2}|i)\right)\right)^{4/(4+\delta)} \sum_{m=0}^{\infty} \alpha(m)^{\delta/(4+\delta)}/T \\
&= 12\left(E\left(|w_{it}|^{(4+\delta)/2}\right)\right)^{4/(4+\delta)} \sum_{m=0}^{\infty} \alpha(m)^{\delta/(4+\delta)}/T \\
&= O\left(\frac{1}{T}\right),
\end{aligned}$$

where the second inequality is Jensen's inequality and the last equality follows from Assumptions 2.4 and 2.5. Hence, the desired result holds for $r = 2$.

Next, we consider the case with $r = 4, 8$. We use Lemma 2.2. From the proof of Lemma 2.2 available in Yokoyama (1980), we have

$$E\left(\left|\sum_{t=1}^{T} w_{it}\right|^r \middle| i\right) \leq K_{r,i}\left(E\left(|w_{it}|^{r+\delta}|i\right)\right)^{r/(r+\delta)} T^{r/2}$$

for some $\delta > 0$, where $K_{r,i}$ is a polynomial of $A_q(\alpha|i)$ for $q \leq r$ and $A_q(\alpha|i) := \sum_{m=0}^{\infty}(m+1)^{q/2-1}\alpha(m|i)^{\delta/(q+\delta)}$. Note that $A_q(\alpha|i) < \infty$ for $q \leq r$ if $A_r(\alpha|i) < \infty$. By Assumption 2.4 or

2.8, there exists a constant $K_r < \infty$ such that $K_{r,i} < K_r$ for all i. Thus, we have

$$\begin{aligned} E\left((\bar{w}_i)^r\right) = E\left(E\left((\bar{w}_i)^r|i\right)\right) &\leq K_r E\left(\left(E\left(|w_{it}|^{r+\delta}|i\right)\right)^{r/(r+\delta)}\right) T^{-r/2} \\ &\leq K_r \left(E\left(E\left(|w_{it}|^{r+\delta}|i\right)\right)\right)^{r/(r+\delta)} T^{-r/2} \\ &= K_r \left(E\left(|w_{it}|^{r+\delta}\right)\right)^{r/(r+\delta)} T^{-r/2} \\ &= O(T^{-r/2}), \end{aligned}$$

where the second inequality is Jensen's inequality and the last equality follows from Assumption 2.5 or 2.9. The proof for $r = 4, 8$ is complete. □

Lemma 2.5. *Suppose that Assumptions 2.1, 2.4, and 2.5 hold. Then, it holds that* $E((\sum_{t=k+1}^{T}(w_{it}w_{i,t-k} - \gamma_{k,i}))^r) \leq CT^{r/2}$ *for some constant* C *and* $r = 2$. *The result holds for* $r = 4$ *if Assumptions 2.4 and 2.5 are replaced by Assumptions 2.8 and 2.9. Furthermore, if Assumption 2.14 holds additionally, the result holds for* $r = 8$ *as well.*

Proof. In view of Lemma 2.3, the lemma follows along the same line as that of Lemma 2.4. □

Lemma 2.6. *Under Assumptions 2.1 and 2.2, we have*

$$E\left((\bar{y}_i - \mu_i)^2\right) = O(T^{-1}).$$

Proof. Note that $\bar{y}_i = \mu_i + \bar{w}_i$ where $\bar{w}_i := T^{-1}\sum_{t=1}^{T} w_{it}$. Therefore, we have

$$E\left((\bar{y}_i - \mu_i)^2\right) = E\left(\bar{w}_i^2\right) = \frac{1}{T}E(V_{T,i}) = O(T^{-1}),$$

where $V_{T,i} := TE((\bar{w}_i)^2|i) = \sum_{j=-T}^{T}\gamma_{j,i}(T-|j|)/T$, the second equality follows from Assumption 2.1, and the last equality follows from Assumption 2.2.

□

Lemma 2.7. *Under Assumptions 2.1, 2.4, and 2.5, we have*

$$E((\hat{\gamma}_{k,i} - \gamma_{k,i})^2) = O(T^{-1}).$$

Proof. Given the estimator $\hat{\gamma}_{k,i}$ has the following decomposition:

$$\begin{aligned} \hat{\gamma}_{k,i} \quad = \quad & \frac{1}{T-k}\sum_{t=k+1}^{T} w_{it}w_{i,t-k} - \frac{T+k}{T-k}(\bar{w}_i)^2 \\ & + \frac{1}{T-k}\sum_{t=1}^{k} w_{it}\bar{w}_i + \frac{1}{T-k}\sum_{t=T-k+1}^{T} w_{it}\bar{w}_i, \end{aligned}$$

we can write:

$$\begin{aligned} E((\hat{\gamma}_{k,i} - \gamma_{k,i})^2) = E\Bigg(\Bigg(& \frac{1}{T-k}\sum_{t=k+1}^{T}(w_{it}w_{i,t-k} - \gamma_{k,i}) - \frac{T+k}{T-k}(\bar{w}_i)^2 \\ & + \frac{1}{T-k}\sum_{t=1}^{k} w_{it}\bar{w}_i + \frac{1}{T-k}\sum_{t=T-k+1}^{T} w_{it}\bar{w}_i\Bigg)^2\Bigg). \end{aligned} \tag{2.39}$$

Owing to Loéve's c_r inequality, we just need to examine the second-order moment of each term in parentheses on the right-hand side of (2.39).

For the first term, we have

$$E\left(\left(\frac{1}{T-k}\sum_{t=k+1}^{T}(w_{it}w_{i,t-k}-\gamma_{k,i})\right)^2\right)=O\left(\frac{1}{T}\right),$$

by Lemma 2.5.

For the second term in (2.39), we first note that $(T+k)/(T-k)=O(1)$. We observe that

$$E(((\bar{w}_i)^2)^2)=E((\bar{w}_i)^4)=O\left(\frac{1}{T^2}\right),$$

by Lemma 2.4. We thus have that

$$E\left(\left(\frac{T+k}{T-k}(\bar{w}_i)^2\right)^2\right)=O\left(\frac{1}{T^2}\right).$$

For the third term, we first observe that by the Cauchy–Schwarz inequality,

$$\begin{aligned}E\left(\left(\frac{1}{T-k}\sum_{t=1}^{k}w_{it}\bar{w}_i\right)^2\right)&=E\left((\bar{w}_i)^2\left(\frac{1}{T-k}\sum_{t=1}^{k}w_{it}\right)^2\right)\\&\le\left(E\left((\bar{w}_i)^4\right)\right)^{1/2}\left(E\left(\left(\frac{1}{T-k}\sum_{t=1}^{k}w_{it}\right)^4\right)\right)^{1/2}.\end{aligned}$$

It is shown in the discussion on the second term that $E\left((\bar{w}_i)^4\right)$ is of order $1/T^2$. Moreover, because k is fixed, it is easy to see that

$$E\left(\left(\frac{1}{T-k}\sum_{t=1}^{k}w_{it}\right)^4\right)=O\left(\frac{1}{T^4}\right).$$

Therefore, it holds that

$$E\left(\left(\frac{1}{T-k}\sum_{t=1}^{k}w_{it}\bar{w}_i\right)^2\right)=O\left(\frac{1}{T^3}\right).$$

The same argument can be used to show that

$$E\left(\left(\frac{1}{T-k}\sum_{t=T-k+1}^{T}w_{it}\bar{w}_i\right)^2\right)=O\left(\frac{1}{T^3}\right).$$

We have thus shown that all of the terms are $O(T^{-1})$ and the statement of the lemma holds. □

Lemma 2.8. *Let a_T and b_T be continuous random variables indexed by T with bounded joint density. Suppose that as $T\to\infty$,*

$$E(|a_T-b_T|^p)=O(T^c), \tag{2.40}$$

for some integer p and real number $c < 0$. It then holds that

$$\sup_x |\Pr(a_T < x) - \Pr(b_T < x)| = O(T^{2c/(2+p)}). \tag{2.41}$$

In particular, if $c = -1$ and $p = 2$, then $2c/(2+p) = -1/2$ and $\sup_x |\Pr(a_T < x) - \Pr(b_T < x)| = O(T^{-1/2})$.

Proof. We have

$$\Pr(a_T < x) = \Pr(a_T < x, b_T < x) + \Pr(a_T < x, b_T \geq x).$$

We take some $\epsilon > 0$. Then, we have

$$\begin{aligned}
&\Pr(a_T < x, b_T \geq x) \\
=&\Pr(a_T < x, b_T \geq x, |a_T - b_T| > \epsilon) + \Pr(a_T < x, b_T \geq x, |a_T - b_T| \leq \epsilon).
\end{aligned}$$

For the first probability on the right-hand side, we have

$$\sup_x \Pr(a_T < x, b_T \geq x, |a_T - b_T| > \epsilon) \leq \Pr(|a_T - b_T| > \epsilon) \leq \frac{E(|a_T - b_T|^p)}{\epsilon^p},$$

by Markov's inequality. For the second probability, we have

$$\begin{aligned}
\sup_x \Pr(a_T < x, b_T \geq x, |a_T - b_T| \leq \epsilon) \leq& \sup_x \Pr(x - \epsilon \leq a_T < x, x \leq b_T \leq x + \epsilon) \\
\leq& \epsilon^2 \sup_x \sup_{x-\epsilon \leq a < x, x \leq b \leq x+\epsilon} f_{a_T, b_T}(a, b) \\
\leq& \epsilon^2 C,
\end{aligned}$$

for some $C > 0$, where f_{a_T, b_T} is the joint density of a_T and b_T. Therefore, we have

$$\sup_x |\Pr(a_T < x) - \Pr(a_T < x, b_T < x)| \leq \frac{E(|a_T - b_T|^p)}{\epsilon^p} + \epsilon^2 C.$$

We now take $\epsilon = T^d$. Then, we have

$$\sup_x |\Pr(a_T < x) - \Pr(a_T < x, b_T < x)| = O\left(\frac{T^c}{T^{dp}} + T^{2d}\right).$$

We note that the above order is minimized by setting $d = c/(2+p)$. Thus, we have

$$\sup_x |\Pr(a_T < x) - \Pr(a_T < x, b_T < x)| = O\left(T^{2c/(2+p)}\right).$$

Similarly, we have

$$\sup_x |\Pr(b_T < x) - \Pr(a_T < x, b_T < x)| = O\left(T^{2c/(2+p)}\right).$$

Therefore, we have

$$\begin{aligned}
&\sup_x |\Pr(a_T < x) - \Pr(b_T < x)| \\
=&\sup_x |(\Pr(a_T < x) - \Pr(a_T < x, b_T < x)) - (\Pr(b_T < x) - \Pr(a_T < x, b_T < x))| \\
\leq&\sup_x |\Pr(a_T < x) - \Pr(a_T < x, b_T < x)| + \sup_x |\Pr(b_T < x) - \Pr(a_T < x, b_T < x)| \\
=&O\left(T^{2c/(2+p)}\right).
\end{aligned}$$

□

Chapter 3

The Effect of Measurement Error in the Sharp Regression Discontinuity Design

3.1 Introduction

This paper develops a nonparametric analysis for the sharp regression discontinuity (RD) design in which the continuous forcing variable may contain classical measurement error. We show that the measurement error causes severe bias for identifying the average treatment effect given the "true" forcing variable at the discontinuity point. We then examine to what extent the average treatment effect can be studied from observed data containing the measurement error. The average treatment effect is approximated using an identified parameter based on the small error variance approximation (SEVA) originally proposed by Chesher (1991). We also develop an estimation procedure for the approximating parameter based on local polynomial regressions and the kernel density estimation.

The RD design was first introduced by Thistlethwaite and Campbell (1960) and has been substantially studied in theoretical econometrics. It is known as a quasi-experimental design, which is a powerful design for treatment effect analyses and program evaluation. Examples of theoretical studies include research on identification (Hahn, Todd, and van der Klaauw, 2001; Lee, 2008; Frandsen, Frölich, and Melly, 2012), estimation (Porter, 2003; Imbens and Kalyanaraman, 2012; Arai and Ichimura, 2014), and inference methods (Lee and Card, 2008; McCrary, 2008; Calonico, Cattaneo, and Titiunik, 2014). In addition, much of the empirical literature has developed analyses based on RD designs because of its utility. For example, many studies have been conducted for education (e.g., Angrist and Lavy, 1999) and health economics (e.g., Card and Shore-Sheppard, 2004). Useful surveys on the RD literature can be seen in Imbens and Lemieux (2008) and Lee and Lemieux (2010).

Despite the vast body of RD literature, studies on RD designs with measurement error are scant (see the paragraph "Related literature" below). In RD designs, a treatment is completely or partly determined by whether a forcing variable[1] is greater than a known threshold. In the

[1]It is also referred to as the "assignment" or "running" variable in the literature.

sharp RD design in which the treatment is completely determined by the forcing variable, the average treatment effect at the threshold is identified by the difference in the means of the outcome marginally above and below the threshold (Hahn et al., 2001). See Figure 3.1 for an intuitive understanding of this. However, if the observed forcing variable contains measurement error, we cannot observe the "true" forcing variable that determines the treatment. Identification analyses for the average treatment effect with the observed mismeasured forcing variable have not been developed enough.

There are many empirical situations based on RD designs in which the observed forcing variable may be mismeasured. Empirical studies based on RD designs often use survey data. For example, Hullegie and Klein (2010) analyze the effect of private insurance coverage on individual health performance (e.g., the number of doctor visits) using a unique public insurance system in Germany. In Germany, employees whose income is below a threshold cannot buy private insurance, so that this unique system provides an RD design. Their forcing variable is individual income, which is found using data from the German Socio-Economic Panel. They also indicate that their forcing variable seems to contain measurement error, because some people buy private insurance despite that their income is below the threshold (i.e., despite their supposed ineligibility to buy private insurance). There are many other applied studies based on RD designs that use survey data to conduct causal analyses, such as Card, Dobkin, and Maestas (2008), Battistin, Brugiavini, Rettore, and Weber (2009), Schanzenbach (2009), and Koch (2013). In such situations, there is always a risk that the forcing variable may be mismeasured, as in other literature in econometrics (see Bound, Brown, and Mathiowetz, 2001 and Schennach, 2013 for surveys on the literature of measurement error in econometrics).

This study first investigates the effect of the forcing variable with classical measurement error in the sharp RD design. We demonstrate that the difference in the conditional means of the outcome given the mismeasured forcing variable marginally above and below the threshold has an identification bias for the average treatment effect given the true forcing variable at the threshold. The identification bias is critical in the sense that even if there is a significant treatment effect, the bias misleads the researchers to the incorrect result of no treatment effect. Furthermore, the measurement error leads the conditional probability of the treatment to be continuous at the threshold. We derive the specific form of the identification bias caused by the measurement error.

To examine the average treatment effect using the mismeasured forcing variable, we then suggest approximating it based on the SEVA originally developed by Chesher (1991). The accuracy of the approximation depends on the magnitude of the variance of the measurement error, σ^2. We show that the average treatment effect is approximated up to the order $O(\sigma^2)$ based on the SEVA when σ is small. In other words, the smaller standard deviation of the measurement error implies a more precise approximation for the average treatment effect. We thus consider that our approximate analysis is appropriate for empirical studies based on survey data in which the forcing variable may contain measurement error caused from incorrect entry

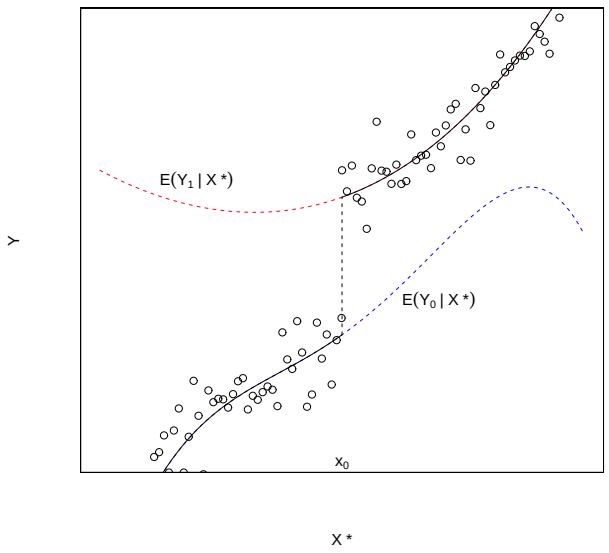


Figure 3.1: The dots indicate pairs of the observed outcome Y_i and the "true" forcing variable X_i^*. The black solid line is the conditional mean of Y_i given X_i^*, which is identified if (Y_i, X_i^*) is observed. The upper dotted red and lower dotted blue lines are the conditional means for the treated and untreated, respectively. The length of the dotted vertical line indicates the average treatment effect at the threshold x_0, $E(Y_{1i} - Y_{0i}|X_i^* = x_0)$, which is identified by the difference in the conditional means of Y_i marginally above and below x_0.

or a memory lapse. Such measurement error can be classical and the variance can be small. Importantly, while σ^2 is generally unknown, σ^2 can be extrapolated or forecast. Additional data for the true forcing variable, such as public data or census data, allow us to conduct an extrapolation under the classical measurement error assumption. Importantly, our approach does not require additional variables such as instruments or repeated measurements, which are often unavailable in empirical situations.

We also provide a nonparametric estimation procedure for the parameter that approximates the average treatment effect based on local polynomial regressions and the kernel density estimation. We derive the consistency and asymptotic normality of the nonparametric estimator. Combining the asymptotic properties with our approximate analysis, the average treatment effect is approximately estimated up to the order $O(\sigma^2)$.

We conduct Monte Carlo simulations to investigate the practical effects of the measurement error on identification for the average treatment effect. The simulations also demonstrate the performance of our approximate analysis. We find that the measurement error critically affects identification of the average treatment effect in our simulation designs. The results of the Monte Carlo simulations also corroborate that our approximate analysis can function even when the magnitude of σ^2 accounts for 20 percent of the variance of a mismeasured forcing variable.

Related literature: As a study in RD literature, this paper is closely related to Battistin et al. (2009), Hullegie and Klein (2010), and Yu (2012).

Battistin et al. (2009) develop a fuzzy RD analysis in which the forcing variable contains measurement error. Their analysis is based on the non-differential measurement error and certain smoothness conditions on the joint distribution of the true and mismeasured forcing variables. They show that under these conditions, the average treatment effect on the treated is identified based on the mismeasured forcing variable using the fuzzy RD estimand, that is, the Wald estimand. However, their analysis does not function under continuous measurement error, because the conditional probability of the treatment given the mismeasured forcing variable is not discontinuous owing to the continuous measurement error (see Remark 3.3). Indeed, our Monte Carlo simulations demonstrate that the Wald estimator is highly unstable, because the measurement error leads the conditional probability of the treatment to be continuous.

Hullegie and Klein (2010) develop a fuzzy RD analysis with a continuous mismeasured forcing variable. Their analysis is based on the linear functional-form specifications and the normally distributed measurement error independent of the observable forcing variable (such measurement error is not classical). They use the parametric specifications to identify and estimate a local average treatment effect in an RD design. In contrast, the present study focuses on a sharp RD design with classical measurement error in a nonparametric manner.

Yu (2012) studies RD designs with an observable continuous forcing variable containing classical measurement error. Although the model studied in the present paper is similar to that in his paper,[2] the approaches developed in both papers differ. He focuses on the conditions under which the average treatment effect at the threshold can be consistently estimated based on the difference in the mean outcomes given the mismeasured forcing variable and treatment at the threshold (see Remark 3.4). He shows that if the measurement error shrinks to zero depending on the sample size under some rate conditions, a local polynomial estimator for the difference is consistent for the average treatment effect. By contrast, this paper first approximates the average treatment effect in the population, which is based on the SEVA, and then develops an estimation for the approximating parameter. As a result, the identification and estimation approaches in both papers differ. For a better understanding, we compare the performance of the analyses developed in both papers using our Monte Carlo simulations, which reveal that our approximate analysis is more successful.

The present paper is also related to Pei (2011) and Dong (2014), but the objectives in those studies differ from that in our paper. Pei (2011) studies a model in which both a true forcing variable and measurement error are discretely distributed with bounded support. He develops an identification analysis for the average treatment effect utilizing the discreteness of the forcing variable. Dong (2014) studies RD designs in which a true unobservable forcing variable is continuous but the observed forcing variable is discretized or rounded, such as age in

[2]More specifically, the setting considered in the present paper is almost the same as "Case 2" in Yu (2012). He also considers other settings in which the treatment is determined using the mismeasured forcing variable and/or in which the treatment is unobservable.

years. She provides modified identification and estimation procedures for the average treatment effect based on parametric polynomial modeling. As she mentions, such a rounding error cannot be classical. In contrast, the present paper develops a nonparametric analysis for the problem of classical measurement error in a sharp RD design with a continuous forcing variable.

As a study in the literature of measurement error, the present paper is related to Chesher (1991), Chesher and Schluter (2002), and Battistin and Chesher (2014). These papers use the SEVA to investigate the effects of measurement error in separate settings. However, to our knowledge, no study applies the SEVA to the problem of measurement error in RD designs. Our study builds on the literature by showing the conditions under which the average treatment effect in the RD design is approximated based on the SEVA.

Organization of rest of the paper: Section 3.2 introduces our setting and the parameter of interest. Section 3.3 examines the effect of the measurement error for identifying the average treatment effect. Section 3.4 develops our approximate analysis based on the SEVA. Section 3.5 develops an estimation method for the approximating parameter. Section 3.6 describes the Monte Carlo simulations. Section 3.7 concludes. All proofs are provided in the Appendix.

Notations: For generic random variables Z_i and W_i, we denote the conditional density (distribution function) of Z_i given $W_i = w$ as $f_{Z|W}(\cdot|w)$ ($F_{Z|W}(\cdot|w)$). We denote the support of Z_i as $supp(Z)$. For a generic function $g(\cdot)$, we denote the left and right limits of $g(x)$ at x_0 as $g(x_0-) \coloneqq \lim_{e \uparrow 0} g(x_0+e)$ and $g(x_0+) \coloneqq \lim_{e \downarrow 0} g(x_0+e)$, respectively. For $s \in \mathbb{N}$, we denote the s-th order (partial) derivative of $g(x)$ with respect to x as $g^{(s)}(x)$. The indicator function $\mathbf{1}(E)$ is 1 if the event E is true and 0 otherwise. We denote a $K \times L$ matrix B with the (k,l) entry $b_{k,l}$ as $B = (b_{k,l})_{(k,l)}$ for $k = 1, \ldots, K$ and $l = 1, \ldots, L$.

3.2 Settings

We observe independent and identically distributed (i.i.d.) random variables $\{(Y_i, D_i, X_i)\}_{i=1}^n$, where $Y_i \in supp(Y) \subset \mathbb{R}$ is an outcome, $D_i \in \{0,1\}$ is a treatment that depends on an unobservable "true" continuous forcing variable $X_i^* \in supp(X^*) \subset \mathbb{R}$, and $X_i \in supp(X) \subset \mathbb{R}$ is the observable continuous forcing variable that may contain measurement error. If unit i is treated, $D_i = 1$, otherwise, $D_i = 0$. We can write $Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}$, where Y_{1i} is the potential outcome when unit i is treated and Y_{0i} is that when untreated. Both Y_{1i} and Y_{0i} cannot be observed for any unit, because no units can be both treated and untreated. This is the standard potential outcome notation.

Suppose that D_i is completely determined using X_i^* as follows:

$$D_i = \mathbf{1}(X_i^* \geq x_0), \tag{3.1}$$

where $x_0 \in supp(X^*)$ is a known fixed threshold. The relationship is commonly referred as the sharp RD design in RD literature (see Lee and Lemieux, 2010). Equation (3.1) means

that all units with $X_i^* \geq x_0$ are treated and all units with $X_i^* < x_0$ are untreated. Here, $E(D_i|X_i^* = x) = \mathbf{1}(x \geq x_0)$ is the deterministic function of x, and it is discontinuous at x_0.

If we can observe the true forcing variable X_i^*, the average treatment effect given $X_i^* = x_0$ is identified under the continuity of $E(Y_{di}|X_i^* = x)$ for $d \in \{0, 1\}$ (Hahn et al., 2001):

$$E(Y_{1i} - Y_{0i}|X_i^* = x_0) = \tau^*, \tag{3.2}$$

where

$$\tau^* \coloneqq E(Y_i|X_i^* = x_0+) - E(Y_i|X_i^* = x_0-) \tag{3.3}$$

$$= E(Y_i|X_i^* = x_0+, D_i = 1) - E(Y_i|X_i^* = x_0-, D_i = 0). \tag{3.4}$$

The left-hand side of (3.2) is the average treatment effect at the threshold, which is the common parameter of interest in the sharp RD design. τ^* is the difference in the conditional means of Y_i given X_i^* (and D_i) at the threshold. The right-hand side of (3.3) is equal to that of (3.4) because of (3.1). If (Y_i, X_i^*) is observed, the right-hand sides of (3.3) and (3.4) are identified, so the average treatment effect is also identified.

Because of the presence of measurement error, we cannot observe the true forcing variable X_i^*, so the right-hand sides of (3.3) and (3.4) cannot be identified. We instead observe X_i that contains measurement error, as follows:

$$X_i = X_i^* + \sigma U_i, \tag{3.5}$$

where the random variable σU_i is continuous measurement error with $E(U_i) = 0$ and $var(U_i) = 1$, and $\sigma \geq 0$ indicates the standard deviation. This additive representation is commonly used in the literature of measurement error (see, e.g., Schennach, 2013). We assume $x_0 \in supp(X)$.

We introduce the following assumptions for the measurement error.

Assumption 3.1. *(i)* U_i *is independent of* $(Y_{1i}, Y_{0i}, D_i, X_i^*)$. *(ii)* $f_U(\cdot)$ *is continuous on bounded support. (iii)* $E(U_i) = 0, var(U_i) = 1, E|U_i|^3 < \infty$.

Assumption 3.1 (i) is the classical measurement error assumption (see Bound et al., 2001 and Schennach, 2013 for the interpretation). This assumption requires joint independence between the measurement error and the other variables. Assumption 3.1 (i) is identical to the independence between U_i and (Y_{1i}, Y_{0i}, X_i^*) in the sharp RD design, because D_i is the deterministic function of X_i^*. Assumption 3.1 (ii) ensures the continuity of U_i with bounded support. The bounded support is required to guarantee the establishment of the approximation developed in Section 3.4. This condition may be restrictive in the theoretical view, but it can be satisfied in many empirical situations. Assumption 3.1 (iii) is a mild moment condition. The existence of the third-order moment is unrestrictive under the bounded support of U_i.

Remark 3.1. We can allow the outcome to contain measurement error. In other words, we can allow a situation in which we observe not the "true" outcome Y_i^* but the mismeasured outcome

Y_i. The analysis in this paper remains unchanged if the measurement error contained in Y_i is independent of $(Y_{1i}, Y_{0i}, D_i, X_i^*, U_i)$ and has mean zero. Thus, we do not explicitly consider that the outcome is mismeasured in this paper.

Remark 3.2. We do not allow a situation in which D_i is also mismeasured, which would require other approaches to analyze the problem caused by the measurement error.

3.3 Identification bias caused by measurement error

This section investigates the effect of the measurement error for identifying the average treatment effect at the threshold. We show that the measurement error leads the difference in the mean outcomes just above and below the threshold and the discontinuity of the conditional probability of the treatment to vanish. We then discuss possible approaches to examine the average treatment effect using the mismeasured forcing variable.

Although the parameter of interest in the RD design is the average treatment effect at the threshold, we focus on studying the effect of the measurement error for identifying τ^*. This is because τ^* equals the average treatment effect under the continuity of $E(Y_{di}|X_i^* = x)$ for $d \in \{0, 1\}$, as discussed in the previous section.

Observing (Y_i, D_i, X_i), it is not uncommon to consider the following parameters:

$$\tau_X \coloneqq E(Y_i|X_i = x_0+) - E(Y_i|X_i = x_0-),$$
$$\tau_{XD} \coloneqq E(Y_i|X_i = x_0+, D_i = 1) - E(Y_i|X_i = x_0-, D_i = 0).$$

τ_X replaces the true forcing variable X_i^* in (3.3) with the mismeasured forcing variable X_i. Similarly, τ_{XD} replaces X_i^* in (3.4) with X_i. τ_X and τ_{XD} are identified using the observable data. Importantly, τ_X and τ_{XD} generally differ, because D_i is not generally a deterministic function of X_i, that is, $D_i \neq \mathbf{1}(X_i \geq x_0)$.

We can guess that τ_X has a severe identification bias for τ^* and that τ_X is close to zero. We observe that $E(Y_i|X_i = x_0+)$ is computed based on a subset of units with $X_i \geq x_0$ (i.e., the right half of Figure 3.2). Units with $X_i \geq x_0$ but $X_i^* < x_0$ may lead $E(Y_i|X_i = x_0+)$ to substantially differ from $E(Y_i|X_i^* = x_0+)$, because the conditional distribution of Y_i can be discontinuous at $X_i^* = x_0$ owing to the RD structure, that is, because the realization of Y_i for $X_i^* < x_0$ can differ from those for $X_i^* \geq x_0$ (see Figure 3.1). Similarly, $E(Y_i|X_i = x_0-)$ is computed based on a subset with the remaining units (i.e., the left half of Figure 3.2) and it can substantially differ from $E(Y_i|X_i^* = x_0-)$ by the influence of the units with $X_i < x_0$ but $X_i^* \geq x_0$. As a result, τ_X could have a severe bias for identifying τ^*. As demonstrated in a later theorem, τ_X is equal to zero because of the bias.

In contrast, we can guess that τ_{XD} does not substantially differ from τ^*. We observe that $E(Y_i|X_i = x_0+, D_i = 1)$ is computed based on a subset of units with $X_i \geq x_0$ and $X_i^* \geq x_0$ (i.e., the upper right in Figure 3.2). Because $E(Y_i|X_i = x_0+, D_i = 1)$ is the conditional mean for a subset of units with $X_i^* \geq x_0$, unlike $E(Y_i|X_i = x_0+)$, it is not affected by the units with

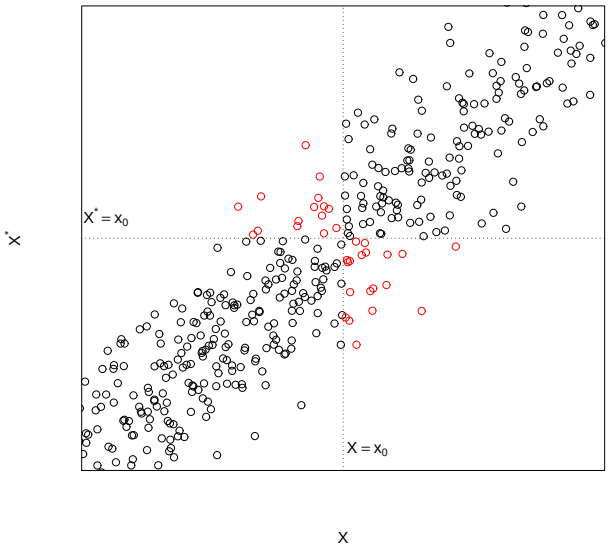


Figure 3.2: The dots are the pairs of (X_i, X_i^*) generated using the data-generating process in Monte Carlo simulations (Section 3.6). The horizontal and vertical axes are X_i and X_i^*, respectively. The dotted lines indicate the discontinuity point. $E(Y_i|X_i = x_0+)$ and $E(Y_i|X_i = x_0-)$ are based on the units in the right and left halves of the graph, respectively. By contrast, $E(Y_i|X_i = x_0+, D_i = 1)$ and $E(Y_i|X_i = x_0-, D_i = 0)$ are based on the units in the upper right and lower left, respectively. The units in the upper left and lower right (the red dots) severely affect identification for the average treatment effect.

$X_i \geq x_0$ and $X_i^* < x_0$. Similarly, $E(Y_i|X_i = x_0-, D_i = 0)$ is computed based on a subset of units with $X_i < x_0$ and $X_i^* < x_0$ (i.e., the lower left in Figure 3.2). Thus, $E(Y_i|X_i = x_0+, D_i = 1)$ and $E(Y_i|X_i = x_0-, D_i = 0)$ may not substantially differ from $E(Y_i|X_i^* = x_0+)$ and $E(Y_i|X_i^* = x_0-)$, respectively. Then, the identification bias of τ_{XD} may be smaller than that of τ_X for τ^*.

Nonetheless, both τ_X and τ_{XD} have an identification bias for τ^* because of the measurement error. To evaluate the identification biases, we introduce the following assumption, which ensures the existence of limits and the use of the dominated convergence theorem. Let $m(x^*) \coloneqq E(Y_{0i}|X_i^* = x^*) + \mathbf{1}(x^* \geq x_0)(E(Y_{1i} - Y_{0i}|X_i^* = x^*) - \tau^*)$.

Assumption 3.2. *(i)* $E(Y_i|X_i^* = x_0+)$, $E(Y_i|X_i^* = x_0-)$, $E(Y_i|X_i = x_0+)$, $E(Y_i|X_i = x_0-)$, $E(Y_i|X_i = x_0+, D_i = 1)$, *and* $E(Y_i|X_i = x_0-, D_i = 0)$ *exist. (ii)* $f_{X^*|X}(x^*|x_0+)$ *and* $f_{X^*|X}(x^*|x_0-)$ *exist for any* $x^* \in \mathbb{R}$. *(iii)* $1 - F_{X^*|X}(x_0|x_0+)$ *and* $F_{X^*|X}(x_0|x_0-)$ *exist and are non-zero. (iv)* $E(Y_{1i}|X_i^* = x^*)f_{X^*|X}(x^*|x)$, $E(Y_{0i}|X_i^* = x^*)f_{X^*|X}(x^*|x)$, *and* $m(x^*)f_{X^*|X}(x^*|x)$ *are dominated by some integrable functions in* $x^* \in \mathbb{R}$ *for* x *near* x_0.

The following theorem shows the specific form of the identification biases of τ_X and τ_{XD} for τ^*. The result (ii) in the theorem is shown in Yu (2012).

Theorem 3.1. *Suppose that Assumptions 3.1 and 3.2 hold and* $\sigma > 0$.

(i) It holds that

$$\begin{aligned}\tau_X = &\int_{x_0}^{\infty} E(Y_{1i}|X_i^* = x^*)\left(f_{X^*|X}(x^*|x_0+) - f_{X^*|X}(x^*|x_0-)\right)dx^* \\ &+ \int_{-\infty}^{x_0} E(Y_{0i}|X_i^* = x^*)\left(f_{X^*|X}(x^*|x_0+) - f_{X^*|X}(x^*|x_0-)\right)dx^* \\ &= 0.\end{aligned} \tag{3.6}$$

(ii) (Yu, 2012) It holds that

$$\tau_{XD} = \tau^* + \left(\frac{\int_{x_0}^{\infty} m(x^*)f_{X^*|X}(x^*|x_0+)dx^*}{1 - F_{X^*|X}(x_0|x_0+)} - \frac{\int_{-\infty}^{x_0} m(x^*)f_{X^*|X}(x^*|x_0-)dx^*}{F_{X^*|X}(x_0|x_0-)}\right). \tag{3.7}$$

The theorem demonstrates that τ_X and τ_{XD} have identification biases for τ^* and the average treatment effect at the threshold. As a result, we cannot precisely evaluate the causal effect of interest based on τ_X or τ_{XD}.

Importantly, the right-hand side of (3.6) vanishes because of the continuity of $f_{X^*|X}(x^*|x)$ at $x = x_0$, which is implied by the continuity of $f_U(\cdot)$ as shown in the proof of the theorem. In other words, τ_X becomes zero because of the effect of the continuous measurement error. This result is remarkable, because this implies that even if there is a substantial causal effect, τ_X misleads researchers into the incorrect conclusion in which there is no causal effect. Indeed, our Monte Carlo simulations demonstrate that an estimator for τ_X is significantly close to zero.

The second term on the right-hand side of (3.7) is the identification bias of τ_{XD} for τ^*. The identification bias does not vanish in general. For example, this identification bias does not vanish even when the treatment effect is constant, that is, when $Y_{1i} - Y_{0i} = b$ for any i and a constant b. In addition, the bias cannot be nonparametrically identified, because it relates the joint distribution of $(Y_{1i}, Y_{0i}, X_i^*, X_i)$ including unobservables (Y_{1i}, Y_{0i}, X_i^*).

Remark 3.3. Battistin et al. (2009) show that in the fuzzy RD design with a mismeasured forcing variable, the average treatment effect for the treated is identified using the Wald estimand:

$$\frac{E(Y_i|X_i = x_0+) - E(Y_i|X_i = x_0-)}{E(D_i|X_i = x_0+) - E(D_i|X_i = x_0-)}. \tag{3.8}$$

under the non-differential measurement error assumption and certain smoothness conditions on the joint distribution of (X_i^*, X_i). However, in a sharp RD design with continuous measurement error, their analysis could not work. To understand this, we observe the conditional mean of D_i (without the classical assumption):

$$E(D_i|X_i = x) = E(\mathbf{1}(X_i^* \geq x_0)|X_i = x) = \int_{x_0}^{\infty} f_{X^*|X}(x^*|x)dx^*.$$

Hence, the difference in the conditional means is

$$E(D_i|X_i = x_0+) - E(D_i|X_i = x_0-) = \int_{x_0}^{\infty}\left(f_{X^*|X}(x^*|x_0+) - f_{X^*|X}(x^*|x_0-)\right)dx^*.$$

This difference vanishes under the continuity of $f_{X^*|X}(x^*|x)$ at $x = x_0$. That is, the discontinuity of the conditional probability is smoothed out because of the continuous measurement error: $E(D_i|X_i = x)$ is not discontinuous at x_0 despite the discontinuity of $E(D_i|X_i^* = x)$. As a result, we cannot identify the average treatment effect using (3.8) under the continuous measurement error. Indeed, our Monte Carlo simulations demonstrate this problem. Hence, we do not recommend focusing on (3.8) in empirical situations in which the researchers are aware of the measurement error and in which they are confident of the discontinuous rule but $E(D_i|X_i)$ is not apparently discontinuous at the threshold.

Remark 3.4. Yu (2012) focuses on a local polynomial estimator for τ_{XD} to study the average treatment effect at the threshold in a sharp RD design with a continuous forcing variable that contains classical measurement error. He first shows that τ_{XD} has identification bias for τ^* that is identical to equation (3.7). He then shows that the local polynomial estimator for τ_{XD} is consistent for τ^* if the measurement error tends to zero depending on the sample size under some rate conditions. We stress that the approaches in his paper and the present paper differ, as we state in the Introduction.

There may be at least three possible approaches to examine τ^* using the mismeasured forcing variable following the literature of measurement error (see Schennach, 2013). First, we could use parameteric specifications, as in Hullegie and Klein (2010). By correcting the identification bias based on the parametric specifications, we can identify τ^*. However, this approach is sensitive to the validity of the parametric specifications: if the parametric specifications are invalid, the identification analysis can be broken.

Second, we may be able to identify τ^* using instrumental variables or repeated measurements. It is well-known in the literature of measurement error that such additional variables are powerful tools for establishing identification with the problems of measurement error. However, valid instrumental variables or repeated measurements are not commonly available in empirical situations based on the RD designs.

Third, we can learn the average treatment effect through approximation methods. The present study uses this approach because the advantages of the approximation approach correspond to those of the RD design: they do not require parametric specifications or additional variables such as instrumental variables. While the approximation approach may not provide the exact identification for the average treatment effect, it provides meaningful information without restrictive requirements.

3.4 The small error variance approximation in the RD design

This section develops an approximation analysis for the average treatment effect at the threshold based on the small error variance approximation (SEVA) originally proposed by Chesher (1991). We show that the average treatment effect is approximated using an identified parameter when the standard deviation of the measurement error σ is small.

We focus on approximating $\tau^* = E(Y_i|X_i^* = x_0+, D_i = 1) - E(Y_i|X_i^* = x_0-, D_i = 0)$ to learn the average treatment effect. In principle, we can also consider approximating $E(Y_i|X_i^* = x_0+) - E(Y_i|X_i^* = x_0-)$ based on the SEVA, although the precision of this approximation is worse. This notion comes from the same reason discussed in the previous section, that is, that $E(Y_i|X_i = x_0+) - E(Y_i|X_i = x_0-)$ vanishes because of continuous measurement error.

Before we state our formal result, we outline our approach for approximating τ^*. Extending the result in Chesher (1991), we show that the conditional density of Y_i given X_i^* and D_i is approximated as follows:

$$f_{Y|X^*D}(y|x,d) = f_{Y|XD}(y|x,d) - \sigma^2 \left(\log^{(1)} f_{X|D}(x|d)\right) f^{(1)}_{Y|XD}(y|x,d) - \frac{\sigma^2}{2} f^{(2)}_{Y|XD}(y|x,d) + o(\sigma^2).$$

for x near x_0, $y \in supp(Y)$, and $d \in \{0,1\}$.[3] The equation shows that the left-hand side is approximated up to the order $O(\sigma^2)$ by the terms on the right-hand side. Because the terms on the right-hand side relate the joint distribution of (Y_i, D_i, X_i) and σ, they are identified by the observable data if we extrapolate or forecast σ. Furthermore, this equation leads to

$$\begin{aligned} E(Y_i|X_i^* = x, D_i = d) =& E(Y_i|X_i = x, D_i = d) - \sigma^2 \left(\log^{(1)} f_{X|D}(x|d)\right) E^{(1)}(Y_i|X_i = x, D_i = d) \\ & - \frac{\sigma^2}{2} E^{(2)}(Y_i|X_i = x, D_i = d) + o(\sigma^2). \end{aligned}$$

The terms on the right-hand side are identified using the data. This indicates that the conditional mean of Y_i given X_i^* and D_i is approximated up to the order $O(\sigma^2)$ by the identified parameter. Accordingly, using this approximation, τ^* and the average treatment effect can be approximated up to the order $O(\sigma^2)$.

To show the approximation result in a rigorous manner, we require additional assumptions.

Assumption 3.3. *$E(Y_i|X_i^* = x_0+, D_i = 1)$, $E(Y_i|X_i^* = x_0-, D_i = 0)$, $E^{(s)}(Y_i|X_i = x_0+, D_i = 1)$, and $E^{(s)}(Y_i|X_i = x_0-, D_i = 0)$ exist for $s = 0, 1, 2$.*

Assumption 3.4. *(i) $f^{(s)}_{Y|X^*D}(y|x,d)$ is bounded in $y \in supp(Y)$ for x near x_0, $d \in \{0,1\}$, and $s = 0, 1, \ldots, 5$. (ii) $f^{(s)}_{X^*|D}(x|d)$ is bounded in x near x_0 for $d \in \{0,1\}$ and $s = 0, 1, \ldots, 5$. (iii) $f^{(s)}_{X|D}(x|d)$ is continuous at $x = x_0$ and bounded near x_0 for $d \in \{0,1\}$ and $s = 0, 1$.*

Assumption 3.5. *$\int_{-\infty}^{\infty} y f^{(s)}_{Y|XD}(y|x,d)dy < \infty$ and $\int_{-\infty}^{\infty} y f^{(t)}_{Y|X^*D}(y|x,d)dy < \infty$ for x near x_0, $d \in \{0,1\}$, $s = 0, 1, 2$, and $t = 0, 1, \ldots, 5$.*

The assumptions are regularity conditions for establishing the approximation for τ^*. Assumptions 3.3–3.5 ensure the existence of the limits, the boundedness of the densities, and the switching of the orders of integration and differentiation. The assumptions guarantee that the order of the approximation becomes $O(\sigma^2)$. We stress that we do not require partial differentiability of the conditional density of Y_i or X_i^* at x_0. Furthermore, the continuity of $f_{X|D}(\cdot|d)$ is implied by the continuity of $f_U(\cdot)$.

[3] As noted in Chesher (1991), while we can allow Y_i to be discrete, X_i^* and X_i must be continuous to establish the approximation. Nonetheless, to avoid complexity, we implicitly assume that Y_i is also continuous in this section.

The following theorem presents an approximation for τ^* based on the SEVA. Let

$$\begin{aligned}\mu(x,d,\sigma) \coloneqq\ & E(Y_i|X_i=x,D_i=d) \\ & -\sigma^2 g(x,d)E^{(1)}(Y_i|X_i=x,D_i=d)-\frac{\sigma^2}{2}E^{(2)}(Y_i|X_i=x,D_i=d),\end{aligned}$$

where $g(x,d) \coloneqq \log^{(1)} f_{X|D}(x|d) = f^{(1)}_{X|D}(x|d)/f_{X|D}(x|d)$ and $d \in \{0,1\}$.

Theorem 3.2. *Suppose that Assumptions 3.1 and 3.3–3.5 hold. When $\sigma \to 0$, it holds that*

$$\tau^* = \mu(x_0+,1,\sigma) - \mu(x_0-,0,\sigma) + o(\sigma^2). \tag{3.9}$$

Theorem 3.2 states that τ^* (and thus the average treatment effect) are approximated up to the order $O(\sigma^2)$ by the difference on the right-hand side when σ is small. The smaller standard deviation of the measurement error implies a more precise approximation for the average treatment effect.

The condition $\sigma \to 0$ means that σ is "sufficiently small" in the mathematical sense. While the original SEVA in Chesher (1991) does not require this condition, we need it for the SEVA in the RD design. The reason why we require the condition (and the bounded support of U_i) is because we approximate the one-sided limits of the conditional expectations. The basic idea behind the SEVA is the convolution of the probability distributions and Taylor's theorem. In our setting, we should apply Taylor's theorem to the conditional densities at every point near the discontinuity point to approximate the one-sided limits of the conditional means. As a result, the condition $\sigma \to 0$ and the bounded support are required to ensure the establishment of the Taylor polynomials at every point near the discontinuity point. However, the condition $\sigma \to 0$ is a mathematical requirement, which does not mean that σ converges in the real world. In practice, the precision of the approximation depends on the magnitude of σ and the data-generating process. We demonstrate the approximate precision using our Monte Carlo simulations, which suggests that our approximation can work even when σ^2 accounts for about 20% of $var(X_i)$.

The terms on the right-hand side of (3.9) are identified by the data (Y_i, D_i, X_i) except for the standard deviation σ. In practice, we can extrapolate σ when we have additional public data or census data on X^* for the population of interest, because $\sigma^2 = var(X) - var(X^*)$ under Assumption 3.1. For example, suppose that we are interested in evaluating a policy program in a state based on survey data and the forcing variable is income, which may contain measurement error. In this situation, we can use public data on income in the state to extrapolate σ by estimating $var(X^*)$. Importantly, this procedure does not require additional variables on the observations $i = 1, \ldots, n$. Alternatively, we can learn the effect of the measurement error through (3.9) by forecasting σ, as in Battistin and Chesher (2014). Because the difference on the right-hand side of (3.9) is monotonic in σ, we can calculate the forecast intervals for the difference by forecasting several values of σ. Hence, Theorem 3.2 allows us to correct or forecast the identification bias because of the measurement error up to the order $O(\sigma^2)$.

Remark 3.5. Our approximate analysis could not be extended to the fuzzy RD design, because $E(D_i|X_i)$ in the fuzzy RD design is not discontinuous at the threshold owing to continuous measurement error. We consider the same setting in Section 3.2, except that D_i is not a deterministic function of X_i^*. Instead, in the fuzzy RD design, the conditional probability of $D_i = 1$ given X_i^* is discontinuous at the threshold: $E(D_i|X_i^* = x_0+) \neq E(D_i|X_i^* = x_0-)$. The fuzzy RD estimand is

$$\frac{E(Y_i|X_i^* = x_0+) - E(Y_i|X_i^* = x_0-)}{E(D_i|X_i^* = x_0+) - E(D_i|X_i^* = x_0-)},$$

which identifies the average treatment effect under the independence assumption between $Y_{1i} - Y_{0i}$ and D_i. Even without the independence assumption, the parameter identifies a local average treatment effect (see Hahn et al. (2001) for details). Here, $E(Y_i|X_i^*) \neq E(Y_i|X_i^*, D_i)$, because D_i is not a deterministic function of X_i^* in the fuzzy RD design.

It might seem that the fuzzy RD estimand is approximated by approximating each term in the estimand based on the SEVA, similar to the sharp RD design. That is, it might seem that by extending the result in Chesher (1991), the average treatment effect in the fuzzy RD design is approximated by

$$\frac{\mu_Y(x_0+, \sigma) - \mu_Y(x_0-, \sigma)}{\mu_D(x_0+, \sigma) - \mu_D(x_0-, \sigma)},$$

where

$$\mu_Y(x, \sigma) \coloneqq E(Y_i|X_i = x) - \sigma^2 \left(\log^{(1)} f_X(x)\right) E^{(1)}(Y_i|X_i = x) - \frac{\sigma^2}{2} E^{(2)}(Y_i|X_i = x),$$
$$\mu_D(x, \sigma) \coloneqq E(D_i|X_i = x) - \sigma^2 \left(\log^{(1)} f_X(x)\right) E^{(1)}(D_i|X_i = x) - \frac{\sigma^2}{2} E^{(2)}(D_i|X_i = x).$$

However, $\mu_D(x_0+, \sigma) - \mu_D(x_0-, \sigma)$ could vanish under continuous measurement error. Under the classical measurement error assumption, we observe that

$$E(D_i|X_i = x) = E(E(D_i|X_i^*, X_i = x)|X_i = x) = \int_{-\infty}^{\infty} E(D_i|X_i^* = x^*) f_{X^*|X}(x^*|x) dx^*.$$

Hence, the difference in the conditional probabilities is

$$E(D_i|X_i = x_0+) - E(D_i|X_i = x_0-) = \int_{-\infty}^{\infty} E(D_i|X_i^* = x^*)(f_{X^*|X}(x^*|x_0+) - f_{X^*|X}(x^*|x_0-)) dx^*.$$

This difference vanishes under the continuity of $f_{X^*|X}(x^*|x)$ at $x = x_0$, which is implied by the continuity of $f_U(\cdot)$. Hence, $\mu_D(x_0+, \sigma) - \mu_D(x_0-, \sigma)$ could also vanish such that we cannot approximate the fuzzy RD estimand based on the SEVA.

Accordingly, we require other approaches to evaluate the effect of measurement error in the fuzzy RD design. This is beyond the scope of this paper.

3.5 Estimation

This section presents a nonparametric estimation procedure for the parameter that approximates the average treatment effect, that is, the difference on the right-hand side of (3.9). We develop the asymptotic properties of the nonparametric estimator.

Using the approximation analysis developed in the previous section, if we can consistently estimate the difference on the right-hand side of (3.9), the average treatment effect is approximately estimated up to the order $O(\sigma^2)$. We thus consider estimating

$$\begin{aligned}
\mu(x_0+,1,\sigma) &= E(Y_i|X_i = x_0+, D_i = 1) \\
&\quad - \sigma^2 g(x_0,1)E^{(1)}(Y_i|X_i = x_0+, D_i = 1) - \frac{\sigma^2}{2}E^{(2)}(Y_i|X_i = x_0+, D_i = 1), \\
\mu(x_0-,0,\sigma) &= E(Y_i|X_i = x_0-, D_i = 0) \\
&\quad - \sigma^2 g(x_0,0)E^{(1)}(Y_i|X_i = x_0-, D_i = 0) - \frac{\sigma^2}{2}E^{(2)}(Y_i|X_i = x_0-, D_i = 0),
\end{aligned}$$

where $g(x,d) \coloneqq f^{(1)}_{X|D}(x|d)/f_{X|D}(x|d)$. In the following, we assume that σ^2 is known, because σ^2 can be extrapolated through $\sigma^2 = var(X) - var(X^*)$ or forecast.

We can consistently estimate $g(x_0,d)$ by $\hat{g}(x_0,d)$ based on the kernel density and density derivative estimators:

$$\hat{g}(x_0,d) \coloneqq \frac{\hat{f}^{(1)}_{X|D}(x_0|d)}{\hat{f}_{X|D}(x_0|d)},$$

where $\hat{f}_{X|D}(x_0|d) \coloneqq (n_d h)^{-1}\sum_{i=1}^n \mathbf{1}(D_i = d)K_h(X_i)$, $\hat{f}^{(1)}_{X|D}(x_0|d) \coloneqq -(n_d h^2)^{-1}\sum_{i=1}^n \mathbf{1}(D_i = d)K^{(1)}_h(X_i)$, n_d is the number of observations with $D_i = d$, $K_h(z) \coloneqq K((z-x_0)/h)$, $K(\cdot)$ is some kernel function, and h is a bandwidth tending to zero as $n \to \infty$. $\hat{f}_{X|D}(x_0|d)$ and $\hat{f}^{(1)}_{X|D}(x_0|d)$ are consistent for $f_{X|D}(x_0|d)$ and $f^{(1)}_{X|D}(x_0|d)$, respectively, under the regularity conditions, similar to those in Silverman (1978) or Li and Racine (2007, Chapter 3) (see also Fan and Gijbels, 1996, Section 2.7). Accordingly, $\hat{g}(x_0,d)$ is consistent for $g(x_0,d)$ under the conditions. We thus assume the conditions implicitly and omit the details for the asymptotic properties of this estimator.

We next focus on estimating $E^{(s)}(Y_i|X_i = x_0+, D_i = 1)$ and $E^{(s)}(Y_i|X_i = x_0-, D_i = 0)$ for $s = 0,1,2$, which are estimated using local polynomial regressions (Fan and Gijbels, 1996). The estimators for $E^{(s)}(Y_i|X_i = x_0+, D_i = 1)$ are given by the following p-th order local polynomial regression:

$$(\hat{\alpha}^+, \hat{\beta}^{+\prime})' \coloneqq \underset{(a,b')'\in\mathbb{R}^{p+1}}{\operatorname{argmin}} \sum_{i=1}^n I_i D_i \left(Y_i - a - b_1(X_i - x_0) - \cdots - b_p(X_i - x_0)^p\right)^2 K_h(X_i), \quad (3.10)$$

where $p \geq 2$ is some positive integer and $I_i \coloneqq \mathbf{1}(X_i \geq x_0)$.[4] Then, $\hat{\alpha}^+$ is the estimator for $E(Y_i|X_i = x_0+, D_i = 1)$, and $\hat{\beta}^+_k$ is that for $(k!)^{-1}E^{(k)}(Y_i|X_i = x_0+, D_i = 1)$ for $k = 1,\ldots,p$. Similarly, the estimators for $E^{(s)}(Y_i|X_i = x_0-, D_i = 0)$ for $s = 0,1,2$ are given by the following p-th order local polynomial regression:

$$(\hat{\alpha}^-, \hat{\beta}^{-\prime})' \coloneqq \underset{(a,b')'\in\mathbb{R}^{p+1}}{\operatorname{argmin}} \sum_{i=1}^n (1-I_i)(1-D_i) \left(Y_i - a - b_1(X_i - x_0) - \cdots - b_p(X_i - x_0)^p\right)^2 K_h(X_i),$$

[4] In practice, the kernel function and bandwidth used for the local polynomial regressions can differ from those used to estimate $f_{X|D}(x_0|d)$ and $f^{(1)}_{X|D}(x_0|d)$.

Then, $\hat{\alpha}^-$ is the estimator for $E(Y_i|X_i = x_0-, D_i = 0)$, and $\hat{\beta}_k^-$ is that for $(k!)^{-1}E^{(k)}(Y_i|X_i = x_0-, D_i = 0)$ for $k = 1, \ldots, p$.

The parameter approximating the average treatment effect is estimated by

$$\hat{\mu}(x_0+, 1, \sigma) - \hat{\mu}(x_0-, 0, \sigma),$$

where

$$\hat{\mu}(x_0+, 1, \sigma) \coloneqq \hat{\alpha}^+ - \sigma^2 \hat{g}(x_0, 1)\hat{\beta}_1^+ - \sigma^2 \hat{\beta}_2^+,$$
$$\hat{\mu}(x_0-, 0, \sigma) \coloneqq \hat{\alpha}^- - \sigma^2 \hat{g}(x_0, 0)\hat{\beta}_1^- - \sigma^2 \hat{\beta}_2^-,$$

which are estimators for $\mu(x_0+, 1, \sigma)$ and $\mu(x_0-, 0, \sigma)$, respectively.

To develop the asymptotic properties of this estimator, we introduce additional assumptions. The assumptions are standard regularity conditions for developing asymptotic properties for the local polynomial estimators, which are analogous to the conditions in Hahn et al. (2001) and Porter (2003). Let $V_i \coloneqq Y_i - E(Y_i|X_i, D_i)$.

Assumption 3.6. *$K(\cdot)$ is continuous, symmetric, and non-negative with compact support. For simplicity, the support is assumed to be $[-M, M]$ for some finite $M > 0$.*

Assumption 3.7. *$f_X(\cdot)$ is bounded, continuous, and bounded away from zero near x_0.*

Assumption 3.8. *$E(D_i|X_i = x_0+)$ and $E(1 - D_i|X_i = x_0-)$ exist and are non-zero.*

Assumption 3.9. *(i) $E(V_i^2|X_i = x, D_i = 1)$ and $E(V_i^2|X_i = x, D_i = 0)$ are bounded near x_0 and $E(V_i^2|X_i = x_0+, D_i = 1)$ and $E(V_i^2|X_i = x_0-, D_i = 0)$ exist. (ii) $E(|V_i|^{2+\zeta}|X_i = x)$ is bounded near x_0 for some $\zeta > 0$.*

Assumption 3.10. *(i) $E(Y_i|X_i = x, D_i = d)$ is $p+1$-times continuously differentiable for x near x_0 and $d \in \{0, 1\}$. (ii) $E^{(k)}(Y_i|X_i = x_0+, D_i = 1)$ and $E^{(k)}(Y_i|X_i = x_0-, D_i = 0)$ exist for $k = 1, \ldots, p+1$. (iii) There exists some $\tilde{M} > 0$ such that $E^{(p+1)}(Y_i|X_i = x, D_i = 1)$ is bounded for $x \in [x_0, x_0 + \tilde{M}]$ and $E^{(p+1)}(Y_i|X_i = x, D_i = 0)$ is bounded for $x \in [x_0 - \tilde{M}, x_0)$.*

To develop asymptotic properties of $\hat{\mu}(x_0+, 1, \sigma) - \hat{\mu}(x_0-, 0, \sigma)$, we first study asymptotic properties of the local polynomial estimators.

Lemma 3.1. *Suppose that Assumptions 3.6–3.10 hold. When $n \to \infty$, $h \to 0$, $nh \to \infty$, and $\sqrt{nh}h^{p+1} \to \tilde{C}$ for some $\tilde{C} \in [0, \infty)$, it holds that*

$$\left(\hat{\alpha}^+, \hat{\beta}^{+\prime}\right)' - \left(\alpha^+, \beta^{+\prime}\right)' \xrightarrow{p} 0,$$
$$\left(\hat{\alpha}^-, \hat{\beta}^{-\prime}\right)' - \left(\alpha^-, \beta^{-\prime}\right)' \xrightarrow{p} 0,$$

and

$$\sqrt{nh}H^{-1}\left(\left(\hat{\alpha}^+, \hat{\beta}^{+\prime}\right)' - \left(\alpha^+, \beta^{+\prime}\right)'\right) \rightsquigarrow N\left(B^+, \Omega^+\right), \tag{3.11}$$
$$\sqrt{nh}H^{-1}\left(\left(\hat{\alpha}^-, \hat{\beta}^{-\prime}\right)' - \left(\alpha^-, \beta^{-\prime}\right)'\right) \rightsquigarrow N\left(B^-, \Omega^-\right), \tag{3.12}$$

where

$$
\begin{aligned}
H &:= diag(1, h^{-1}, \ldots, h^{-p}), \\
B^+ &:= \tilde{C} E^{(p+1)}(Y_i | X_i = x_0+, D_i = 1)(\Gamma^+)^{-1}(\gamma_{p+1}, \ldots, \gamma_{2p+1})', \\
B^- &:= \tilde{C} E^{(p+1)}(Y_i | X_i = x_0-, D_i = 0)(\Gamma^-)^{-1}\left((-1)^{p+1}\gamma_{p+1}, \ldots, (-1)^{2p+1}\gamma_{2p+1}\right)', \\
\Omega^+ &:= \frac{E(V_i^2 | X_i = x_0+, D_i = 1)}{E(D_i | X_i = x_0+) f_X(x_0)}(\Gamma^+)^{-1}\Delta^+(\Gamma^+)^{-1}, \\
\Omega^- &:= \frac{E(V_i^2 | X_i = x_0-, D_i = 0)}{E(1 - D_i | X_i = x_0-) f_X(x_0)}(\Gamma^-)^{-1}\Delta^-(\Gamma^-)^{-1}, \\
\Gamma^+ &:= (\gamma_{k+l-2})_{(k,l)}, \qquad \Gamma^- := \left((-1)^{k+l+1}\gamma_{k+l-2}\right)_{(k,l)}, \qquad \text{for} \quad k, l = 1, \ldots, p+1, \\
\Delta^+ &:= (\delta_{k+l-2})_{(k,l)}, \qquad \Delta^- := \left((-1)^{k+l+1}\delta_{k+l-2}\right)_{(k,l)}, \qquad \text{for} \quad k, l = 1, \ldots, p+1, \\
\gamma_q &:= \int_0^M u^q K(u) du, \qquad \delta_q := \int_0^M u^q K^2(u) du, \qquad \text{for} \quad q = 0, \ldots, 2p+1.
\end{aligned}
$$

Lemma 3.1 shows that the vectors of local polynomial estimators are consistent and asymptotically normal. The asymptotic distributions are not centered at zero because of the presence of the asymptotic biases. The asymptotic biases of the vectors of the local polynomial estimators are of order $O(Hh^{p+1})$ and depend on the one-sided derivatives of $E(Y_i | X_i, D_i)$.

Lemma 3.1 states that the convergence rates of the vectors of the local polynomial estimators are $1/(\sqrt{nh}H^{-1})$. Specifically, the convergence rates of $\hat{\alpha}^+$ and $\hat{\alpha}^-$ are of order $1/\sqrt{nh}$, those of $\hat{\beta}_1^+$ and $\hat{\beta}_1^-$ are of order $1/(\sqrt{nh}h)$, and those of $\hat{\beta}_2^+$ and $\hat{\beta}_2^-$ are of order $1/(\sqrt{nh}h^2)$. These results are consistent with the results in the literature of local polynomial regressions, such as those in Fan and Gijbels (1992), Ruppert and Wand (1994), and Masry (1996a,b). From these results, we expect the convergence rate of $\hat{\mu}(x_0+, 1, \sigma) - \hat{\mu}(x_0-, 0, \sigma)$ to be of order $1/(\sqrt{nh}h^2)$.

The asymptotic properties of $\hat{\mu}(x_0+, 1, \sigma) - \hat{\mu}(x_0-, 0, \sigma)$ are developed in the following theorem.

Theorem 3.3. *Suppose that Assumptions 3.6–3.10 hold and* $\hat{g}(x_0, d) \xrightarrow{p} g(x_0, d)$ *for* $d \in \{0, 1\}$. *When* $n \to \infty$, $h \to 0$, $nh \to \infty$, *and* $\sqrt{nh}h^{p+1} \to \tilde{C}$ *for some* $\tilde{C} \in [0, \infty)$, *it holds that*

$$
\hat{\mu}(x_0+, 1, \sigma) - \hat{\mu}(x_0-, 0, \sigma) - (\mu(x_0+, 1, \sigma) - \mu(x_0-, 0, \sigma)) \xrightarrow{p} 0,
$$

and

$$
\sqrt{nh}h^2 (\hat{\mu}(x_0+, 1, \sigma) - \hat{\mu}(x_0-, 0, \sigma) - (\mu(x_0+, 1, \sigma) - \mu(x_0-, 0, \sigma))) \rightsquigarrow N(B, \Omega),
$$

where $B := \sigma^2 e_3'(B^- - B^+)$, $\Omega := \sigma^4 e_3'(\Omega^+ + \Omega^-) e_3$, *and* $e_3 := (0, 0, 1, 0, \ldots, 0)'$ *is the* $p+1$ *vector and* B^+, B^-, Ω^+, *and* Ω^- *are defined in Lemma 3.1.*

Theorem 3.3 shows that the estimator for the parameter approximating the average treatment effect is consistent for the parameter and asymptotically normal. The asymptotic distribution is not centered at zero because of the presence of the asymptotic bias. The convergence speed of the estimator is of order $1/(\sqrt{nh}h^2)$, which is expected by the discussion

above: the convergence rate of the estimator is determined by those of the estimators for $E^{(2)}(Y_i|X_i = x_0+, D_i = 1)$ and $E^{(2)}(Y_i|X_i = x_0-, D_i = 0)$. This convergence rate is slower than $1/\sqrt{nh}$, although this result is standard. This convergence speed is a limitation of the local polynomial estimators, which cannot be overcome if we use the local polynomial regressions. We can overcome the slow convergence rate using parametric methods such as parametric polynomial regressions, although we do not pursue this issue here because this paper focuses on nonparametric methods.

Remark 3.6. The selection of the kernel functions and bandwidths are practically concerned. In particular, the precision of the nonparametric estimator $\hat{\mu}(x_0+, 1, \sigma) - \hat{\mu}(x_0-, 0, \sigma)$ largely depends on selecting the bandwidths, as other nonparametric estimators do. We explain the details of bandwidth selection in our Monte Carlo simulations (Section 3.6). We note that the bandwidth selection for RD designs developed in Imbens and Kalyanaraman (2012) and Arai and Ichimura (2014) cannot directly apply our setting. This is because our estimand, the difference on the right-hand side of (3.9), is not typical in the sharp RD design.

3.6 Monte Carlo simulations

This section presents the results of the Monte Carlo simulations. We first describe the simulation designs and the implementation of our approximate analysis, and then we report the results.

The simulations are conducted with R 3.1.1 for Windows 7. 1000 replications are used for the simulation. We set the sample size to $n = 2500$, which may look somewhat large, although this is required to execute higher-order local polynomial regressions.

3.6.1 Designs

We consider two designs for the potential outcomes:

Design A:
$$Y_{1i} = 1.52 + 0.84X_i^* - 3.0(X_i^*)^2 + 7.99(X_i^*)^3 - 9.01(X_i^*)^4 + 3.56(X_i^*)^5 + e_i,$$
$$Y_{0i} = 0.48 + 1.27X_i^* + 7.18(X_i^*)^2 + 20.21(X_i^*)^3 + 21.54(X_i^*)^4 + 7.33(X_i^*)^5 + e_i,$$

Design B:
$$Y_{1i} = 0.5 + 0.84X_i^* - 0.3(X_i^*)^2 - 2.397(X_i^*)^3 - 0.901(X_i^*)^4 + 3.56(X_i^*)^5 + e_i,$$
$$Y_{0i} = 0 + 1.27X_i^* - 3.59(X_i^*)^2 + 14.147(X_i^*)^3 + 23.694(X_i^*)^4 + 10.995(X_i^*)^5 + e_i,$$

where $X_i^* \sim i.i.d.\ 2Beta(2, 4) - 0.7$ and $e_i \sim i.i.d.N(0, 0.1295^2)$ in both designs, which implies $E(X_i^*) = -1/30$, $var(X_i^*) = 32/252$. The treatment is $D_i = \mathbf{1}(X_i^* \geq 0)$, that is, $x_0 = 0$, in each design. Design A is similar to Imbens and Kalyanaraman (2012, Lee design), Arai and Ichimura (2014, Design 1), and Calonico et al. (2014, Model 1), which is motivated by Lee (2008)'s data. Design B is analogous to Calonico et al. (2014, Model 3). However, the average treatment effects and $E(X_i^*)$ are bigger here, which reveal the riskiness of the mismeasured forcing variable. For illustration, the conditional means are plotted in Figure 3.3.

The observable mismeasured forcing variable is generated as $X_i = X_i^* + \sigma U_i$, where U_i is the i.i.d. truncated normally distributed random variable with mean 0 and standard deviation

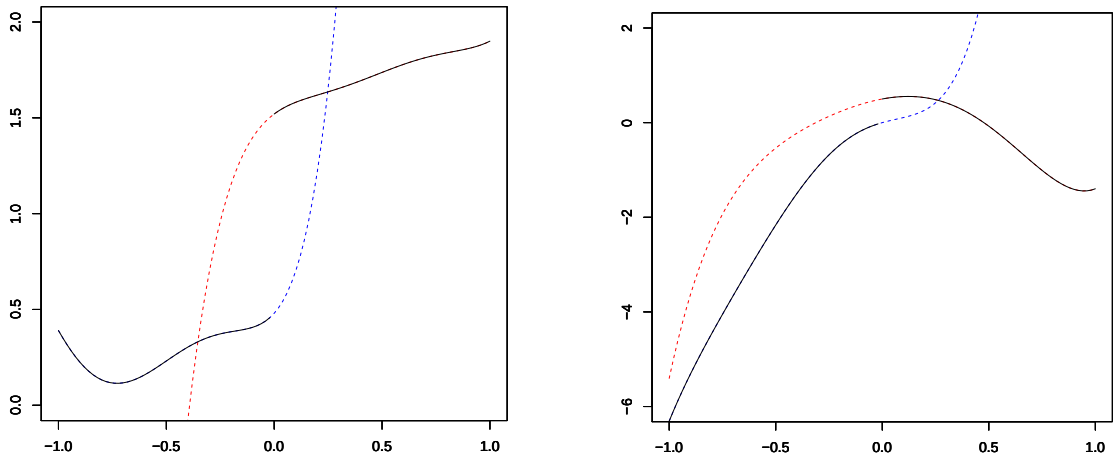
1, whose support is $[-3, 3]$. U_i is independent of the other variables. We consider three values for the standard deviation of the measurement error: $\sigma = 0.12, 0.15, 0.18$. Under each σ, the magnitudes of σ^2 account for about 10%, 15%, and 20% of $var(X_i)$, respectively. For illustration, the densities of X_i^* and X_i for each σ are plotted in Figure 3.4.

We evaluate the performance of four estimators. The first is the estimator based on the SEVA developed in Section 3.5 (we denote it as "ESEVA"): $\hat{\mu}(x_0+, 1, \sigma) - \hat{\mu}(x_0-, 0, \sigma)$. For the kernel density estimation, we use the Epanechnikov kernel function $K_1(u) = 3/4(1-u^2)\mathbf{1}(|u| \leq 1)$ and the normal-scale rule bandwidth given by $h_d = 2.34\hat{\sigma}_{X,d} n_d^{-1/5}$, where $\hat{\sigma}_{X,d}$ is the square root of the sample variance of X_i for observations with $D_i = d$ for $d \in \{0, 1\}$ and n_d is the number of observations with $D_i = d$. For the kernel density derivative estimation, we employ the kernel function proposed by Jones (1994), $K_2^{(1)}(u) = u(1-u)^2\mathbf{1}(|u| \leq 1)/4$ and the normal scale rule bandwidth $h_d = \hat{\sigma}_{X,d}(112\sqrt{\pi}/n_d)^{1/7}$. We use the local linear regression to estimate the one-sided limits of the conditional expectation and the second-order local polynomial regressions to estimate the first and second derivatives. We use separate-order local polynomial regressions because of the different convergence rates of the estimators and the automatic boundary adaptive property of the local polynomial regressions (Fan and Gijbels, 1996).[5] For all local polynomial regressions, we employ the triangle kernel function $K_3(u) = (1-|u|)\mathbf{1}(|u| \leq 1)$. The bandwidth for the p-th order local polynomial regression required to estimate the ν-th right derivative is selected by the plug-in method developed in Fan and Gijbels (1996, p.67), that is,

$$h_{\nu,p} = C_{\nu,p}(K) \left(\frac{\hat{\sigma}^2(x_0+)}{(\hat{E}^{(p+1)}(Y_i|X_i = x_0+, D_i = 1))^2 \hat{f}_X(x_0) \hat{E}(D_i|X_i = x_0+) n^+} \right)^{1/(2p+3)},$$

for $(\nu, p) = (0, 1), (1, 2), (2, 2)$. $\hat{\sigma}^2(x_0+)$ is the local linear estimator for the conditional variance of Y_i given $D_i = 1$ and $X_i = x_0+$. $\hat{f}_X(x_0)$ is the kernel density estimator for $f_X(x_0)$. $\hat{E}(D_i|X_i = x_0+)$ is the local linear estimator for $E(D_i|X_i = x_0+)$. n^+ is the number of observations with $D_i = 1$ and $X_i \geq x_0$. $C_{\nu,p}(K)$ is a constant depending on the kernel function whose definition is given in Fan and Gijbels (1996, p.67). For the triangle kernel, we set $C_{0,1}(K_3) = 2.9925$, $C_{1,2}(K_3) = 3.5218$, and $C_{2,2}(K_3) = 3.1077$. To select the bandwidth for the second order local polynomial regression, we need a pilot estimate for $\hat{E}^{(3)}(Y_i|X_i = x_0+, D_i = 1)$, which is estimated using the third-order local polynomial regression. Selecting the bandwidths for the left derivatives is analogous. The variance of the measurement error σ^2 is estimated through the difference in the sample variance of X_i and that of X^* using artificial additional data on X^* whose sample size is 5000.

[5] We can also use third-order polynomial regressions to estimate the second derivatives. However, we find that the performance of these estimators is worse than that of estimators based on the second-order polynomial regressions. This is because the plug-in bandwidth for the third-order polynomial regressions is of order $n^{-1/9}$, which leads to oversmoothing bandwidth under sample size 2500. We require a larger sample size to employ the plug-in bandwidth of order $n^{-1/9}$. Hence, we employ the second-order polynomial regressions, which lead to the plug-in bandwidth of order $n^{-1/7}$.



(a) Design A (b) Design B

Figure 3.3: The conditional means in each design are plotted. The dotted red and blue lines are $E(Y_{1i}|X_i^* = x)$ and $E(Y_{0i}|X_i^* = x)$, respectively. The solid black line is $E(Y_i|X_i^* = x)$.

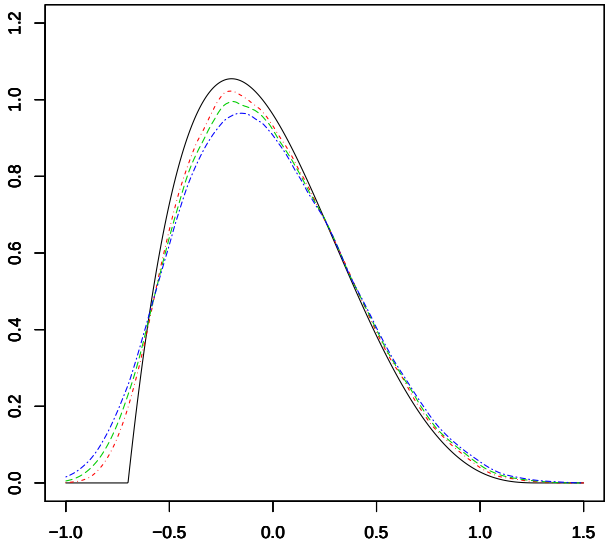


Figure 3.4: The densities of X_i^* and X_i for each σ are plotted. The black line is the density of X_i^*, the red one is that of X_i for $\sigma = 0.12$, the green one is that for $\sigma = 0.15$, and the blue one is that for $\sigma = 0.18$.

		ESEVA		$\hat{\tau}_X$		$\hat{\tau}_{XD}$		Fuzzy	
σ	true	bias	std	bias	std	bias	std	bias	std
0.12	1.04	0.0710	0.0343	-1.0083	0.1121	0.1028	0.0224	-0.0755	7.5655
0.15	1.04	0.0799	0.0324	-0.9984	0.1080	0.1196	0.0208	0.2164	4.1668
0.18	1.04	0.0920	0.0316	-1.0144	0.1044	0.1367	0.0205	0.8930	19.5753

Table 3.1: Monte Carlo simulation results with design A

		ESEVA		$\hat{\tau}_X$		$\hat{\tau}_{XD}$		Fuzzy	
σ	true	bias	std	bias	std	bias	std	bias	std
0.12	0.5	0.0853	0.0919	-0.4925	0.0789	0.2539	0.0497	0.4399	20.6762
0.15	0.5	0.1044	0.1031	-0.4916	0.0901	0.3336	0.0562	-1.5143	33.8111
0.18	0.5	0.1586	0.1087	-0.4915	0.0994	0.4236	0.0605	0.1016	8.9724

Table 3.2: Monte Carlo simulation results with design B

The second is the estimator for τ_X, $\hat{\tau}_X \coloneqq \tilde{\alpha}_Y^+ - \tilde{\alpha}_Y^-$, where

$$(\tilde{\alpha}_Y^+, \tilde{\beta}_Y^+)' \coloneqq \underset{(a,b)' \in \mathbb{R}^2}{\operatorname{argmin}} \sum_{i=1}^{n} I_i \left(Y_i - a - b_1(X_i - x_0)\right)^2 K\left(\frac{X_i - x_0}{h}\right),$$
$$(\tilde{\alpha}_Y^-, \tilde{\beta}_Y^-)' \coloneqq \underset{(a,b)' \in \mathbb{R}^2}{\operatorname{argmin}} \sum_{i=1}^{n} (1 - I_i) \left(Y_i - a - b_1(X_i - x_0)\right)^2 K\left(\frac{X_i - x_0}{h}\right).$$

That is, $\hat{\tau}_X$ is the estimator based on the local linear regressions. For the estimator, we employ the triangle kernel function and the plug-in bandwidth discussed in Fan and Gijbels (1996).

The third is the estimator for τ_{XD}, $\hat{\tau}_{XD} \coloneqq \hat{\alpha}^+ - \hat{\alpha}^-$, based on the local linear regressions developed in Lemma 3.1 (i.e., $p = 1$). For this estimator, we employ the triangle kernel function and the plug-in bandwidth discussed previously.

The fourth is the estimator in the fuzzy RD design discussed in Remark 3.3, that is, the estimator of (3.8) (we denote it as "Fuzzy"). Specifically, the estimator is $(\tilde{\alpha}_Y^+ - \tilde{\alpha}_Y^-)/(\tilde{\alpha}_D^+ - \tilde{\alpha}_D^-)$, where $\tilde{\alpha}_D^+$ and $\tilde{\alpha}_D^-$ are obtained by the minimization problems here in which Y_i is replaced with D_i. For the estimator, we employ the triangle kernel function and the plug-in bandwidth in Fan and Gijbels (1996).

3.6.2 Results

The results of the Monte Carlo simulations with designs A and B are reported in Tables 3.1 and 3.2, respectively. The column labeled "true" reports the true average treatment effect at the threshold. The bias for the average treatment effect and the standard deviation of each estimator are presented in the tables.

The simulation results demonstrate that the approximate analysis based on the SEVA is informative for learning the average treatment effect. In both designs, the biases of ESEVA are moderate for each σ. However, the biases of ESEVA with design B are somewhat larger than those with design A. Nonetheless, the biases of ESEVA are considerably smaller than those of the remaining estimators in all cases. The biases of ESEVA increase as σ increases, although

σ	mean	std
0.12	0.0276	0.0946
0.15	0.0372	0.0929
0.18	0.0231	0.0888

Table 3.3: Estimates of $E(D_i|X_i = x_0+) - E(D_i|X_i = x_0-)$

this is expected by our analysis developed in Section 3.4. Accordingly, the simulation results corroborate our approximate analysis based on the SEVA.

The standard deviations of ESEVA are moderate in both designs for all σ. However, as σ becomes larger with design B, the standard deviations increase. This is expected using our approximation analysis: when σ is large, the effects of the second and third terms in $\hat{\mu}(x_0+, 1, \sigma)$ and $\hat{\mu}(x_0-, 0, \sigma)$ on the standard deviation increase. Nonetheless, the standard deviations become smaller as n increases, because the asymptotic variance of ESEVA is of order $1/(\sqrt{nh}h^2)$.

The simulation results reveal that $\hat{\tau}_X$ has severe bias for identifying the average treatment effect. The bias of $\hat{\tau}_X$ is critical even for small σ. As expected by the analysis in Section 3.3, the estimates of $\hat{\tau}_X$ in both designs are close to 0, which leads us to the misleading consequence in which there is no treatment effect. This result indicates that $\hat{\tau}_X$ is not consistent for the average treatment effect because of the identification bias caused by the measurement error.

The biases of $\hat{\tau}_{XD}$ are relatively moderate compared with those of $\hat{\tau}_X$. However, the biases of $\hat{\tau}_{XD}$ are larger than those of ESEVA in all cases. In particular, the biases of $\hat{\tau}_{XD}$ with design B are about three times as large as those of ESEVA. Furthermore, the mean squared errors of $\hat{\tau}_{XD}$ are bigger than those of ESEVA in all cases. These results suggest that ESEVA functions better than $\hat{\tau}_{XD}$ in all cases.

The performance of Fuzzy is poor. The estimator is unstable and both the bias and the standard deviation are incoherent in each setting. This is because the measurement error causes the discontinuity of the conditional mean of D_i to vanish, as discussed in Remark 3.3. The mean and standard deviation of the estimates of the discontinuity of $E(D_i|X_i = x)$ at x_0 with design A are reported in Table 3.3,[6] which reveals that the discontinuity vanishes because of the measurement error even for small σ. According to the simulation results, we do not recommend using the fuzzy RD estimand in situations in which the discontinuity size of $E(D_i|X_i)$ is apparently small despite a confident discontinuous rule.

To summarize, the simulation results corroborate our theoretical analysis. The measurement error leads the difference in the mean outcomes just above and below the threshold and the discontinuous size of the conditional means of D_i to vanish. In addition, the approximate analysis based on the SEVA works more successfully than the remaining estimators: the biases and the mean squared errors of ESEVA are smaller than those of $\hat{\tau}_X$, $\hat{\tau}_{XD}$, and Fuzzy.

[6] Because the data-generating processes of (D_i, X_i^*, X_i) are the same in each design, we have similar results for the estimates of the discontinuity with each design. We thus report the results only with design A.

3.7 Conclusion

This paper presents a nonparametric analysis in the sharp RD design in which the forcing variable contains measurement error. We show that the average treatment effect given the "true" forcing variable at the discontinuity point cannot be identified based on the difference in the mean outcomes given the mismeasured forcing variable. We present the exact form of the identification bias, which leads us to the misleading consequence in which there is no treatment effect even if there exists a significant treatment effect. To examine the average treatment effect using the mismeasured forcing variable, we propose approximating it using the small error variance approximation originally developed by Chesher (1991). We develop an estimation method for the parameter that approximates the average treatment effect based on local polynomial regressions and the kernel density estimation. Monte Carlo simulations reveal that the identification bias caused by the measurement error is critical, and they corroborate the performance of our approximation analysis.

Future work: While this paper focuses only on the sharp RD design, it is worth investigating the effect of measurement error in the fuzzy RD design in which the forcing variable may contain measurement error. The conditional probability of the treatment may vanish because of the continuous measurement error such that the small error variance approximation cannot be executed in the fuzzy RD design, as we discuss in Remark 3.5. Thus, we would need other approaches to examine the causal effect in the fuzzy RD design with a mismeasured forcing variable.

3.8 Appendix

This appendix presents proofs of the theorems and lemma in the text.

3.8.1 Proof of Theorem 3.1

Proof of (3.6): First,

$$\begin{aligned} Y_i &= E(Y_i|X_i^*) + Y_i - E(Y_i|X_i^*) \\ &= E(D_i Y_{1i}|X_i^*) + E((1-D_i)Y_{0i}|X_i^*) + W_i \\ &= \mathbf{1}(X_i^* \geq x_0)E(Y_{1i}|X_i^*) + \mathbf{1}(X_i^* < x_0)E(Y_{0i}|X_i^*) + W_i \end{aligned}$$

where $W_i \coloneqq Y_i - E(Y_i|X_i^*)$. We thus have

$$\begin{aligned} E(Y_i|X_i = x_0+) &= \int_{x_0}^{\infty} E(Y_{1i}|X_i^* = x^*) f_{X^*|X}(x^*|x_0+)dx^* \\ &\quad + \int_{-\infty}^{x_0} E(Y_{0i}|X_i^* = x^*) f_{X^*|X}(x^*|x_0+)dx^* + E(W_i|X_i = x_0+), \qquad (3.13) \\ E(Y_i|X_i = x_0-) &= \int_{x_0}^{\infty} E(Y_{1i}|X_i^* = x^*) f_{X^*|X}(x^*|x_0-)dx^* \end{aligned}$$

$$+\int_{-\infty}^{x_0} E(Y_{0i}|X_i^*=x^*)f_{X^*|X}(x^*|x_0-)dx^* + E(W_i|X_i=x_0-), \tag{3.14}$$

by Assumption 3.2 and the dominated convergence theorem.

Under Assumption 3.1, $f_{X^*|X}(x^*|x)$ is continuous at $x=x_0$ for $x^*\in\mathbb{R}$ such that $f_{X^*|X}(x^*|x_0+) = f_{X^*|X}(x^*|x_0-)$. This is because

$$\begin{aligned} f_{X^*|X}(x^*|x) &= \frac{f_{X^*X}(x^*,x)}{f_X(x)} \\ &= \frac{f_{X^*}(x^*)f_U(\frac{x-x^*}{\sigma})}{\int_{-\infty}^{\infty} f_{X^*}(x^*)f_U(\frac{x-x^*}{\sigma})dx^*}, \end{aligned}$$

by Assumption 3.1 and the convolution of the probability distributions. Hence, the continuity of $f_{X^*|X}(x^*|\cdot)$ follows from that of $f_U(\cdot)$ and the dominated convergence theorem.

To show (3.6), it thus suffices to show that for any x,

$$E(W_i|X_i=x)=0. \tag{3.15}$$

Because $E(W_i|X_i=x) = E(W_i|X_i=x,D_i=1)\Pr(D_i=1|X_i=x) + E(W_i|X_i=x,D_i=0)\Pr(D_i=0|X_i=x)$ by the law of iterated expectations, we show $E(W_i|X_i=x,D_i=d)=0$ for $d\in\{0,1\}$. To this end, we first compute $f_{W|XD}(w|x,1)$:

$$\begin{aligned} f_{W|XD}(w|x,1) &= \frac{f_{XW|D}(x,w|1)}{f_{X|D}(x|1)} \\ &= \frac{\sigma^{-1}\int_{-\infty}^{\infty} f_{X^*W|D}(x^*,w|1)f_{U|D}(\frac{x-x^*}{\sigma}|1)dx^*}{\sigma^{-1}\int_{-\infty}^{\infty} f_{X^*|D}(x^*|1)f_{U|D}(\frac{x-x^*}{\sigma}|1)dx^*} \\ &= \frac{\int_{x_0}^{\infty} f_{X^*W}(x^*,w)f_U(\frac{x-x^*}{\sigma})dx^*/(1-F_{X^*}(x_0))}{\int_{x_0}^{\infty} f_{X^*}(x^*)f_U(\frac{x-x^*}{\sigma})dx^*/(1-F_{X^*}(x_0))} \\ &= \frac{\int_{x_0}^{\infty} f_{W|X^*}(w|x^*)f_{X^*}(x^*)f_U(\frac{x-x^*}{\sigma})dx^*}{\int_{x_0}^{\infty} f_{X^*}(x^*)f_U(\frac{x-x^*}{\sigma})dx^*}, \end{aligned}$$

where the second and third equalities follow from Assumption 3.1 and the convolution of the probability distributions. Then, we have

$$\begin{aligned} E(W_i|X_i=x,D_i=1) &= \frac{\int w\int_{x_0}^{\infty} f_{W|X^*}(w|x^*)f_{X^*}(x^*)f_U(\frac{x-x^*}{\sigma})dx^*dw}{\int_{x_0}^{\infty} f_{X^*}(x^*)f_U(\frac{x-x^*}{\sigma})dx^*} \\ &= \frac{\int_{x_0}^{\infty} E(W_i|X_i^*=x^*)f_{X^*}(x^*)f_U(\frac{x-x^*}{\sigma})dx^*}{\int_{x_0}^{\infty} f_{X^*}(x^*)f_U(\frac{x-x^*}{\sigma})dx^*} \\ &= 0, \end{aligned} \tag{3.16}$$

where the second equality follows from Fubini's theorem and the third equality follows from $E(W_i|X_i^*)=0$. Similarly, we have

$$E(W_i|X_i=x,D_i=0)=0. \tag{3.17}$$

Therefore, we obtain (3.15). Consequently, we have (3.6) by (3.13), (3.14), and (3.15).

Proof of (3.7): The proof is almost identical to that in Yu (2012). First,

$$\begin{aligned} Y_i &= E(Y_i|X_i^*) + Y_i - E(Y_i|X_i^*) \\ &= E(Y_{0i}|X_i^*) + \mathbf{1}(X_i^* \geq x_0)\{E(Y_{1i} - Y_{0i}|X_i^*) - \tau^*\} + \mathbf{1}(X_i^* \geq x_0)\tau^* + W_i \\ &= m(X_i^*) + \mathbf{1}(X_i^* \geq x_0)\tau^* + W_i \\ &= m(X_i^*) + D_i\tau^* + W_i, \end{aligned} \tag{3.18}$$

where $W_i \coloneqq Y_i - E(Y_i|X_i^*)$ and $m(X_i^*) \coloneqq E(Y_{0i}|X_i^*) + \mathbf{1}(X_i^* \geq x_0)\big(E(Y_{1i} - Y_{0i}|X_i^*) - \tau^*\big)$.

To show (3.7), we compute

$$\begin{aligned} &E(Y_i|X_i = x_0+, D_i = 1) \\ &= E(m(X_i^*)|X_i = x_0+, D_i = 1) + \tau^* + E(W_i|X_i = x_0+, D_i = 1), \end{aligned} \tag{3.19}$$

and

$$E(Y_i|X_i = x_0-, D_i = 0) = E(m(X_i^*)|X_i = x_0-, D_i = 0) + E(W_i|X_i = x_0-, D_i = 0). \tag{3.20}$$

By (3.16) and (3.17), we have $E(W_i|X_i = x_0+, D_i = 1) = E(W_i|X_i = x_0-, D_i = 0) = 0$. To evaluate $E(m(X_i^*)|X_i = x, D_i = 1)$, we compute $f_{X^*|XD}(x^*|x, 1)$:

$$\begin{aligned} f_{X^*|XD}(x^*|x,1) &= \frac{f_{X^*X|D}(x^*, x|1)}{f_{X|D}(x|1)} \\ &= \frac{\mathbf{1}(x^* \geq x_0) f_{X^*X}(x^*, x)/(1 - F_{X^*}(x_0))}{f_{X|D}(x|1)} \\ &= \frac{\mathbf{1}(x^* \geq x_0) f_{X^*|X}(x^*|x) f_X(x)/(1 - F_{X^*}(x_0))}{f_{X|D}(x|1)} \\ &= \frac{\mathbf{1}(x^* \geq x_0) f_{X^*|X}(x^*|x) f_X(x)/(1 - F_{X^*}(x_0))}{\sigma^{-1}\int_{x_0}^{\infty} f_{X^*}(x^*) f_U(\frac{x - x^*}{\sigma}) dx^*/(1 - F_{X^*}(x_0))}, \end{aligned}$$

where the first and third equalities follow from Bayes' theorem and the fourth equality follows from Assumption 3.1 and the convolution of the probability distributions. Then, we have

$$\begin{aligned} E(m(X_i^*)|X_i = x, D_i = 1) &= \frac{\int m(x^*)\mathbf{1}(x^* \geq x_0) f_{X^*|X}(x^*|x) f_X(x) dx^*}{\sigma^{-1}\int_{x_0}^{\infty} f_{X^*}(x^*) f_U(\frac{x - x^*}{\sigma}) dx^*} \\ &= \sigma \frac{\int_{x_0}^{\infty} m(x^*) f_{X^*|X}(x^*|x) f_X(x) dx^*}{\int_{x_0}^{\infty} f_{X^*U}(x^*, \frac{x - x^*}{\sigma}) dx^*} \\ &= \frac{\int_{x_0}^{\infty} m(x^*) f_{X^*|X}(x^*|x) f_X(x) dx^*}{\int_{x_0}^{\infty} f_{X^*X}(x^*, x) dx^*} \\ &= \frac{\int_{x_0}^{\infty} m(x^*) f_{X^*|X}(x^*|x) dx^*}{\int_{x_0}^{\infty} f_{X^*|X}(x^*|x) dx^*}, \end{aligned} \tag{3.21}$$

by the convolution of the probability distributions. Similarly, we obtain

$$E(m(X_i^*)|X_i = x, D_i = 0) = \frac{\int_{-\infty}^{x_0} m(x^*) f_{X^*|X}(x^*|x) dx^*}{\int_{-\infty}^{x_0} f_{X^*|X}(x^*|x) dx^*}. \tag{3.22}$$

Consequently, we obtain the desired result by (3.16), (3.17), (3.19), (3.20), (3.21), and (3.22).

□

3.8.2 Proof of Theorem 3.2

In this proof, for generic A and B, we write $A = B + o(\sigma^2)$ by $A \approx B$ for notational simplicity. We first show that $E(Y_i|X_i^* = x_0+, D_i = 1) \approx \mu(x_0+, 1, \sigma)$. Fix $\varepsilon > 0$. Under Assumptions 3.3 and 3.4, there exists some $e > 0$ such that

$$
\begin{aligned}
&|E(Y_i|X_i^* = x_0+, D_i = 1) - E(Y_i|X_i^* = x_0 + e, D_i = 1)| < \varepsilon, \\
&|\mu(x_0+, 1, \sigma) - \mu(x_0 + e, 1, \sigma)| < \varepsilon.
\end{aligned}
$$

By the triangle inequality, it thus holds that

$$
\begin{aligned}
&|E(Y_i|X_i^* = x_0+, D_i = 1) - \mu(x_0+, 1, \sigma)| \\
\leq &|E(Y_i|X_i^* = x_0+, D_i = 1) - E(Y_i|X_i^* = x_0 + e, D_i = 1)| \\
&+ |E(Y_i|X_i^* = x_0 + e, D_i = 1) - \mu(x_0 + e, 1, \sigma)| + |\mu(x_0 + e, 1, \sigma) - \mu(x_0+, 1, \sigma)| \\
< &|E(Y_i|X_i^* = x_0 + e, D_i = 1) - \mu(x_0 + e, 1, \sigma)| + 2\varepsilon.
\end{aligned}
$$

Hence, we obtain the desired result if we show

$$
E(Y_i|X_i^* = x_0 + e, D_i = 1) \approx \mu(x_0 + e, 1, \sigma), \tag{3.23}
$$

for any $e > 0$, because $\varepsilon > 0$ is arbitrary. The proof of (3.23) is similar to that in Chesher (1991). We set $x = x_0 + e$ for notational simplicity.

To prove (3.23), we calculate an approximation for $f_{Y|XD}(y|x, 1)$. To this end, we first compute the approximation for $f_{YX|D}(y, x|1)$. For any y and u, we have

$$
\begin{aligned}
f_{YXU|D}(y, x, u|1) &= f_{YX^*U|D}(y, x - \sigma u, u|1) \\
&= f_{Y|X^*UD}(y|x - \sigma u, u, 1) f_{X^*U|D}(x - \sigma u, u|1) \\
&= f_{Y|X^*D}(y|x - \sigma u, 1) f_{X^*|D}(x - \sigma u|1) f_U(u),
\end{aligned}
$$

where the second equality follows from Bayes' theorem, and the third equality follows from Assumption 3.1. Applying Taylor's theorem around $\sigma = 0$, it holds that for sufficiently small $\sigma > 0$

$$
\begin{aligned}
&f_{YXU|D}(y, x, u|1) \\
&\approx f_{Y|X^*D}(y|x, 1) f_{X^*|D}(x|1) f_U(u) \\
&- \sigma u f_U(u) \left\{ f^{(1)}_{Y|X^*D}(y|x, 1) f_{X^*|D}(x|1) + f_{Y|X^*D}(y|x, 1) f^{(1)}_{X^*|D}(x|1) \right\} \\
&+ \frac{1}{2}\sigma^2 u^2 f_U(u) \left\{ f^{(2)}_{Y|X^*D}(y|x, 1) f_{X^*|D}(x|1) + 2 f^{(1)}_{Y|X^*D}(y|x, 1) f^{(1)}_{X^*|D}(x|1) + f_{Y|X^*D}(y|x, 1) f^{(2)}_{X^*|D}(x|1) \right\},
\end{aligned}
$$

under Assumptions 3.1 and 3.4. Integrating the both sides with respect to u, we have

$$
\begin{aligned}
f_{YX|D}(y, x|1) \approx &f_{Y|X^*D}(y|x, 1) f_{X^*|D}(x|1) \\
&+ \frac{\sigma^2}{2} \left\{ f^{(2)}_{Y|X^*D}(y|x, 1) f_{X^*|D}(x|1) \right. \\
&\left. + 2 f^{(1)}_{Y|X^*D}(y|x, 1) f^{(1)}_{X^*|D}(x|1) + f_{Y|X^*D}(y|x, 1) f^{(2)}_{X^*|D}(x|1) \right\},
\end{aligned} \tag{3.24}
$$

by Assumptions 3.1 and 3.4.

Similar to Equation (2·5) in Chesher (1991), we have the approximation of $1/f_{X|D}(x|1)$ for sufficiently small $\sigma > 0$:

$$\frac{1}{f_{X|D}(x|1)} \approx \frac{1}{f_{X^*|D}(x|1)} - \frac{\sigma^2}{2}\frac{f^{(2)}_{X^*|D}(x|1)}{(f_{X^*|D}(x|1))^2}, \tag{3.25}$$

under Assumptions 3.1 and 3.4.

Therefore, multiplying (3.24) by (3.25), we have

$$f_{Y|XD}(y|x,1) \approx f_{Y|X^*D}(y|x,1) + \frac{\sigma^2}{2}\left\{2f^{(1)}_{Y|X^*D}(y|x,1)\left(\log^{(1)} f_{X^*|D}(x|1)\right) + f^{(2)}_{Y|X^*D}(y|x,1)\right\}.$$

Furthermore, under Assumption 3.4, this approximation leads to

$$f^{(s)}_{Y|XD}(y|x,1) = f^{(s)}_{Y|X^*D}(y|x,1) + O(\sigma^2),$$

for $s = 1, 2$. We thus have

$$f_{Y|XD}(y|x,1) \approx f_{Y|X^*D}(y|x,1) + \frac{\sigma^2}{2}\left\{2f^{(1)}_{Y|XD}(y|x,1)\left(\log^{(1)} f_{X^*|D}(x|1)\right) + f^{(2)}_{Y|XD}(y|x,1)\right\}.$$

This leads to

$$\begin{aligned} E(Y_i|X_i = x, D_i = 1) \approx{}& E(Y_i|X_i^* = x, D_i = 1) + \sigma^2\left(\log^{(1)} f_{X^*|D}(x|1)\right)E^{(1)}(Y_i|X_i = x, D_i = 1) \\ &+ \frac{\sigma^2}{2}E^{(2)}(Y_i|X_i = x, D_i = 1), \end{aligned}$$

under Assumptions 3.1 and 3.3–3.5. Thus, we have

$$\begin{aligned} E(Y_i|X_i^* = x, D_i = 1) \approx{}& E(Y_i|X_i = x, D_i = 1) - \sigma^2\left(\log^{(1)} f_{X^*|D}(x|1)\right)E^{(1)}(Y_i|X_i = x, D_i = 1) \\ &- \frac{\sigma^2}{2}E^{(2)}(Y_i|X_i = x, D_i = 1), \end{aligned}$$

It holds that

$$\begin{aligned} f_{X^*|D}(x|1) &= f_{X|D}(x|1) + O(\sigma^2), \\ f^{(1)}_{X^*|D}(x|1) &= f^{(1)}_{X|D}(x|1) + O(\sigma^2), \end{aligned}$$

under Assumptions 3.1, 3.4, and 3.5 similar to Equation (2·4) in Chesher (1991). Therefore, it holds that

$$E(Y_i|X_i^* = x, D_i = 1) \approx \mu(x, 1, \sigma).$$

Accordingly, we obtain (3.23) and show $E(Y_i|X_i^* = x_0+, D_i = 1) \approx \mu(x_0+, 1, \sigma)$.

Similarly, we can show that

$$E(Y_i|X_i^* = x_0-, D_i = 0) \approx \mu(x_0-, 0, \sigma).$$

Consequently, we obtain the desired result.

□

3.8.3 Proof of Theorem 3.3

By Lemma 3.1, the consistency of $\hat{g}(x_0, d)$ for $g(x_0, d)$, and Slutsky's theorem, it holds that

$$\sqrt{nh}h^2 \left(\hat{\mu}\left(x_0+, 1, \sigma\right) - \mu\left(x_0+, 1, \sigma\right)\right) \rightsquigarrow N\left(-\sigma^2 e_3' B^+, \sigma^4 e_3' \Omega^+ e_3\right), \tag{3.26}$$

$$\sqrt{nh}h^2 \left(\hat{\mu}\left(x_0-, 0, \sigma\right) - \mu\left(x_0-, 0, \sigma\right)\right) \rightsquigarrow N\left(-\sigma^2 e_3' B^-, \sigma^4 e_3' \Omega^- e_3\right). \tag{3.27}$$

Because the left-hand sides of (3.26) and (3.27) are independent, we obtain the desired result using the continuous mapping theorem.

□

3.8.4 Proof of Lemma 3.1

In this proof, we denote a generic constant as C. We only provide the proof of (3.11), because that of (3.12) is analogous. The proof is an extension of those of Hahn, Todd, and van der Klaauw (1999) and Porter (2003). The minimization problem (3.10) is rewritten as

$$\min_{(a,b')' \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} I_i D_i (Y_i^* - (a - \alpha^+) - (b_1 - \beta_1^+)(X_i - x_0) - \cdots - (b_p - \beta_p^+)(X_i - x_0)^p)^2 K\left(\frac{X_i - x_0}{h}\right),$$

where

$$Y_i^{*+} \coloneqq Y_i - \alpha^+ - \beta_1^+(X_i - x_0) - \cdots - \beta_p^+(X_i - x_0)^p,$$
$$\alpha^+ \coloneqq E(Y_i | X_i = x_0+, D_i = 1), \qquad \beta_k^+ \coloneqq \frac{1}{k!} E^{(k)}(Y_i | X_i = x_0+, D_i = 1) \quad \text{for } k = 1, \ldots, p+1.$$

Define

$$\tilde{Z}_i \coloneqq (1, X_i - x_0, \ldots, (X_i - x_0)^p)', \qquad \tilde{Z} \coloneqq (\tilde{Z}_1, \ldots, \tilde{Z}_n)',$$
$$r \coloneqq (a, b_1, \ldots, b_p)', \qquad \gamma^+ \coloneqq (\alpha^+, \beta_1^+, \ldots, \beta_p^+)', \qquad \hat{\gamma}^+ \coloneqq (\hat{\alpha}^+, \hat{\beta}_1^+, \ldots, \hat{\beta}_p^+)'$$
$$Y^{*+} \coloneqq (Y_1^{*+}, \ldots, Y_n^{*+})', \qquad A_h^+ \coloneqq diag\left(K\left(\frac{X_1 - x_0}{h}\right) I_1 D_1, \ldots, K\left(\frac{X_n - x_0}{h}\right) I_n D_n\right).$$

Then, the minimization problem is

$$\begin{aligned}
&\operatorname*{argmin}_{r \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} I_i D_i \left(Y_i^{*+} - (r - \gamma^+)' \tilde{Z}_i\right)^2 K\left(\frac{X_i - x_0}{h}\right) \\
&= \operatorname*{argmin}_{r \in \mathbb{R}^{p+1}} \left(Y^{*+} - \tilde{Z}(r - \gamma^+)\right)' A_h^+ \left(Y^{*+} - \tilde{Z}(r - \gamma^+)\right).
\end{aligned}$$

By the first-order condition, we have

$$\begin{aligned}
\hat{\gamma}^+ - \gamma^+ &= (\tilde{Z}' A_h^+ \tilde{Z})^{-1} \tilde{Z}' A_h^+ Y^{*+} \\
&= H(H\tilde{Z}' A_h^+ \tilde{Z} H)^{-1} H \tilde{Z}' A_h^+ Y^{*+} \\
&= H(Z' A_h^+ Z)^{-1} Z' A_h^+ Y^{*+},
\end{aligned}$$

where $H \coloneqq diag(1, h^{-1}, \ldots, h^{-p})$, $Z \coloneqq (Z_1, \ldots, Z_n)'$, and $Z_i \coloneqq (1, (X_i - x_0)/h, \ldots, (X_i - x_0)^p/h^p)'$. It holds that

$$\begin{aligned} H^{-1}(\hat{\gamma}^+ - \gamma^+) &= (Z' A_h^+ Z)^{-1} Z' A_h^+ Y^{*+} \\ &= \left(\frac{1}{nh} \sum_{i=1}^n Z_i Z_i' K \left(\frac{X_i - x_0}{h} \right) I_i D_i \right)^{-1} \frac{1}{nh} \sum_{i=1}^n Z_i Y_i^{*+} K \left(\frac{X_i - x_0}{h} \right) I_i D_i. \end{aligned}$$

Therefore, we have the following decomposition:

$$\begin{aligned} &\sqrt{nh} H^{-1}(\hat{\gamma}^+ - \gamma^+) \\ &= \left(\frac{1}{nh} \sum_{i=1}^n Z_i Z_i' K_h(X_i) I_i D_i \right)^{-1} && (3.28) \\ &\left(\frac{1}{\sqrt{nh}} \sum_{i=1}^n \left(Z_i Y_i^{*+} K_h(X_i) I_i D_i - E\left(Z_i Y_i^{*+} K_h(X_i) I_i D_i \middle| X_i, D_i \right) \right) \right. && (3.29) \\ &+ \frac{1}{\sqrt{nh}} \sum_{i=1}^n \left(E\left(Z_i Y_i^{*+} K_h(X_i) I_i D_i \middle| X_i, D_i \right) - E\left(Z_i Y_i^{*+} K_h(X_i) I_i D_i \right) \right) && (3.30) \\ &+ \frac{1}{\sqrt{nh}} \sum_{i=1}^n E\left(Z_i Y_i^{*+} K_h(X_i) I_i D_i \right) - \tilde{B}_n^+ && (3.31) \\ &\left. + \tilde{B}_n^+ \right), && (3.32) \end{aligned}$$

where $\tilde{B}_n^+ \coloneqq \sqrt{nh} h^{p+1} \beta_{p+1} f_X(x_0) E(D_i | X_i = x_0+)(\gamma_{p+1}, \ldots, \gamma_{2p+1})'$. In the following, we study each term separately. Term (3.28) is shown to converge in probability to some constant. Term (3.29) is shown to converge in distribution to the normal distribution. Terms (3.30) and (3.31) are shown to be asymptotically negligible. The multiplication of (3.28) with (3.32) converges to B^+ as $\sqrt{nh} h^{p+1} \to \tilde{C} \in [0, \infty)$ by the following proof.

Term (3.28): We show that

$$\left(\frac{1}{nh} \sum_{i=1}^n Z_i Z_i' K_h(X_i) I_i D_i \right)^{-1} \xrightarrow{p} \frac{1}{E(D_i | X_i = x_0+) f_X(x_0)} (\Gamma^+)^{-1}, \tag{3.33}$$

where $\Gamma^+ \coloneqq (\gamma_{k+l-2})_{(k,l)}$ for $k, l = 1, \ldots, p+1$ and $\gamma_q \coloneqq \int_0^M u^q K(u) du$ for $q = 0, \ldots, 2p$. For $q = 0, \ldots, 2p$,

$$\begin{aligned} &E\left(\frac{1}{nh} \sum_{i=1}^n \left(\frac{X_i - x_0}{h} \right)^q K_h(X_i) I_i D_i \right) \\ &= h^{-1} E\left(\left(\frac{X_i - x_0}{h} \right)^q K_h(X_i) I_i D_i \right) \\ &= h^{-1} E\left(\left(\frac{X_i - x_0}{h} \right)^q K_h(X_i) I_i E(D_i | X_i) \right) \\ &= h^{-1} \int_{x_0}^{x_0 + Mh} \left(\frac{x - x_0}{h} \right)^q K\left(\frac{x - x_0}{h} \right) E(D_i | X_i = x) f_X(x) dx \end{aligned}$$

$$
\begin{aligned}
&= \int_0^M u^q K(u) E(D_i | X_i = x_0 + uh) f_X(x_0 + uh) du \\
&= E(D_i | X_i = x_0+) f_X(x_0) \int_0^M u^q K(u) du + o(1) \\
&= E(D_i | X_i = x_0+) f_X(x_0) \gamma_q + o(1),
\end{aligned}
$$

where the first equality follows from the i.i.d. assumption, the second equality follows from the law of iterated expectations, the third equality follows from the definition of I_i and Assumption 3.6, and the fifth equality follows from Assumptions 3.7 and 3.8 and the dominated convergence theorem. We also have

$$
\begin{aligned}
var\left(\frac{1}{nh} \sum_{i=1}^n \left(\frac{X_i - x_0}{h} \right)^q K_h(X_i) I_i D_i \right) &= \frac{1}{nh^2} var\left(\left(\frac{X_i - x_0}{h} \right)^q K_h(X_i) I_i D_i \right) \\
&\leq \frac{1}{nh^2} E\left(\left(\frac{X_i - x_0}{h} \right)^{2q} K_h^2(X_i) I_i D_i \right) \\
&\leq \frac{1}{nh^2} E\left(\left(\frac{X_i - x_0}{h} \right)^{2q} K_h^2(X_i) I_i \right) \\
&= \frac{1}{nh} \int_0^M u^{2q} K^2(u) f_X(x_0 + uh) du \\
&= \frac{1}{nh} f_X(x_0) \int_0^M u^{2q} K^2(u) du + o\left(\frac{1}{nh} \right) \\
&= o(1),
\end{aligned}
$$

where the first equality follows from the i.i.d. assumption, the second inequality follows from $D_i \leq 1$, the second equality follows from Assumption 3.6, and the third equality follows from Assumptions 3.6 and 3.7 and from the dominated convergence theorem. Therefore, we have shown (3.33) using Markov's inequality and the continuous mapping theorem.

Term (3.29): We show that

$$
\begin{aligned}
&\frac{1}{\sqrt{nh}} \sum_{i=1}^n \left(Z_i Y_i^{*+} K_h(X_i) I_i D_i - E\left(Z_i Y_i^{*+} K_h(X_i) I_i D_i \middle| X_i, D_i \right) \right) \\
&\rightsquigarrow N\left(0, E(V_i^2 | X_i = x_0+, D_i = 1) E(D_i | X_i = x_0+) f_X(x_0) \Delta^+ \right),
\end{aligned} \tag{3.34}
$$

where $\Delta^+ \coloneqq (\delta_{k+l-2})_{(k,l)}$ with $k, l = 1, \ldots, p+1$ and $\delta_l \coloneqq \int_0^M u^l K^2(u) du$ for $l = 0, \ldots, 2p$. To this end, we use the Cramer–Wald device. We observe that

$$
\begin{aligned}
&\frac{1}{\sqrt{nh}} \sum_{i=1}^n \left(Z_i Y_i^{*+} K_h(X_i) I_i D_i - E\left(Z_i Y_i^{*+} K_h(X_i) I_i D_i \middle| X_i, D_i \right) \right) \\
&= \frac{1}{\sqrt{nh}} \sum_{i=1}^n Z_i I_i K_h(X_i) D_i (Y_i^{*+} - E(Y_i^{*+} | X_i, D_i)) \\
&= \frac{1}{\sqrt{nh}} \sum_{i=1}^n Z_i I_i K_h(X_i) D_i (Y_i - E(Y_i | X_i, D_i)) \\
&= \frac{1}{\sqrt{nh}} \sum_{i=1}^n Z_i I_i K_h(X_i) D_i V_i,
\end{aligned}
$$

where $V_i := Y_i - E(Y_i|D_i, X_i)$. Let λ be a nonzero finite vector and

$$U_{n,i} := \frac{1}{\sqrt{nh}} \lambda' Z_i I_i K\left(\frac{X_i - x_0}{h}\right) D_i V_i.$$

By the law of iterated expectations, $E(U_{n,i}) = 0$. For the variance, it holds that for $l = 0, \ldots, 2p$,

$$\begin{aligned}
&E\left(\frac{1}{nh}\left(\frac{X_i - x_0}{h}\right)^l I_i K_h^2(X_i) D_i V_i^2\right) \\
&= \frac{1}{nh} E\left(\left(\frac{X_i - x_0}{h}\right)^l I_i K_h^2(X_i) E(D_i V_i^2 | X_i)\right) \\
&= \frac{1}{nh} E\left(\left(\frac{X_i - x_0}{h}\right)^l I_i K_h^2(X_i) E(V_i^2 | X_i, D_i = 1) E(D_i | X_i)\right) \\
&= \frac{1}{nh} \int_{x_0}^{x_0 + Mh} \left(\frac{x - x_0}{h}\right)^l K^2\left(\frac{x - x_0}{h}\right) E(V_i^2 | X_i = x, D_i = 1) E(D_i | X_i = x) f_X(x) dx \\
&= \frac{1}{n} \int_0^M u^l K^2(u) E(V_i^2 | X_i = x_0 + hu, D_i = 1) E(D_i | X_i = x_0 + hu) f_X(x_0 + hu) du \\
&= \frac{1}{n} E(V_i^2 | X_i = x_0+, D_i = 1) E(D_i | X_i = x_0+) f_X(x_0) \int_0^M u^l K^2(u) du + o(n), \\
&= \frac{1}{n} E(V_i^2 | X_i = x_0+, D_i = 1) E(D_i | X_i = x_0+) f_X(x_0) \delta_l + o(n),
\end{aligned}$$

where the first equality follows from the law of iterated expectations and the fifth equality follows from Assumptions 3.6–3.9 and the dominated convergence theorem. Thus, we have

$$\sum_{i=1}^n var(U_{n,i}) \to E(V_i^2 | X_i = x_0+, D_i = 1) E(D_i | X_i = x_0+) f_X(x_0) \lambda' \Delta^+ \lambda.$$

We next check Lyapunov's condition. Considering some $\zeta > 0$, it holds that

$$\begin{aligned}
&\sum_{i=1}^n E|U_{n,i}|^{2+\zeta} \\
&= \sum_{i=1}^n \left(\frac{1}{nh}\right)^{\frac{\zeta}{2}} \frac{1}{nh} E\left(|\lambda' Z_i|^{2+\zeta} \left|K\left(\frac{X_i - x_0}{h}\right)\right|^{2+\zeta} I_i D_i |V_i|^{2+\zeta}\right) \\
&\leq \left(\frac{1}{nh}\right)^{\frac{\zeta}{2}} \frac{1}{h} E\left(|\lambda' Z_i|^{2+\zeta} \left|K\left(\frac{X_i - x_0}{h}\right)\right|^{2+\zeta} I_i |V_i|^{2+\zeta}\right) \\
&\leq C\left(\frac{1}{nh}\right)^{\frac{\zeta}{2}} \frac{1}{h} E\left(\sum_{l=0}^p \left|\lambda_l \left(\frac{X_i - x_0}{h}\right)^l\right|^{2+\zeta} \left|K\left(\frac{X_i - x_0}{h}\right)\right|^{2+\zeta} I_i E(|V_i|^{2+\zeta} | X_i)\right) \\
&\leq C\left(\frac{1}{nh}\right)^{\frac{\zeta}{2}} \left(\sup_{x \in [0, x_0 + Mh]} E(|V_i|^{2+\zeta} | X_i = x)\right) \int_0^M |K(u)|^{2+\zeta} f_X(x_0 + uh) \sum_{l=0}^p |\lambda_l u^l|^{2+\zeta} du \\
&= O\left(\left(\frac{1}{nh}\right)^{\frac{\zeta}{2}}\right) = o(1),
\end{aligned}$$

where the first inequality follows from $D_i \leq 1$, the second inequality follows from the law of iterated expectations and Loéve's C_r inequality, and the second equality follows from Assumptions 3.6, 3.7, and 3.9. Therefore, by Lyapunov CLT and the Cramer–Wald device, we have shown (3.34).

Term (3.30): We show that

$$\begin{aligned}&\frac{1}{\sqrt{nh}}\left(\sum_{i=1}^{n} E\left(Z_i Y_i^{*+} K_h(X_i) I_i D_i \middle| X_i, D_i\right) - E\left(Z_i Y_i^{*+} K_h(X_i) I_i D_i\right)\right) \\ &= O_p(h^{p+1}) = o_p(1).\end{aligned} \tag{3.35}$$

To this end, we first define

$$\mu_j^+(x) := E(Y_i | X_i = x, D_i = 1) - \left(\alpha^+ + \beta_1^+(x - x_0) + \cdots + \beta_j^+(x - x_0)^j\right),$$

for $j \in \mathbb{N}$, and we use

$$\frac{1}{h^{p+1}} \sup_{x \in (x_0, x_0 + Mh]} \left|\mu_{p+1}^+(x)\right| = o(1), \tag{3.36}$$

by Assumption 3.10 and Taylor's theorem. It is clear that the expectation of the left-hand side of (3.35) is zero by the law of iterated expectations. The variance is

$$\begin{aligned}var\left(\frac{1}{\sqrt{nh}}\sum_{i=1}^{n} E\left(Z_i Y_i^{*+} K_h(X_i) I_i D_i \middle| X_i, D_i\right)\right) &= \frac{1}{h} var\left(E\left(Z_i Y_i^{*+} K_h(X_i) I_i D_i \middle| X_i, D_i\right)\right) \\ &\leq \frac{1}{h} E\left(E\left(\left(Z_i Y_i^{*+} K_h(X_i) I_i D_i\right)^2 \middle| X_i, D_i\right)\right) \\ &\leq \frac{1}{h} E\left(Z_i^2 (Y_i^{*+})^2 K_h^2(X_i) I_i\right),\end{aligned}$$

where the first equality follows from the i.i.d. assumption and the second inequality follows from the law of iterated expectations. For the elements, for $l = 0, \ldots, p$, we have

$$\begin{aligned}&\frac{1}{h} E\left(\left(\frac{X_i - x_0}{h}\right)^{2l} (Y_i^{*+})^2 K_h^2(X_i) I_i\right) \\ =&\frac{1}{h} E\left(\left(\frac{X_i - x_0}{h}\right)^{2l} (\mu_{p+1}^+(X_i) + \beta_{p+1}(X_i - x_0)^{p+1})^2 K_h^2(X_i) I_i\right) \\ \leq& C\frac{1}{h} E\left(\left(\frac{X_i - x_0}{h}\right)^{2l} \left(\left(\mu_{p+1}^+(X_i)\right)^2 + \beta_{p+1}^2 (X_i - x_0)^{2(p+1)}\right) K_h^2(X_i) I_i\right) \\ \leq& C\left(\sup_{x \in (x_0, x_0 + Mh]} \left(\mu_{p+1}^+(x)\right)^2\right) \frac{1}{h} E\left(\left(\frac{X_i - x_0}{h}\right)^{2l} K_h^2(X_i) I_i\right) \\ &+ C\beta_{p+1}^2 \frac{1}{h} E\left(\left(\frac{X_i - x_0}{h}\right)^{2l} (X_i - x_0)^{2(p+1)} K_h^2(X_i) I_i\right) \\ =& Ch^{2(p+1)} \left(\frac{1}{h^{p+1}} \sup_{x \in (x_0, x_0 + Mh]} \left|\mu_{p+1}^+(x)\right|\right)^2 \frac{1}{h} E\left(\left(\frac{X_i - x_0}{h}\right)^{2l} K_h^2(X_i) I_i\right) \\ &+ Ch^{2(p+1)} \beta_{p+1}^2 \frac{1}{h} E\left(\left(\frac{X_i - x_0}{h}\right)^{2(l+p+1)} K_h^2(X_i) I_i\right) \\ =& o\left(h^{2(p+1)}\right) + O\left(h^{2(p+1)}\right),\end{aligned}$$

where the first inequality follows from Loéve's C_r inequality and where the last equality follows from Assumptions 3.6 and 3.7 and (3.36). Therefore, (3.35) holds by Chebyshev's inequality.

Term (3.31): We show that

$$\begin{aligned}&\frac{1}{\sqrt{nh}}\sum_{i=1}^{n} E\left(Z_i Y_i^{*+} K_h(X_i) I_i D_i\right)\\&-\sqrt{nh}h^{p+1}\beta_{p+1}^{+} f_X(x_0) E(D_i|X_i = x_0+)(\gamma_{p+1},\ldots,\gamma_{2p+1})' = o(1).\end{aligned} \tag{3.37}$$

We first observe that

$$\begin{aligned}&\frac{1}{\sqrt{nh}}\sum_{i=1}^{n} E\left(Z_i Y_i^{*+} K_h(X_i) I_i D_i\right)\\=&\frac{\sqrt{n}}{\sqrt{h}} E\left(Z_i K_h(X_i) I_i E\left(Y_i^{*+} D_i|X_i, D_i\right)\right)\\=&\frac{\sqrt{n}}{\sqrt{h}} E\left(Z_i K_h(X_i) I_i E\left(Y_i^{*+}|X_i, D_i = 1\right) E(D_i|X_i)\right)\\=&\frac{\sqrt{n}}{\sqrt{h}} E\left(Z_i K_h(X_i) I_i \left(\mu_{p+1}^{+}(X_i) + \beta_{p+1}(X_i - x_0)^{p+1}\right) E(D_i|X_i)\right),\end{aligned}$$

where the first and second equalities follow from the law of iterated expectations. Then, for $l = 0,\ldots,p$, we have

$$\begin{aligned}&\left|\frac{1}{\sqrt{nh}}\sum_{i=1}^{n} E\left(\left(\frac{X_i - x_0}{h}\right)^l Y_i^{*+} K_h(X_i) I_i D_i\right) - \sqrt{nh}h^{p+1}\beta_{p+1}^{+} f_X(x_0) E(D_i|X_i = x_0+)\gamma_{l+p+1}\right|\\=&\left|\frac{\sqrt{n}}{\sqrt{h}} E\left(\left(\frac{X_i - x_0}{h}\right)^l K_h(X_i) I_i \left(\mu_{p+1}^{+}(X_i) + \beta_{p+1}^{+}(X_i - x_0)^{p+1}\right) E(D_i|X_i)\right)\right.\\&\left.- \sqrt{nh}h^{p+1}\beta_{p+1}^{+} f_X(x_0) E(D_i|X_i = x_0+)\gamma_{l+p+1}\right|\\\le&\left|\frac{\sqrt{n}}{\sqrt{h}} h^{p+1}\beta_{p+1}^{+} E\left(\left(\frac{X_i - x_0}{h}\right)^{l+p+1} K_h(X_i) I_i E(D_i|X_i)\right)\right.\\&\left.- \sqrt{nh}h^{p+1}\beta_{p+1}^{+} f_X(x_0) E(D_i|X_i = x_0+)\gamma_{l+p+1}\right|\\&+\left|\frac{\sqrt{n}}{\sqrt{h}} E\left(\left(\frac{X_i - x_0}{h}\right)^l K_h(X_i) I_i \mu_{p+1}^{+}(X_i) E(D_i|X_i)\right)\right|\\\le&\left|\sqrt{nh}h^{p+1}\beta_{p+1}^{+}\left(\int_0^M u^{l+p+1} K(u) E(D_i|X_i = x_0 + uh) f_X(x_0 + uh) du\right.\right.\\&\left.\left.- f_X(x_0) E(D_i|X_i = x_0+)\gamma_{l+p+1}\right)\right|\\&+\left|\sqrt{nh}h^{p+1}\left(\frac{1}{h^{p+1}}\sup_{x\in(x_0,x_0+Mh]}\left|\mu_{p+1}^{+}(x)\right|\right)\int_0^M u^l K(u) E(D_i|X_i = x_0 + uh) f_X(x_0 + uh) du\right|\\=&o(\sqrt{nh}h^{p+1}) = o(1),\end{aligned}$$

where the first inequality follows from the triangle inequality and the last equality follows from Assumptions 3.6–3.8 and the dominated convergence theorem. Thus, we obtain (3.37).

Consequently, we have the desired result by Slutsky's theorem.

□

Chapter 4

Identification in Weakly Separable Models with a Binary Endogenous Variable

4.1 Introduction

This paper presents an identification analysis for the structural function in a triangular simultaneous equations model with a binary endogenous explanatory variable. We focus on identifying the structural function at specified values of the explanatory variables and specific quantiles of the unobservable variables, in the manner of Chesher (2003, 2005) and Jun et al. (2011). The model in which we study the identification is nonparametric and weakly separable in the sense that the unobservable variable is weakly separable but nonadditive with the explanatory variables in the structural function, as in the model presented by Vytlacil and Yildiz (2007). The presented modeling allows us to identify the values of the structural function by using informative bounds.

Models with nonadditive unobservables have been studied by microeconometric works over the past decade. This field of scholarship has become important for at least two reasons. First, economic theory rarely leads to the functional forms and/or additive separability of structural functions with respect to the unobservable variables. Second, such models allow researchers to understand the heterogeneity of economic agents. Unlike in conventional additively separable models, the ceteris paribus effects in models with nonadditive unobservables can have stochastic variation and be heterogeneous across economic agents. Because these prominent features enrich econometric analyses, research based on models with nonadditive unobservables has been developed and several articles have been devoted to the study of identification and estimation in models with nonadditive unobservables. Examples of such works include Matzkin (2003, 2008), Chernozhukov and Hansen (2005), Chernozhukov, Imbens, and Newey (2007), Chesher (2007b, 2010), Hoderlein and Mammen (2007, 2009), Imbens and Newey (2009), Altonji, Ichimura, and Otsu (2012), and Sasaki (2014). The novel contribution to the literature of the present paper is providing a new identification analysis for the structural function in this setting.

Specifically, we consider the following structural equations model:

$$Y = h_1(g(D, X), \rho_1),$$
$$D = h_2(Z, \rho_2),$$

where Y is an observable scalar dependent variable that can be continuous or discrete, D is an observable binary explanatory variable that may be endogenous, X is a vector of the observable explanatory variables, $Z \coloneqq (W', X')'$ is a vector of the observable instruments, ρ_1 and ρ_2 are unobservable scalar random variables, and ρ_2 is normalized uniformly distributed on $(0, 1)$. The function h_1 is an unknown function in which ρ_1 is weakly separable but possibly nonadditive with the explanatory variables due to the presence of an unknown function g. The function h_2 is an unknown function that is not restricted to being additively separable in ρ_2.

In the above simultaneous equations model, we focus on identifying the values of the structural function $h_1(g(0, x_0), r_1^*)$ and $h_1(g(1, x_1), r_1^*)$, where x_0 and x_1 are certain values of X and $r_1^* \coloneqq Q_{\rho_1|\rho_2}(\tau_1|\tau_2)$ is the conditional τ_1 quantile of ρ_1 given $\rho_2 = \tau_2$, where $\tau_1, \tau_2 \in (0, 1)$. The values of the structural function present valuable information for economic units as discussed in Chesher (2003, 2005).

Our identification analysis consists of two steps. Firstly, we show that the values of the structural function are partially identified by the conditional quantiles of Y given D and Z that are no less or no greater than the values of the structural function. This first step is based on the modified control variate approach presented in Chesher (2005). Secondly, by utilizing the results of the first step and the weak separability of the structural function, we establish more informative bounds for identifying the values of the structural function. For example, the approach for identifying $h_1(g(0, x_0), r_1^*)$ is based on finding the values of X, namely $\underline{x}_0$ and $\overline{x}_0$, satisfying $h_1(g(0, x_0), r_1^*) \geq h_1(g(1, \underline{x}_0), r_1^*)$ or $h_1(g(0, x_0), r_1^*) \leq h_1(g(0, \bar{x}_0), r_1^*)$. In other words, we utilize the values of X that lead the values of the structural function to be no greater or no less than the value of the structural function we wish to identify. Certain subsets of such values of X are identified based on an extension of Vytlacil and Yildiz (2007, Lemma 4.1) using the instrumental variables Z. By combining the identified values of X with the results derived from the first step, we find a tighter identification bound for $h_1(g(0, x_0), r_1^*)$. If there are many such values of X, we generally obtain a further more informative bound for $h_1(g(0, x_0), r_1^*)$. Indeed, the simple example presented in Section 4.4 indicates that there are such values of X even when the support of Z includes a few elements and that the identification analysis leads to an informative identification bound for a value of the structural function. Similarly, we show that a tighter identification bound for $h_1(g(1, x_1), r_1^*)$ is established.

A binary endogenous variable is involved in many important applications. For example, consider analyzing how a college degree affects annual income. In this setting, the binary variable describes whether a person graduates from college, the explanatory variables and instruments include demographic profiles, and the unobservable variables are any acquired and innate abilities that may be correlated with a college degree. Identifying the values of the structural

function in this example allows researchers to infer counterfactual values for annual income when changing college graduate status, demographic profiles, or abilities.

Related literature: The present paper is most closely related to Chesher (2003, 2005) and Jun et al. (2011). Chesher (2003, 2005) provides local identification conditions for structural functions in nonseparable triangular simultaneous equations models with continuous endogenous variables and with a discrete endogenous variable, respectively. He exploits control variate approaches in order to identify the values of the structural function under local order conditions and local rank conditions. However, the local rank condition in Chesher (2005) is never satisfied when the endogenous variable is binary. Intuitively, the reason is that the local rank condition requires the order of the discrete endogenous variable. Since the binary random variable has a qualitative meaning, the local rank condition is not satisfied.

Jun et al. (2011) also study identification for the values of the structural function in a nonseparable equations system with discrete endogenous variables that can be binary, although differences exist between the present paper and theirs. First, while Jun et al. (2011) allow the general nonseparability of structural functions, the unobservables in their paper are restricted to be uniform random variables. Uniform unobservables can be restrictive when combined with the monotonicity of the distribution function of the unobservables, as they also discuss. By contrast, although the present paper does not require that ρ_1 is a uniform random variable, we assume the weak separability of the structural function and that the endogenous variable is binary.

Second, more importantly, the identification approach and identification result in Jun et al. (2011) differ from those in the present paper. Jun et al. (2011) introduce a new rank condition based on a Dynkin system of measurable sets to identify the values of the structural function. By using this new rank condition, they show that the values of the structural function are partially identified by tighter bounds than those proposed by Chesher (2005) and that the identification is also established for binary endogenous variables. On the contrary, by focusing on a binary endogenous variable, the present paper utilizes the values of the covariates such $\underline{x}_0$ and $\overline{x}_0$, which allow us to identify the values of the structural function by using informative bounds. Accordingly, the models, identification approaches, and identification results in both papers are different.

To improve our understanding of these distinctions, we compare the different identification results by using a simple example (see Section 4.4). In this example, we show that the identification analysis in Jun et al. (2011) cannot lead to the identification bound for a value of the structural function, while our identification analysis can.

The present paper is also related to Vytlacil and Yildiz (2007), Shaikh and Vytlacil (2011), and Jun, Pinkse, and Xu (2012). However, while these papers also study identification in weakly separable models, their parameters of interest are different from the objective in the present paper. These papers focus on identifying the average treatment effects, whereas we identify the

values of the structural function.

Organization of the paper: Section 4.2 defines our model and introduces the parameter of interest in a formal manner. Section 4.3 provides our main identification results for the values of the structural function. Section 4.4 demonstrates the identification analysis by using a simple example. Section 4.5 concludes. The proofs of the theorems and lemmas are presented in the Appendix.

Notations: For the generic random variables A and B, we denote the support of A given $B = b$ as $supp(A|B = b)$, the conditional distribution function of A given $B = b$ as $F_{A|B}(a|b) \coloneqq \Pr(A \leq a|B = b)$, and the conditional $p \in (0,1)$ quantile of A given $B = b$ as $Q_{A|B}(p|b) \coloneqq \inf\{a : F_{A|B}(a|b) \geq p\}$. We write the indicator function as $\mathbf{1}(E)$, i.e., $\mathbf{1}(E) = 1$ if the event E is true and $\mathbf{1}(E) = 0$ otherwise. We denote the sign function as $sign(\cdot)$, i.e., $sign(c) = 1$ if $c > 0$, $sign(c) = 0$ if $c = 0$, and $sign(c) = -1$ if $c < 0$.

4.2 The Model

This section introduces the model in which we study the identification of the structural function and the parameter of interest.

We observe the random variables (Y, D, X, Z) related to the following triangular simultaneous equations model as mentioned in the Introduction:

$$\begin{aligned} Y &= h_1(g(D, X), \rho_1), \\ D &= h_2(Z, \rho_2). \end{aligned} \tag{4.1}$$

The dependent variable $Y \in supp(Y) \subset \mathbb{R}$ is a random variable that can be continuous or discrete. The random vector $X \in supp(X) \subset \mathbb{R}^{K_X}$ is a vector of the covariates. The random vector $Z \coloneqq (W', X')' \in supp(Z) \subset \mathbb{R}^{K_Z}$ is a vector of the instrumental variables. The random variables $\rho_1 \in supp(\rho_1) \subset \mathbb{R}$ and $\rho_2 \in (0,1)$ are the unobservable variables. We assume that ρ_1 and ρ_2 are jointly continuously distributed and ρ_2 is normalized uniformly distributed on $(0,1)$ independently of Z. The binary random variable $D \in \{0,1\}$ may be endogenous in the sense that D may be correlated with ρ_1. We denote the conditional probability of $D = d$ given $Z = z$ as $p_d(z) \coloneqq \Pr(D = d|Z = z)$ for $d \in \{0,1\}$ and $z \in supp(Z)$. The functions $h_1 : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$, $g : \{0,1\} \times \mathbb{R}^{K_X} \mapsto \mathbb{R}$, and $h_2 : \mathbb{R}^{K_Z} \times [0,1] \mapsto \{0,1\}$ are unknown structural functions. This model is analogous to the model considered in Vytlacil and Yildiz (2007).

The global exclusion restriction is required in model (4.1). In other words, there must be, at least, a scalar observable variable W that is excluded from the structural function h_1 but is included in the structural function h_2. The global exclusion restriction is also used in Vytlacil and Yildiz (2007) and Jun et al. (2011), whereas Chesher (2005) uses the local exclusion restriction to identify the values of the structural function. Although the global exclusion

restriction can be restrictive in practice, we require one because it allows us to identify the values of the structural function h_1 by using informative bounds.

To develop our identification analysis, we make the following assumptions.

Assumption 4.1. *(i) The function $h_1(t,\cdot)$ is weakly monotonic, normalized caglad*[1] *and nondecreasing, for any $t \in \mathbb{R}$. (ii) The function $h_2(z,\cdot)$ is caglad and nondecreasing for any $z \in \mathbb{R}^{K_Z}$.*

Assumption 4.2. *The random vectors Z and $(\rho_1, \rho_2)'$ are independent.*

Assumption 4.1 is the same as the condition required in Chesher (2005) and Jun et al. (2011). The monotonicity of h_1 guarantees the presence of an inverse function of h_1 with respect to the second argument. It is important to note that under this assumption, the function h_2 evaluated at $Z = z$ can be written as

$$h_2(z,\rho_2) = \begin{cases} 0 & \text{if} \quad 0 \leq \rho_2 \leq p_0(z), \\ 1 & \text{if} \quad p_0(z) < \rho_2 \leq 1. \end{cases} \tag{4.2}$$

This expression of h_2 comes from $h_2(z,\rho_2) = Q_{D|Z}(\rho_2|z)$, which is derived from the monotonicity of h_2 and the normalization of ρ_2. It allows us to clarify the relationship between the structural function h_1 and the conditional quantile function of Y given D and Z. We note that the function h_2 is identified by the observed data since $Q_{D|Z}(r_2|z)$ is identified for any $z \in supp(Z)$ and $r_2 \in (0,1)$.

Assumption 4.2 requires independence between the instrumental vector and unobservable random variables. In other words, the assumption requires the exogeneity of Z. This assumption can be restrictive in practice, although the independence assumption is often introduced in studies of nonparametric identification because of its power (Vytlacil and Yildiz, 2007 and Jun et al., 2011 also use the independence assumption, for instance).

The parameters of interest are the values of the structural function $h_1(g(d,x), r_1^*)$ for $d \in \{0,1\}$, $x \in supp(X)$, and $r_1^* := Q_{\rho_1|\rho_2}(\tau_1|\tau_2)$, where $\tau_1, \tau_2 \in (0,1)$ are the probability indicators selected by the researchers. These parameters present beneficial information for economic agents without assuming the parametric specifications of the functions and the unobservables and without additive separability in the structural functions. For instance, identifying the values of the structural function allows researchers to infer counterfactual values for economic units. In addition, identifying $h_1(g(1,x), r_1^*)$ and $h_1(g(0,x), r_1^*)$ naturally leads us to identify the ceteris paribus effect $h_1(g(1,x), r_1^*) - h_1(g(0,x), r_1^*)$, which is often of interest in empirical settings.

Model (4.1) is more restrictive than the general nonseparable models studied in Chesher (2003, 2005) and Jun et al. (2011), although it is general enough to include many of the models considered in the econometric literature. The specific examples and advantages of weakly separable models are discussed in Vytlacil and Yildiz (2007). We provide a simple example as an illustration.

[1] The caglad function is a function that is left continuous and that has right limits at each point in the range.

Example 4.1. Consider the triangular simultaneous equations model:

$$\begin{aligned} Y =& \mathbf{1}(\alpha D + X'\beta + \rho_1 > 0), \\ D =& \mathbf{1}(Z'\gamma + \rho_2 > 0). \end{aligned}$$

This simultaneous model, studied in Heckman (1978), is a typical parametric example of the weakly separable model (4.1). In this example, $h_1(t, r_1) = \mathbf{1}(t + r_1 > 0)$, $g(d, x) = \alpha d + x'\beta$, and $h_2(z, r_2) = \mathbf{1}(z'\gamma + r_2 > 0)$. The parameter of interest $h_1(g(d, x), r_1^*)$ is equal to $\mathbf{1}(\alpha d + x'\beta + r_1^* > 0)$ for $d \in \{0, 1\}$ and $x \in supp(X)$. The ceteris paribus effect with respect to D is $h_1(g(1, x), r_1^*) - h_1(g(0, x), r_1^*) = \mathbf{1}(\alpha d + x'\beta + r_1^* > 0) - \mathbf{1}(x'\beta + r_1^* > 0)$.

4.3 Identification for the values of the structural function

This section develops identification for the values of the structural function h_1 in two steps. Firstly, we provide partial identification results in which the values of the structural function are no greater or no less than the conditional quantiles of Y given D and Z (Section 4.3.1). Secondly, we provide more informative bounds for the values by utilizing the results in the first step and the weak separability of the structural function (Section 4.3.2).

4.3.1 Partial identification under weak conditions

We first provide an identification result for $h_1(g(0, x_0), r_1^*)$ and then one for $h_1(g(1, x_1), r_1^*)$ because their identification results differ, where $x_0, x_1 \in supp(X)$, $r_1^* := Q_{\rho_1|\rho_2}(\tau_1|\tau_2)$, and $\tau_1, \tau_2 \in (0, 1)$ are the probability indicators selected by the researchers.

In the following, we show that $h_1(g(0, x_0), r_1^*)$ is partially identified by $Q_{Y|DZ}(\tau_1|0, z_0)$ based on the modified control variate approach of Chesher (2005), where $z_0 = (w_0', x_0')'$ is selected by the researchers as long as z_0 satisfies the following condition. To establish partial identification, we require the following assumptions.

Assumption 4.3. *There exists a value of instruments $z_0 = (w_0', x_0')' \in supp(Z)$ satisfying $0 < p_0(z_0) \leq \tau_2$.*

Assumption 4.4. *The conditional distribution function $F_{\rho_1|\rho_2}(r_1|r_2)$ is nondecreasing with respect to $r_2 \in (0, \tau_2]$ for all $r_1 \in \mathbb{R}$.*

Assumption 4.3 requires that $p_0(z_0)$ is no greater than τ_2 and that there are treated units and untreated units conditional on $Z = z_0$. Importantly, in practice, we can examine whether the assumption holds since we can estimate $p_0(z_0)$ from the data.

Assumption 4.4 restricts the direction of the impact of ρ_2 on the conditional distribution of ρ_1. This assumption and expression (4.2) guarantee the presence of particular inequalities between the structural function h_1 and the conditional quantile function of Y given D and Z. Importantly, Assumption 4.4 requires the local (as opposed to the global) monotonicity of the

conditional distribution function.[2] The monotonicity assumption is also required in Chesher (2005) and Jun et al. (2011).[3]

Under Assumptions 4.1–4.4, we present an upper bound for the value of the structural function $h_1(g(0,x_0),r_1^*)$ by the conditional quantile of Y given D and Z. The following result is an extension of the identification result in Chesher (2005).

Lemma 4.1. *Suppose that Assumptions 4.1–4.4 hold. Then, it holds that* [4]

$$\inf supp(Y) \leq h_1(g(0,x_0),r_1^*) \leq Q_{Y|DZ}(\tau_1|0,z_0). \tag{4.3}$$

Lemma 4.1 states that the value of the structural function $h_1(g(0,x_0),r_1^*)$ is partially identified by the conditional quantile $Q_{Y|DZ}(\tau_1|0,z_0)$ since $Q_{Y|DZ}(\tau_1|0,z_0)$ is identified by the observable data. The width of the interval depends on the data-generating process and the values z_0, τ_1, and τ_2. Even if the interval is wide, the partial identification result in the lemma could be informative in the sense discussed in Manski (2009). This partial identification result presents information on the structural function under weak conditions.

We next move onto developing partial identification for the value of the structural function $h_1(g(1,x_1),r_1^*)$, where $x_1 \in supp(X)$ and, again, $r_1^* := Q_{\rho_1|\rho_2}(\tau_1|\tau_2)$ for $\tau_1,\tau_2 \in (0,1)$. The following assumptions guarantee that $h_1(g(1,x_1),r_1^*)$ is partially identified by a conditional quantile of Y given D and Z. The assumptions are analogous to Assumptions 4.3 and 4.4.

Assumption 4.5. *There exists a value of instruments* $z_1 = (w_1', x_1')' \in supp(Z)$ *satisfying* $\tau_2 \leq p_0(z_1) < 1$.

Assumption 4.6. *The conditional distribution function* $F_{\rho_1|\rho_2}(r_1|r_2)$ *is nondecreasing with respect to* $r_2 \in [\tau_2, 1)$ *for all* $r_1 \in \mathbb{R}$.[5]

The following lemma presents a partial identification result for the value of the structural function $h_1(g(1,x_1),r_1^*)$.

Lemma 4.2. *Suppose that Assumptions 4.1, 4.2, 4.5, and 4.6 hold. Then, it holds that* [6]

$$Q_{Y|DZ}(\tau_1|1,z_1) \leq h_1(g(1,x_1),r_1^*) \leq \sup supp(Y). \tag{4.4}$$

Lemma 4.2 states that the value of the structural function $h_1(g(1,x_1),r_1^*)$ is partially identified by the conditional quantile $Q_{Y|DZ}(\tau_1|1,z_1)$ since $Q_{Y|DZ}(\tau_1|1,z_1)$ is identified by the observable data. The difference between the results in Lemmas 4.1 and 4.2 is whether the values of the structural function are no greater or no less than the conditional quantiles.

[2] In fact, we allow $F_{\rho_1|\rho_2}(r_1|\cdot)$ to be nonincreasing instead of Assumption 4.4. We also note that the identification results in the lemmas and theorems depend on the direction of the monotonicity. For simplicity, we assume that $F_{\rho_1|\rho_2}(r_1|\cdot)$ is nondecreasing.

[3] To be exact, Jun et al. (2011) assume that the conditional distribution function is nonincreasing with respect to the conditioning variable.

[4] If we assume that $F_{\rho_1|\rho_2}(r_1|\cdot)$ is nonincreasing instead of Assumption 4.4, we have $Q_{Y|DZ}(\tau_1|0,z_0) \leq h_1(g(0,x_0),r_1^*) \leq \sup supp(Y)$ instead of (4.3).

[5] Again, we allow $F_{\rho_1|\rho_2}(r_1|\cdot)$ to be nonincreasing instead of Assumption 4.6 and the identification results depend on the direction of the monotonicity.

[6] If we assume that $F_{\rho_1|\rho_2}(r_1|\cdot)$ is nonincreasing instead of Assumption 4.6, we have $\inf supp(Y) \leq h_1(g(1,x_1),r_1^*) \leq Q_{Y|DZ}(\tau_1|1,z_1)$ instead of (4.4).

Remark 4.1. The identification results (4.3) and (4.4) are established for any $\tau_1 \in (0,1)$ under these assumptions, which do not require any restriction on τ_1.

Remark 4.2. In fact, the results in Lemmas 4.1 and 4.2 can be shown without the global exclusion restriction and without Assumption 4.2. That is, even when there is no additional instruments W (i.e., when $Z = X$) and X is not independent of ρ_1, the results in Lemmas 4.1 and 4.2 are established if we set $r_1^* = Q_{\rho_1|\rho_2 X}(\tau_1|\tau_2, x_0)$ or $r_1^* = Q_{\rho_1|\rho_2 X}(\tau_1|\tau_2, x_1)$, respectively. However, because the analyses presented in the following subsection require the global exclusion restriction and Assumption 4.2, we also need them in this subsection in order to ensure the consistency of the discussion.

Remark 4.3. We have point identification for the values of the structural function under the exogeneity of D. That is, if the conditional distribution function $F_{\rho_1|\rho_2}(r_1|\cdot)$ is constant for all $r_1 \in \mathbb{R}$, we have the point identification $h_1(g(0,x_0), r_1^*) = Q_{Y|DZ}(\tau_1|0,z_0)$ and $h_1(g(1,x_1), r_1^*) = Q_{Y|DZ}(\tau_1|1,z_1)$. This finding is consistent with the result in Chesher (2005).

4.3.2 More informative bounds

This subsection presents more informative bounds for identifying the values of the structural function based on the results in Lemmas 4.1 and 4.2 and the weak separability of the structural function. The identification analysis requires some additional assumptions, under which we have tighter identification bounds for the values of the structural function than the bounds in Lemmas 4.1 and 4.2. Again, for clarity, we firstly focus on identification for $h_1(g(0,x_0), r_1^*)$ and then move onto identification for $h_1(g(1,x_1), r_1^*)$.

Before we state the formal result, we explain the basic idea of more informative bounds for identifying $h_1(g(0,x_0), r_1^*)$. Results (4.3) and (4.4) in the lemmas lead us to guess that if we know $h_1(g(0,x_0), r_1^*) \geq h_1(g(1,\underline{x}_0), r_1^*)$ for some $\underline{x}_0 \in supp(X)$, the value of the structural function $h_1(g(0,x_0), r_1^*)$ could be intervally identified by

$$Q_{Y|DZ}(\tau_1|1,\underline{z}_0) \leq h_1(g(1,\underline{x}_0), r_1^*) \leq h_1(g(0,x_0), r_1^*) \leq Q_{Y|DZ}(\tau_1|0,z_0),$$

where $\underline{z}_0 = (\underline{w}_0', \underline{x}_0')' \in supp(Z)$ with $\tau_2 \leq p_0(\underline{z}_0) < 1$. In general, this bound is more informative than bound (4.3). Further, if in addition we know $h_1(g(0,x_0), r_1^*) \leq h_1(g(0,\overline{x}_0), r_1^*)$ and $Q_{Y|DZ}(\tau_1|0,z_0) \geq Q_{Y|DZ}(\tau_1|0,\overline{z}_0)$ for some $\overline{z}_0 = (\overline{w}_0', \overline{x}_0')' \in supp(Z)$ with $0 < p_0(\overline{z}_0) \leq \tau_2$, we could have the more informative bound

$$Q_{Y|DZ}(\tau_1|1,\underline{z}_0) \leq h_1(g(0,x_0), r_1^*) \leq Q_{Y|DZ}(\tau_1|0,\overline{z}_0).$$

If there are many values of Z such as $\underline{z}_0$ or $\overline{z}_0$, we could have a further more informative bound for identifying $h_1(g(0,x_0), r_1^*)$.

In fact, we cannot identify all of the values of X such as $\underline{x}_0$ and $\overline{x}_0$. However, certain subsets of such values of X can be identified by using the instrumental variables Z thanks to the weak separability of the structural function. The method is based on an extension

of Vytlacil and Yildiz (2007, Lemma 4.1). To proceed with our analysis, we introduce some additional assumptions.

Assumption 4.7. *For any $r_2 \in (0,1)$, $E(h_1(\cdot,\rho_1)|\rho_2 = r_2)$ is strictly increasing.*[7]

Assumption 4.8. *The function $h_1(\cdot, r_1)$ is nondecreasing for all $r_1 \in supp(\rho_1)$.*[8]

Assumption 4.7 is identical to the assumption employed in Vytlacil and Yildiz (2007). Importantly, this assumption requires neither the strict monotonicity of the structural function $h_1(\cdot, r_1)$ nor the monotonicity of the function g.

Assumption 4.8 requires the weak monotonicity of $h_1(\cdot, r_1)$. While this may be restrictive in practice, many models in the econometric literature satisfy this condition, e.g., nonparametric models with an additive error and parametric discrete choice models such as probit models. It is important to note that under Assumption 4.8, $h_1(g(0,x_0), r_1^*) \geq h_1(g(1,\underline{x}_0), r_1^*)$ is satisfied if and only if $g(0,x_0) \geq g(1,\underline{x}_0)$ and $h_1(g(0,x_0), r_1^*) \leq h_1(g(0,\overline{x}_0), r_1^*)$ is satisfied if and only if $g(0,x_0) \leq g(0,\overline{x}_0)$.

For fixed $x_0 \in supp(X)$, we define the following sets, which are used to get a tighter bound for identifying $h_1(g(0,x_0), r_1^*)$:

$$\begin{aligned}
\underline{\mathcal{Z}}_0(x_0) \coloneqq \big\{ & z = (w', x')' \in supp(Z) : \; h_1(g(1,x), r_1^*) \leq h_1(g(0,x_0), r_1^*); \; \tau_2 \leq p_0(z) < 1; \\
& \exists \hat{w}_0, \hat{\hat{w}}_0, \tilde{w}_0, \tilde{\tilde{w}}_0 \in supp(W) : \; p_0((\hat{w}_0', x_0')') = p_0((\hat{\hat{w}}_0', x')') =: \underline{p}_0, \\
& p_0((\tilde{w}_0', x_0')') = p_0((\tilde{\tilde{w}}_0', x')') =: \underline{q}_0 \;\text{ with }\; \underline{p}_0 \neq \underline{q}_0 \;\text{ and }\; 0 < \underline{p}_0, \underline{q}_0 < 1 \big\}, \\
\overline{\mathcal{Z}}_0(x_0) \coloneqq \big\{ & z = (w', x')' \in supp(Z) : \; h_1(g(0,x), r_1^*) \geq h_1(g(0,x_0), r_1^*); \; 0 < p_0(z) \leq \tau_2; \\
& \exists \dot{w}_0, \ddot{w}_0 \in supp(W) : p_0((\dot{w}_0', x_0')') = p_0((\ddot{w}_0', x')') =: \overline{p}_0 \;\text{ with }\; \overline{p}_0 > 0 \big\}.
\end{aligned}$$

$\underline{\mathcal{Z}}_0(x_0)$ is the set of elements $\underline{z}_0 = (\underline{w}_0', \underline{x}_0')' \in supp(Z)$ satisfying $h_1(g(1,\underline{x}_0), r_1^*) \leq h_1(g(0,x_0), r_1^*)$, $\tau_2 \leq p_0(\underline{z}_0) < 1$, and the existence of corresponding values $\hat{w}_0, \hat{\hat{w}}_0, \tilde{w}_0, \tilde{\tilde{w}}_0 \in supp(W)$ that satisfy the certain conditions. Thus, Lemma 4.2 guarantees the inequality for $\underline{z}_0 \in \underline{\mathcal{Z}}_0(x_0)$:

$$Q_{Y|DZ}(\tau_1|1,\underline{z}_0) \leq h_1(g(1,\underline{x}_0), r_1^*) \leq h_1(g(0,x_0), r_1^*).$$

Similarly, $\overline{\mathcal{Z}}_0(x_0)$ is the set of elements $\overline{z}_0 = (\overline{w}_0', \overline{x}_0')' \in supp(Z)$ satisfying $h_1(g(0,\overline{x}_0), r_1^*) \geq h_1(g(0,x_0), r_1^*)$, $0 < p_0(\overline{z}_0) \leq \tau_2$, and the existence of corresponding values $\dot{w}_0, \ddot{w}_0 \in supp(W)$ that satisfy the certain conditions. Thus, Lemma 4.1 leads the relationship for $\overline{z}_0 \in \overline{\mathcal{Z}}_0(x_0)$:

$$Q_{Y|DZ}(\tau_1|0,\overline{z}_0) \geq h_1(g(0,\overline{x}_0), r_1^*) \geq h_1(g(0,x_0), r_1^*).$$

Hence, if we can identify the sets $\underline{\mathcal{Z}}_0(x_0)$ and $\overline{\mathcal{Z}}_0(x_0)$, these inequalities lead us to find a tighter bound for identifying $h_1(g(0,x_0), r_1^*)$.

The existence of the values $\hat{w}_0, \hat{\hat{w}}_0, \tilde{w}_0, \tilde{\tilde{w}}_0 \in supp(W)$ allows us to identify $\underline{\mathcal{Z}}_0(x_0)$.[9] The condition for $\underline{\mathcal{Z}}_0(x_0)$ requires, for $\underline{z}_0 = (\underline{w}_0', \underline{x}_0')' \in \underline{\mathcal{Z}}_0(x_0)$, the existence of $\hat{w}_0, \hat{\hat{w}}_0, \tilde{w}_0, \tilde{\tilde{w}}_0 \in$

[7] The direction of the strict monotonicity is normalization.

[8] The direction of the monotonicity is normalization.

[9] Note that we allow $\hat{w}_0 = \hat{\hat{w}}_0$ and/or $\tilde{w}_0 = \tilde{\tilde{w}}_0$ in the definition of $\underline{\mathcal{Z}}_0(x_0)$.

$supp(W)$ such that the conditional probability of $D = 0$ given $W = \hat{w}_0$ and $X = x_0$ is equal to the conditional probability of $D = 0$ given $W = \hat{\hat{w}}_0$ and $X = \underline{x}_0$, such that the conditional probability of $D = 0$ given $W = \tilde{w}_0$ and $X = x_0$ is equal to the conditional probability of $D = 0$ given $W = \tilde{\tilde{w}}_0$ and $X = \underline{x}_0$, and such that the probabilities are not identical. In other words, the conditional distribution of D has to be sensitive to the values of W holding $X = x_0$ or $X = \underline{x}_0$ fixed. Therefore, the conditions are related to the sensitivity of the conditional distribution of D given Z and the size of the support of Z. Accordingly, intuitively, the size of $\underline{\mathcal{Z}}_0(x_0)$ could be large when the conditional distribution of D given $Z = z$ is sensitive to the variation in $z \in supp(Z)$ and the support of Z is abundant.

Likewise, the existence of the values $\dot{w}_0, \ddot{w}_0 \in supp(W)$ also allows us to identify $\overline{\mathcal{Z}}_0(x_0)$.[10] The condition for $\overline{\mathcal{Z}}_0(x_0)$ is also related to the sensitivity of the conditional distribution of D to the values of Z. It requires that for $\overline{z}_0 = (\overline{w}_0', \overline{x}_0')' \in \overline{\mathcal{Z}}_0(x_0)$, there exist the values $\dot{w}_0, \ddot{w}_0 \in supp(W)$ such that the conditional probability of $D = 0$ given $W = \dot{w}_0$ and $X = x_0$ is identical to the conditional probability of $D = 0$ given $W = \ddot{w}_0$ and $X = \overline{x}_0$. The size of $\overline{\mathcal{Z}}_0(x_0)$ also depends on the size of the support of Z and the sensitivity of the conditional distribution of D to the values of Z. We note that $z_0 \in \overline{\mathcal{Z}}_0(x_0)$ under Assumption 4.3.

To identify the sets $\underline{\mathcal{Z}}_0(x_0)$ and $\overline{\mathcal{Z}}_0(x_0)$, we define

$$f_1(p, q, x) := -\frac{1}{q-p}\big(E(DY|X = x, p_0(Z) = q) - E(DY|X = x, p_0(Z) = p)\big),$$
$$f_0(p, q, x) := \frac{1}{q-p}\big(E((1-D)Y|X = x, p_0(Z) = q) - E((1-D)Y|X = x, p_0(Z) = p)\big),$$

for $x \in supp(X)$ and $p, q \in supp(p_0(Z)|X = x) \cap (0, 1)$ with $p \neq q$. f_1 and f_0 are similar to the functions used in Vytlacil and Yildiz (2007).[11] We also define

$$m_0(p, x, x') := \frac{1}{p}\big(E((1-D)Y|X = x, p_0(Z) = p) - E((1-D)Y|X = \tilde{x}, p_0(Z) = p)\big),$$

for $x, \tilde{x} \in supp(X)$ and $p \in supp(p_0(Z)|X = x) \cap supp(p_0(Z)|X = \tilde{x}) \cap (0, 1)$.

To proceed with our analysis, we make the following assumption.

Assumption 4.9. *The sets $\underline{\mathcal{Z}}_0(x_0)$ and $\overline{\mathcal{Z}}_0(x_0)$ are nonempty.*

The sets $\underline{\mathcal{Z}}_0(x_0)$ and $\overline{\mathcal{Z}}_0(x_0)$ are identified by the following lemma, which is an extension of Vytlacil and Yildiz (2007, Lemma 4.1).

Lemma 4.3. *Suppose that Assumptions 4.2 and 4.7–4.9 hold. For fixed $x_0 \in supp(X)$, it holds that, for any $x \in supp(X)$ and $p, q \in supp(p_0(Z)|X = x_0) \cap supp(p_0(Z)|X = x) \cap (0, 1)$ with $p \neq q$,*

$$sign\big(g(0, x_0) - g(1, x)\big) = sign\big(f_0(p, q, x_0) - f_1(p, q, x)\big), \tag{4.5}$$

[10] Note that we allow $\dot{w}_0 = \ddot{w}_0$ in the definition of $\overline{\mathcal{Z}}_0(x_0)$.

[11] The slight difference comes from the difference of the normalizations on the structural function for the binary endogenous variable in each paper. Vytlacil and Yildiz (2007) normalize $D = \mathbf{1}(\rho_2 \geq \Pr(D = 1|Z))$, which is slightly different from (4.2), although there is no essential difference between these normalizations.

and

$$sign\big(g(0,x_0)-g(0,x)\big)=sign\big(m_0(p,x_0,x)\big). \tag{4.6}$$

Therefore, the sets $\underline{\mathcal{Z}}_0(x_0)$ *and* $\overline{\mathcal{Z}}_0(x_0)$ *are identified.*

The set $\underline{\mathcal{Z}}_0(x_0)$ is identified because, for fixed $x_0 \in supp(X)$ and $\underline{z}_0 = (\underline{w}_0', \underline{x}_0')' \in \underline{\mathcal{Z}}_0(x_0)$, result (4.5) and Assumptions 4.8 and 4.9 lead to

$$\begin{aligned} sign\big(h_1(g(0,x_0),r_1^*)-h_1(g(1,\underline{x}_0),r_1^*)\big) =& sign\big(g(0,x_0)-g(1,\underline{x}_0)\big) \\ =& sign\big(f_0(\underline{p}_0,\underline{q}_0,x_0)-f_1(\underline{p}_0,\underline{q}_0,\underline{x}_0)\big), \end{aligned}$$

and because $f_0(\underline{p}_0,\underline{q}_0,x_0)$ and $f_1(\underline{p}_0,\underline{q}_0,\underline{x}_0)$ are identified. Likewise, the set $\overline{\mathcal{Z}}_0(x_0)$ is identified because result (4.6) and Assumptions 4.8 and 4.9 lead to the analogous relationship.

Importantly, if the exclusion restriction is not satisfied, Lemma 4.3 does not hold. In other words, unless there is at least one additional instrumental variable W, the sets $\underline{\mathcal{Z}}_0(x_0)$ and $\overline{\mathcal{Z}}_0(x_0)$ are not identified. This is because when there are no additional instruments (i.e., when $Z = X$), no values of $f_1(p,q,x)$ and $f_0(p,q,x)$ are well-defined.

We now have a more informative bound for identifying $h_1(g(0,x_0),r_1^*)$.

Theorem 4.1. *Suppose that Assumptions 4.1, 4.2, 4.4, and 4.6–4.9 hold. Then, it holds that*

$$\sup_{z\in\underline{\mathcal{Z}}_0(x_0)} Q_{Y|DZ}(\tau_1|1,z) \le h_1(g(0,x_0),r_1^*) \le \inf_{z\in\overline{\mathcal{Z}}_0(x_0)} Q_{Y|DZ}(\tau_1|0,z). \tag{4.7}$$

Theorem 4.1 presents the identification bound for $h_1(g(0,x_0),r_1^*)$, which is generally tighter than the bound in Lemma 4.1. As the elements in $\underline{\mathcal{Z}}_0(x_0)$ and $\overline{\mathcal{Z}}_0(x_0)$ increase, the identification bound becomes more informative. Indeed, if there are values $\underline{z}_0 \in \underline{\mathcal{Z}}_0(x_0)$ and $\overline{z}_0 \in \overline{\mathcal{Z}}_0(x_0)$ satisfying $Q_{Y|DZ}(\tau_1|1,\underline{z}_0) = Q_{Y|DZ}(\tau_1|0,\overline{z}_0)$, the value $h_1(g(0,x_0),r_1^*)$ is point-identified. This situation could arise when $\underline{\mathcal{Z}}_0(x_0)$ and $\overline{\mathcal{Z}}_0(x_0)$ include abundant elements, that is, when the support of Z is abundant and the conditional distribution of D is sensitive to the values of Z.

Similarly, for the value $h_1(g(1,x_1),r_1^*)$, we have a tighter identification bound than the bound in Lemma 4.2 based on a similar discussion. We define, for any $x,\tilde{x} \in supp(X)$ and $p \in supp(p_0(Z)|X=x) \cap supp(p_0(Z)|X=\tilde{x}) \cap (0,1)$,

$$m_1(p,x,\tilde{x}) := \frac{1}{1-p}\big(E(DY|X=x,p_0(Z)=p)-E(DY|X=\tilde{x},p_0(Z)=p)\big),$$

and, for fixed $x_1 \in supp(X)$, we define the following sets:

$$\begin{aligned} \underline{\mathcal{Z}}_1(x_1) :=& \big\{ z=(w',x')' \in supp(Z):\ h_1(g(1,x),r_1^*) \le h_1(g(1,x_1),r_1^*);\ \tau_2 \le p_0(z) < 1; \\ & \exists \dot{w}_1,\ddot{w}_1 \in supp(W): p_0((\dot{w}_1',x_1')') = p_0((\ddot{w}_1',x')') =: \underline{p}_1 \ \text{ with } \ \underline{p}_1 < 1 \big\}, \\ \overline{\mathcal{Z}}_1(x_1) :=& \big\{ z=(w',x')' \in supp(Z):\ h_1(g(0,x),r_1^*) \ge h_1(g(1,x_1),r_1^*);\ 0 < p_0(z) \le \tau_2; \\ & \exists \hat{w}_1,\hat{\tilde{w}}_1,\tilde{w}_1,\tilde{\tilde{w}}_1 \in supp(W):\ p_0((\hat{w}_1',x_1')') = p_0((\hat{\tilde{w}}_1',x')') =: \overline{p}_1, \\ & p_0((\tilde{w}_1',x_1')') = p_0((\tilde{\tilde{w}}_1',x')') =: \overline{q}_1 \ \text{ with } \ \overline{p}_1 \ne \overline{q}_1 \ \text{ and } \ 0 < \overline{p}_1,\overline{q}_1 < 1 \big\}. \end{aligned}$$

The definitions of $\underline{\mathcal{Z}}_1(x_1)$ and $\overline{\mathcal{Z}}_1(x_1)$ are analogous to those of $\overline{\mathcal{Z}}_0(x_0)$ and $\underline{\mathcal{Z}}_0(x_0)$, respectively. The sets $\underline{\mathcal{Z}}_1(x_1)$ and $\overline{\mathcal{Z}}_1(x_1)$ are used to give a tighter identification bound for $h_1(g(1,x_1),r_1^*)$. We note that $z_1 \in \underline{\mathcal{Z}}_1(x_1)$ under Assumption 4.5.

Like the identification for $h_1(g(0,x_0),r_1^*)$, we need the following assumption.

Assumption 4.10. *The sets $\underline{\mathcal{Z}}_1(x_1)$ and $\overline{\mathcal{Z}}_1(x_1)$ are nonempty.*

The following lemma shows that the sets $\underline{\mathcal{Z}}_1(x_1)$ and $\overline{\mathcal{Z}}_1(x_1)$ are identified.

Lemma 4.4. *Suppose that Assumptions 4.2, 4.7, 4.8, and 4.10 hold. For fixed $x_1 \in supp(X)$, it holds that, for any $x \in supp(X)$ and $p,q \in supp(p_0(Z)|X=x_1) \cap supp(p_0(Z)|X=x) \cap (0,1)$ with $p \neq q$,*

$$sign\big(g(1,x_1)-g(0,x)\big) = sign\big(f_1(p,q,x_1)-f_0(p,q,x)\big), \tag{4.8}$$

and

$$sign\big(g(1,x_1)-g(1,x)\big) = sign\big(m_1(p,x_1,x)\big). \tag{4.9}$$

Therefore, the sets $\underline{\mathcal{Z}}_1(x_1)$ and $\overline{\mathcal{Z}}_1(x_1)$ are identified.

Based on the results in Lemmas 4.1, 4.2, and 4.4, we have a more informative identification bound for $h_1(g(1,x_1),r_1^*)$.

Theorem 4.2. *Suppose that Assumptions 4.1, 4.2, 4.4, 4.6–4.8, and 4.10 hold. Then, it holds that*

$$\sup_{z\in\underline{\mathcal{Z}}_1(x_1)} Q_{Y|DZ}(\tau_1|1,z) \le h_1(g(1,x_1),r_1^*) \le \inf_{z\in\overline{\mathcal{Z}}_1(x_1)} Q_{Y|DZ}(\tau_1|0,z). \tag{4.10}$$

Theorem 4.2 allows us to identify the value of the structural function $h_1(g(1,x_1),r_1^*)$ by the tighter bound than the bound in Lemma 4.2. This is analogous to the result in Theorem 4.1.

4.4 A simple example demonstrating the power of the identification analysis

This section considers a simple example in order to demonstrate the power of the proposed identification analysis and compare the difference between the identification analysis in the present paper and that in Jun et al. (2011).

Example 4.2. Suppose that the data-generating process is as follows:[12]

$$Y = h_1(g(D,X),\rho_1) = \mathbf{1}(0.1-0.3D-0.2X+\rho_1>0),$$
$$D = h_2(Z,\rho_2) = \mathbf{1}(-0.5+0.1W+0.1X+\rho_2>0),$$

[12]In fact, in this example, we do not need to specify any functional forms for the structural functions to demonstrate the proposed identification result. We need only inequalities (4.11) and (4.12) to demonstrate it. However, we introduce the parametric specifications for the concreteness of the discussion.

where $Z = (W, X)'$, $W \in \{0, 1, 2\}$, $X \in \{0, 1\}$, $supp(Z) = \{(0, 0), (0, 1), (1, 0), (1, 1), (2, 0), (2, 1)\}$, and ρ_2 is uniformly distributed on $(0, 1)$. We do not specify the joint distribution of ρ_1 and ρ_2 except for Assumptions 4.2, 4.4, 4.6, and 4.7. We note that the model satisfies Assumptions 4.1 and 4.8.

Suppose that we would like to identify $h_1(g(0, 1), r_1^*) = \mathbf{1}(-0.1+r_1^* > 0)$ for $r_1^* = Q_{\rho_1|\rho_2}(\tau_1|0.4)$ and some $\tau_1 \in (0, 1)$, i.e., we focus on the value of the structural function at $d = 0$, $x_0 = 1$, and $\tau_2 = 0.4$. We first observe that the probability $p_0(z_0)$ at $z_0 = (w_0, x_0)' = (0, 1)'$ is 0.4 and that this satisfies Assumption 4.3. Thus, Lemma 4.1 leads to

$$\inf supp(Y) \leq h_1(g(0, 1), r_1^*) \leq Q_{Y|DZ}(\tau_1|0, (0, 1)').$$

Although this bound could be informative, Theorem 4.1 leads us to find a tighter bound in general. To this end, we observe the other values of the structural function $h_1(g(0, 0), r_1^*) = \mathbf{1}(0.1 + r_1^* > 0)$, $h_1(g(1, 0), r_1^*) = \mathbf{1}(-0.2 + r_1^* > 0)$, and $h_1(g(1, 1), r_1^*) = \mathbf{1}(-0.4 + r_1^* > 0)$. Therefore, for any r_1^*,

$$h_1(g(0, 0), r_1^*) \geq h_1(g(0, 1), r_1^*) \geq h_1(g(1, 0), r_1^*) \geq h_1(g(1, 1), r_1^*). \tag{4.11}$$

We also observe that the values of $p_0(z)$ are $p_0((0, 0)') = 0.5$, $p_0((0, 1)') = 0.4$, $p_0((1, 0)') = 0.4$, $p_0((1, 1)') = 0.3$, $p_0((2, 0)') = 0.3$, and $p_0((2, 1)') = 0.2$. It thus holds that

$$p_0((0, 0)') > p_0((0, 1)') = p_0((1, 0)') = \tau_2 > p_0((1, 1)') = p_0((2, 0)') > p_0((2, 1)'). \tag{4.12}$$

Therefore, from the definitions of $\underline{\mathcal{Z}}_0(x_0)$ and $\overline{\mathcal{Z}}_0(x_0)$, we have

$$\begin{aligned}
\underline{\mathcal{Z}}_0(1) &= \{(0, 0)', (0, 1)', (1, 0)'\}, \\
\overline{\mathcal{Z}}_0(1) &= \{(0, 1)', (1, 0)', (1, 1)', (2, 0)', (2, 1)'\}.
\end{aligned}$$

Note that in practice, these sets are identified by Lemma 4.3. Thus, according to Theorem 4.1, we have the following identification result:

$$\max_{z \in \underline{\mathcal{Z}}_0(1)} Q_{Y|DZ}(\tau_1|1, z) \leq h_1(g(0, 1), r_1^*) \leq \min_{z \in \overline{\mathcal{Z}}_0(1)} Q_{Y|DZ}(\tau_1|0, z). \tag{4.13}$$

If there are some $\underline{z}_0 \in \underline{\mathcal{Z}}_0(1)$ and $\overline{z}_0 \in \overline{\mathcal{Z}}_0(1)$ satisfying $Q_{Y|DZ}(\tau_1|1, \underline{z}_0) = Q_{Y|DZ}(\tau_1|0, \overline{z}_0)$, the value of the structural function $h_1(g(0, 1), r_1^*)$ is point-identified.

We have another identification result for $h_1(g(0, 1), r_1^*)$ using the identification result developed in Jun et al. (2011, Theorem 1) when assuming ρ_1 is uniformly distributed on $(0, 1)$. We can show that

$$\inf supp(Y) \leq h_1(g(0, 1), r_1^*) \leq \min_{V \in \mathcal{J}^+((0,1)', 0.4)} h_1(g(0, 1), Q_{\rho_1|\rho_2}(\tau_1|V)), \tag{4.14}$$

where $\mathcal{J}^+((0, 1)', 0.4) = \{(0, 0.2], (0, 0.3], (0, 0.4], (0.2, 0.3], (0.2, 0.4], (0.3, 0.4], (0, 0.2] \cup (0.3, 0.4]\}$ and the upper bound is identified. This is shown as follows. As in Jun et al. (2011, pp. 124–125), we define $\mathcal{V}((d, x)') := \{V : V \neq \emptyset;\ \exists w \in supp(W) : V((d, x)', w) = V\}$, where $V((d, x)', w) :=$

$\{r_2 \in (0,1) : h_2(x,w,r_2) = d\}$. Then, we have $\mathcal{V}((0,1)') = \{(0,0.2],(0,0.3],(0,0.4]\}$. By applying the Dynkin system in Jun et al. (2011, Definition 1), we have the Dynkin system of $\mathcal{V}((0,1)')$:

$$\mathcal{D}(\mathcal{V}((0,1)')) = \{(0,0.2],(0,0.3],(0,0.4],(0.2,0.3],(0.2,0.4],(0.3,0.4],(0,0.2]\cup(0.3,0.4]\}.$$

As in Jun et al. (2011, equation (12)), we also define $\mathcal{J}^-((d,x)',\tau_2) \coloneqq \{V \in \mathcal{D}(\mathcal{V}((d,x)')) : V \geq \tau_2\}$ and $\mathcal{J}^+((d,x)',\tau_2) \coloneqq \{V \in \mathcal{D}(\mathcal{V}((d,x)')) : V \leq \tau_2\}$,[13] where $V \geq \tau_2$ ($V \leq \tau_2$) means that all elements in V are no less (no greater) than τ_2. By definition, we have

$$\mathcal{J}^-((0,1)',0.4) = \emptyset,$$
$$\mathcal{J}^+((0,1)',0.4) = \{(0,0.2],(0,0.3],(0,0.4],(0.2,0.3],(0.2,0.4],(0.3,0.4],(0,0.2]\cup(0.3,0.4]\}.$$

Therefore, we have (4.14) by Jun et al. (2011, Theorem 1).

While (4.14) does not tell us the lower bound for identifying $h_1(g(0,1),r_1^*)$, (4.13) leads to the identification bound. Therefore, especially if $h_1(g(0,1),r_1^*) = 1$, while (4.14) can never lead to the point identification for $h_1(g(0,1),r_1^*)$, (4.13) could lead to the point identification for $h_1(g(0,1),r_1^*)$.

Despite the simplicity of the above example, it demonstrates the power of the identification analysis developed in the previous section. The example tells us that even when the support of Z includes a few elements, the sets $\underline{\mathcal{Z}}_0(x_0)$ and $\overline{\mathcal{Z}}_0(x_0)$ lead to informative identification bounds. Indeed, in the example, the numbers of the elements of $\underline{\mathcal{Z}}_0(x_0)$ and $\overline{\mathcal{Z}}_0(x_0)$ are no less than half of the size of the support of Z and the identification analysis could lead to point identification for the value of the structural function. Therefore, especially if the support of Z includes rich elements, the sets $\underline{\mathcal{Z}}_0(x_0)$ and $\overline{\mathcal{Z}}_0(x_0)$ could have abundant elements and the identification analysis could present meaningful information for the structural function.

The example also suggests that there is a case in which the identification analysis in Jun et al. (2011) cannot lead to the identification bound for a value of the structural function, while our identification analysis can. In other words, the analysis in the present paper may lead to a more informative identification result than that in Jun et al. (2011).

Nevertheless, we should note that we cannot know which identification analyses are generally superior because each result is established based on different quantities. Which identification analyses are more informative depends on the data-generating process and value of the structural function we would like to identify.

4.5 Conclusion

This paper presents a new identification analysis for the structural function in weakly separable models (Vytlacil and Yildiz, 2007) with a binary endogenous explanatory variable. We

[13] We note that these definitions are slightly different from equation (12) in Jun et al. (2011). The difference comes from the difference in the assumptions on the direction of the monotonicity of $F_{\rho_1|\rho_2}(r_1|\cdot)$. While we assume that $F_{\rho_1|\rho_2}(r_1|\cdot)$ is nondecreasing, Jun et al. (2011) assume that $F_{\rho_1|\rho_2}(r_1|\cdot)$ is nonincreasing.

focus on identifying the values of the structural function at specified values of the explanatory variables and specific quantiles of the unobservables, in the manner of Chesher (2005) and Jun et al. (2011). Our identification analysis consists of two steps. We firstly show that under weak assumptions, the values of the structural function are partially identified using the conditional quantiles of the dependent variable given the binary endogenous variable and instrumental variables that are no less or no greater than the values of the structural function. This first step is based on the modified control variate approach in Chesher (2005). Secondly, we show that by utilizing the results of the first step and the weak separability of the structural function, the values of the structural function are intervally identified by more informative bounds. This approach is based on identifying the values of the explanatory variables that lead to the values of the structural function being no greater or no less than the value of the structural function we wish to identify. We show that certain subsets of these values of explanatory variables are identified based on an extension of Vytlacil and Yildiz (2007, Lemma 4.1) using the instrumental variables. Combining such values with the results in the first step leads to more informative identification bounds for the values of the structural function. Indeed, the simple example developed herein demonstrates that our identification analysis leads to an informative identification bound for a value of the structural function that is not intervally identified by the existing identification analysis.

Although the present paper focuses on identification, the values of the structural function, in principle, could be nonparametrically estimated through Theorems 4.1 and 4.2. However, developing asymptotic properties of the estimation method would be challenging because the estimation requires estimating the sets $\underline{\mathcal{Z}}_0(x_0)$, $\overline{\mathcal{Z}}_0(x_0)$, $\underline{\mathcal{Z}}_1(x_1)$, and $\overline{\mathcal{Z}}_1(x_1)$ and because the estimation consists of several steps. Such an estimation method is being developed in a separate paper by the author.

4.6 Appendix

This appendix presents the proofs of the theorems and lemmas.

4.6.1 Proofs of the theorems

Proof of Theorem 4.1. We first note that under Assumptions 4.2, 4.7, 4.8, and 4.9, the sets $\underline{\mathcal{Z}}_0(x_0)$ and $\overline{\mathcal{Z}}_0(x_0)$ are identified by Lemma 4.3. Therefore, it holds that for any $\underline{z}_0 = (\underline{w}_0', \underline{x}_0')' \in \underline{\mathcal{Z}}_0(x_0)$ and $\overline{z}_0 = (\overline{w}_0', \overline{x}_0')' \in \overline{\mathcal{Z}}_0(x_0)$

$$Q_{Y|DZ}(\tau_1|1, \underline{z}_0) \leq h_1(g(1, \underline{x}_0), r_1^*) \leq h_1(g(0, x_0), r_1^*) \leq h_1(g(0, \overline{x}_0), r_1^*) \leq Q_{Y|DZ}(\tau_1|0, \overline{z}_0),$$

according to the definitions of $\underline{\mathcal{Z}}_0(x_0)$ and $\overline{\mathcal{Z}}_0(x_0)$ and Lemmas 4.1 and 4.2. Thus, we have the desired result. □

Proof of Theorem 4.2. Note that under Assumptions 4.2, 4.7, 4.8, and 4.10, the sets $\underline{\mathcal{Z}}_1(x_1)$ and $\overline{\mathcal{Z}}_1(x_1)$ are identified by Lemma 4.4. Therefore, it holds that for any $\underline{z}_1 = (\underline{w}_1', \underline{x}_1')' \in \underline{\mathcal{Z}}_1(x_1)$

and $\overline{z}_1 = (\overline{w}_1', \overline{x}_1')' \in \overline{\mathcal{Z}}_1(x_1)$

$$Q_{Y|DZ}(\tau_1|1, \underline{z}_1) \leq h_1(g(1, \underline{x}_1), r_1^*) \leq h_1(g(1, x_1), r_1^*) \leq h_1(g(0, \overline{z}_1), r_1^*) \leq Q_{Y|DZ}(\tau_1|0, \overline{z}_1),$$

according to the definitions of $\underline{\mathcal{Z}}_1(x_1)$ and $\overline{\mathcal{Z}}_1(x_1)$ and Lemmas 4.1 and 4.2. Thus, we have the desired result. □

4.6.2 Proofs of the lemmas

Proof of Lemma 4.1. Although this proof is similar to that in Chesher (2005), we present it here for the benefit of readers. Before we start the proof, we define an inverse function of h_1 under Assumption 4.1. Define

$$h_1^{-1}(g(d, x), y) \coloneqq \sup\{r_1 : h_1(g(d, x), r_1) \leq y\}.$$

We note that the inverse function h_1^{-1} is cadlag[14] and nondecreasing in y and it satisfies

$$h_1(g(d, x), h_1^{-1}(g(d, x), y)) \leq y,$$

with equality holding when $h_1(g(d, x), \rho_1)$ is strictly increasing in ρ_1, and

$$h_1(g(d, x), \rho_1) = \inf\{y : h_1^{-1}(g(d, x), y) \geq \rho_1\}. \tag{4.15}$$

The remaining proof consists of three steps.

Step 1: Under Assumptions 4.1 and 4.2, we show the following equality:

$$F_{Y|DZ}(y|0, z_0) = \frac{1}{p_0(z_0)} \int_0^{p_0(z_0)} F_{\rho_1|\rho_2}(h_1^{-1}(g(0, x_0), y)|r_2) dr_2. \tag{4.16}$$

Firstly, the conditional distribution function of Y given $D = 0$ and $Z = z_0$ is

$$F_{Y|DZ}(y|0, z_0) = \frac{\Pr(Y \leq y \cap D = 0|Z = z_0)}{\Pr(D = 0|Z = z_0)}.$$

Since h_1 is a nondecreasing function with respect to ρ_1, for the event in the numerator on the right-hand side of the above equation, we have

$$\{Y \leq y \cap D = 0|Z = z_0\} \Longleftrightarrow \{\rho_1 \leq h_1^{-1}(g(0, x_0), y) \cap 0 \leq \rho_2 \leq p_0(z_0)|Z = z_0\}.$$

Since $(\rho_1, \rho_2)'$ is independent of Z under Assumption 4.2, it holds that

$$\Pr(Y \leq y \cap D = 0|Z = z_0) = \int_0^{p_0(z_0)} F_{\rho_1|\rho_2}\left(h_1^{-1}(g(0, x_0), y)|r_2\right) dr_2.$$

Therefore, we have (4.16).

[14] The caglad function is a function that is right continuous and that has left limits at each point in the range.

Step 2: Under Assumptions 4.1, 4.2, and 4.4, we show the following inequality:

$$Q_{Y|DZ}(\tau_1|0,z_0) \geq h_1\left(g(0,x_0), Q_{\rho_1|\rho_2}(\tau_1|p_0(z_0))\right). \tag{4.17}$$

Since $F_{\rho_1|\rho_2}(r_1|\cdot)$ is nondecreasing, from (4.16) we have

$$F_{Y|DZ}(y|0,z_0) \leq F_{\rho_1|\rho_2}\left(h_1^{-1}\left(g\left(0,x_0\right),y\right)|p_0(z_0)\right). \tag{4.18}$$

We define

$$\underline{q} \coloneqq \inf\left\{y : F_{\rho_1|\rho_2}\left(h_1^{-1}(g(0,x_0),y)|p_0(z_0)\right) \geq \tau_1\right\}.$$

Note that by definition

$$Q_{Y|DZ}(\tau_1|0,z_0) \coloneqq \inf\{y : F_{Y|DZ}(y|0,z_0) \geq \tau_1\}.$$

Therefore, inequality (4.18) means

$$Q_{Y|DX}(\tau_1|0,z_0) \geq \underline{q}, \tag{4.19}$$

We next derive the relationship between $\underline{q}$ and the structural function h_1. We note that for any y,

$$F_{\rho_1|\rho_2}(h_1^{-1}(g(0,z_0),y)|p_0(z_0)) \geq \tau_1,$$

is satisfied if and only if

$$h_1^{-1}(g(0,z_0),y) \geq Q_{\rho_1|\rho_2}(\tau_1|p_0(z_0)).$$

Therefore, it holds that

$$\underline{q} = \inf\{y : h_1^{-1}(g(0,x_0),y) \geq Q_{\rho_1|\rho_2}(\tau_1|p_0(z_0))\},$$

and, according to equation (4.15), this means

$$\underline{q} = h_1\left(g(0,x_0), Q_{\rho_1|\rho_2}(\tau_1|p_0(z_0))\right).$$

Therefore, from equation (4.19), (4.17) holds.

Step 3: Under Assumptions 4.1–4.4, we show

$$\inf supp(Y) \leq h_1(g(0,x_0),r_1^*) \leq Q_{Y|DZ}(\tau_1|0,z_0).$$

We first note that if $F_{\rho_1|\rho_2}(\tau_1|\cdot)$ is nondecreasing, then $Q_{\rho_1|\rho_2}(r_1|\cdot)$ is nonincreasing. Thus, under Assumption 4.3, we have

$$Q_{\rho_1|\rho_2}(\tau_1|p_0(z_0)) \geq Q_{\rho_1|\rho_2}(\tau_1|\tau_2),$$

and since $h_1(t,\cdot)$ is nondecreasing,

$$h_1\left(g(0,x_0), Q_{\rho_1|\rho_2}(\tau_1|p_0(z_0))\right) \geq h_1\left(g(0,x_0), r_1^*\right). \tag{4.20}$$

Therefore, inequalities (4.17) and (4.20) lead to

$$Q_{Y|DZ}(\tau_1|0,z_0) \geq h_1(g(0,x_0), r_1^*).$$

Since it clearly holds that

$$\inf supp(Y) \leq h_1(g(0,x_0), r_1^*),$$

we find the desired result. □

Proof of Lemma 4.2. This proof is similar to that of Lemma 4.1. Again, the proof consists of three steps.

Step 1: Under Assumptions 4.1 and 4.2, we show the following equality:

$$F_{Y|DZ}(y|1,z_1) = \frac{1}{p_1(z_1)} \int_{p_0(z_1)}^{1} F_{\rho_1|\rho_2}(h_1^{-1}(g(1,x_1),y)|r_2)dr_2. \tag{4.21}$$

Firstly, we observe

$$F_{Y|DZ}(y|1,z_1) = \frac{\Pr(Y \leq y \cap D=1|Z=z_1)}{\Pr(D=1|Z=z_1)}.$$

Since $h_1(t,\cdot)$ is nondecreasing, it holds

$$\{Y \leq y \cap D=1|Z=z_1\} \Longleftrightarrow \{\rho_1 \leq h_1^{-1}(g(1,x_1),y) \cap p_0(z_1) < \rho_2 \leq 1|Z=z_1\},$$

meaning that the probability is

$$\Pr(Y \leq y \cap D=1|Z=z_1) = \int_{p_0(z_1)}^{1} F_{\rho_1|\rho_2}(h_1^{-1}(g(1,x_1),y)|r_2)dr_2,$$

under Assumption 4.2. Therefore, we have equation (4.21).

Step 2: Under Assumptions 4.1, 4.2, and 4.6, we show the following inequality:

$$h_1(g(1,x_1), Q_{\rho_1|\rho_2}(\tau_1|p_0(z_1))) \geq Q_{Y|DZ}(\tau_1|1,z_1). \tag{4.22}$$

Since $F_{\rho_1|\rho_2}(r_1|\cdot)$ is nondecreasing, (4.21) leads to

$$F_{Y|DZ}(y|1,z_1) \geq F_{\rho_1|\rho_2}(h_1^{-1}(g(1,x_1),y)|p_0(z_1)). \tag{4.23}$$

Define

$$\bar{q} \coloneqq \inf\{y|F_{\rho_1|\rho_2}(h_1^{-1}(g(1,x_1),y)|p_0(z_1)) \geq \tau_1\}.$$

By definition, $Q_{Y|DZ}(\tau_1|1,z_1) \coloneqq \inf\{y : F_{Y|DZ}(y|1,z_1) \geq \tau_1\}$ and inequality (4.23) means

$$Q_{Y|DZ}(\tau_1|1,z_1) \leq \bar{q}, \tag{4.24}$$

We next derive the relationship between $\bar{q}$ and h_1. We note that for any y,

$$F_{\rho_1|\rho_2}(h_1^{-1}(g(1,x_1),y)|p_0(z_1)) \geq \tau_1$$

is satisfied if and only if

$$h_1^{-1}(g(1,x_1),y) \geq Q_{\rho_1|\rho_2}(\tau_1|p_0(z_1)).$$

Therefore, it holds that

$$\bar{q} = \inf\{y : h_1^{-1}(g(1,x_1),y) \geq Q_{\rho_1|\rho_2}(\tau_1|p_0(z_1))\},$$

and, according to equation (4.15), this means

$$\bar{q} = h_1(g(1,x_1), Q_{\rho_1|\rho_2}(\tau_1|p_0(z_1))).$$

Therefore, from equation (4.24), we have (4.22).

Step 3: Under Assumptions 4.1, 4.2, 4.5, and 4.6, we show

$$Q_{Y|DZ}(\tau_1|1,z_1) \leq h_1(g(1,x_1),r_1^*) \leq \sup supp(Y).$$

Since $Q_{\rho_1|\rho_2}(\tau_1|\cdot)$ is nonincreasing under Assumption 4.6, we have under Assumption 4.5

$$Q_{\rho_1|\rho_2}(\tau_1|p_0(z_1)) \leq Q_{\rho_1|\rho_2}(\tau_1|\tau_2),$$

and since $h_1(t,\cdot)$ is nondecreasing,

$$h_1(g(1,x_1), Q_{\rho_1|\rho_2}(\tau_1|p_0(z_1))) \leq h_1(g(1,x_1),r_1^*). \tag{4.25}$$

Hence, inequalities (4.22) and (4.25) mean

$$Q_{Y|DZ}(\tau_1|1,z_1) \leq h_1(g(1,x_1),r_1^*). \tag{4.26}$$

Since it clearly holds that

$$h_1(g(1,x_1),r_1^*) \leq \sup supp(Y),$$

we get the desired result. □

Proof of Lemma 4.3. We first note that, for $\underline{z}_0 = (\underline{w}_0', \underline{x}_0')' \in \underline{\mathcal{Z}}_0(x_0)$, there exist the values $\hat{w}_0, \hat{\tilde{w}}_0, \tilde{w}_0, \tilde{\tilde{w}}_0 \in supp(W)$ such that $p_0((\hat{w}_0', x_0')') = p_0((\hat{\tilde{w}}_0', \underline{x}_0')') =: \underline{p}_0$ and $p_0((\tilde{w}_0', x_0')') = p_0((\tilde{\tilde{w}}_0', \underline{x}_0')') =: \underline{q}_0$ with $\underline{p}_0 \neq \underline{q}_0$ and $0 < \underline{p}_0, \underline{q}_0 < 1$. We also note that, for $\overline{z}_0 = (\overline{w}_0', \overline{x}_0') \in$

$\overline{\mathcal{Z}}_0(x_0)$, there exist the values $\dot{w}_0, \ddot{w}_0 \in supp(W)$ such that $p_0((\dot{w}_0', x_0')') = p_0((\ddot{w}_0', \overline{x}_0')') =: \overline{p}_0$ with $\overline{p}_0 > 0$. Therefore, if we show results (4.5) and (4.6), we have

$$\begin{aligned} sign\big(h_1(g(0,x_0),r_1^*) - h_1(g(1,\underline{x}_0),r_1^*)\big) =& sign\big(g(0,x_0) - g(1,\underline{x}_0)\big) \\ =& sign\big(f_0(\underline{p}_0,\underline{q}_0,x_0) - f_1(\underline{p}_0,\underline{q}_0,\underline{x}_0)\big), \end{aligned}$$

and

$$\begin{aligned} sign\big(h_1(g(0,x_0),r_1^*) - h_1(g(0,\overline{x}_0),r_1^*)\big) =& sign\big(g(0,x_0) - g(0,\overline{x}_0)\big) \\ =& sign\big(m_0(\overline{p}_0,x_0,\overline{x}_0)\big), \end{aligned}$$

under Assumption 4.8. Thus, if we show (4.5) and (4.6), the sets $\underline{\mathcal{Z}}_0(x_0)$ and $\overline{\mathcal{Z}}_0(x_0)$ are identified because $f_0(\underline{p}_0,\underline{q}_0,x_0)$, $f_1(\underline{p}_0,\underline{q}_0,\underline{x}_0)$, and $m_0(\overline{p}_0,x_0,\overline{x}_0)$ are identified by the observable data. Hence, we focus on showing (4.5) and (4.6).

Proof of (4.5): The proof is similar to that in Vytlacil and Yildiz (2007). Without loss of generality, we consider $p < q$. Under Assumption 4.2 and model (4.2), we have

$$\begin{aligned} f_0(p,q,x_0) =& \frac{1}{q-p}\left(E((1-D)Y|X=x_0,p_0(Z)=q) - E((1-D)Y|X=x_0,p_0(Z)=p)\right) \\ =& \frac{1}{q-p}\left(\int_0^q E(h_1(g(0,x_0),\rho_1)|X=x_0,\rho_2=r_2)dr_2\right. \\ & \left. - \int_0^p E(h_1(g(0,x_0),\rho_1)|X=x_0,\rho_2=r_2)dr_2\right) \\ =& \frac{1}{q-p}\int_p^q E(h_1(g(0,x_0),\rho_1)|\rho_2=r_2)dr_2, \end{aligned}$$

and

$$\begin{aligned} f_1(p,q,x) =& -\frac{1}{q-p}\left(E(DY|X=x,p_0(Z)=q) - E(DY|X=x,p_0(Z)=p)\right) \\ =& -\frac{1}{q-p}\left(\int_q^1 E(h_1(g(1,x),\rho_1)|X=x,\rho_2=r_2)dr_2\right. \\ & \left. - \int_p^1 E(h_1(g(1,x),\rho_1)|X=x,\rho_2=r_2)dr_2\right) \\ =& \frac{1}{q-p}\int_p^q E(h_1(g(1,x),\rho_1)|\rho_2=r_2)dr_2. \end{aligned}$$

Thus, we have

$$f_0(p,q,x_0) - f_1(p,q,x) = \frac{1}{q-p}\int_p^q E(h_1(g(0,x_0),\rho_1) - h_1(g(1,x),\rho_1)|\rho_2=r_2)dr_2.$$

Therefore, (4.5) holds under Assumption 4.7.

Proof of (4.6): Like the above proof, we observe

$$\begin{aligned}
&m_0(p, x_0, x) \\
=&\frac{1}{p}\big(E((1-D)Y|X = x_0, p_0(Z) = p) - E((1-D)Y|X = x, p_0(Z) = p)\big) \\
=&\frac{1}{p}\int_0^p \big(E(h_1(g(0, x_0), \rho_1)|X = x_0, \rho_2 = r_2) - E(h_1(g(0, x), \rho_1)|X = x, \rho_2 = r_2)\big)dr_2 \\
=&\frac{1}{p}\int_0^p E(h_1(g(0, x_0), \rho_1) - h_1(g(0, x), \rho_1)|\rho_2 = r_2)dr_2.
\end{aligned}$$

Therefore, (4.6) is established under Assumption 4.7. □

Proof of Lemma 4.4. The proof is similar to that of Lemma 4.3. We note that, for $\underline{z}_1 = (\underline{w}_1', \underline{x}_1') \in \underline{\mathcal{Z}}_1(x_1)$, there exist the values $\dot{w}_1, \ddot{w}_1 \in supp(W)$ such that $p_0((\dot{w}_1', x_1')') = p_0((\ddot{w}_1', \underline{x}_1')') =: \underline{p}_1$ with $\underline{p}_1 < 1$. We also note that, for $\overline{z}_1 = (\overline{w}_1', \overline{x}_1') \in \overline{\mathcal{Z}}_1(x_1)$, there exist the values $\hat{w}_1, \hat{\dot{w}}_1, \tilde{w}_1, \tilde{\ddot{w}}_1 \in supp(W)$ such that $p_0((\hat{w}_1', x_1')') = p_0((\hat{\dot{w}}_1', \overline{x}_1')') =: \overline{p}_1$ and $p_0((\tilde{w}_1', x_1')') = p_0((\tilde{\ddot{w}}_1', \overline{x}_1')') =: \overline{q}_1$ with $\overline{p}_1 \neq \overline{q}_1$ and $0 < \overline{p}_1, \overline{q}_1 < 1$. Therefore, if we show (4.8) and (4.9), we have

$$\begin{aligned}
sign\big(h_1(g(1, x_1), r_1^*) - h_1(g(0, \overline{x}_1), r_1^*)\big) =& sign\big(g(1, x_1) - g(0, \overline{x}_1)\big) \\
=& sign\big(f_1(\overline{p}_1, \overline{q}_1, x_1) - f_0(\overline{p}_1, \overline{q}_1, \overline{x}_1)\big),
\end{aligned}$$

and

$$sign\big(h_1(g(1, x_1), r_1^*) - h_1(g(1, \underline{x}_1), r_1^*)\big) = sign\big(g(1, x_1) - g(1, \underline{x}_1)\big) = sign\big(m_1(\underline{p}_1, x_1, \underline{x}_1)\big),$$

under Assumption 4.8. Thus, if we show (4.8) and (4.9), the sets $\overline{\mathcal{Z}}_1(x_1)$ and $\underline{\mathcal{Z}}_1(x_1)$ are identified because $f_1(\overline{p}_1, \overline{q}_1, x_1)$, $f_0(\overline{p}_1, \overline{q}_1, \overline{x}_1)$, and $m_1(\underline{p}_1, x_1, \underline{x}_1)$ are identified by the observable data. Since (4.8) is clear from (4.5), we present the proof of (4.9).

Proof of (4.9): We observe that under Assumption 4.2,

$$\begin{aligned}
&m_1(p, x_1, x) \\
=&\frac{1}{1-p}\big(E(DY|X = x_1, p_0(Z) = p) - E(DY|X = x, p_0(Z) = p)\big) \\
=&\frac{1}{1-p}\int_p^1 \big(E(h_1(g(1, x_1), \rho_1)|X = x_1, \rho_2 = r_2) - E(h_1(g(1, x), \rho_1)|X = x, \rho_2 = r_2)\big)dr_2 \\
=&\frac{1}{1-p}\int_p^1 E(h_1(g(1, x_1), \rho_1) - h_1(g(1, x), \rho_1)|\rho_2 = r_2)dr_2.
\end{aligned}$$

Therefore, (4.9) is established under Assumption 4.7. □

Acknowledgements

I am especially thankful to my advisors Yoshihiko Nishiyama and Ryo Okui for their advice, support, and encouragement. I was greatly fortunate in receiving their sustained and tender guidance during my PhD years at Kyoto University. I also gratefully appreciate Ryo Okui for the joint work in the second chapter of this thesis.

The authors of the second chapter would like to thank Hidehiko Ichimura, Kengo Kato, Katsumi Shimotsu, Hiroyuki Kasahara, Kazuhiko Hayakawa, Yoon-Jae Whang, Simon Lee, Yuya Sasaki, Taisuke Otsu, Hiroaki Kaido, Roger Koenker, Yoosoon Chang, Hashem Pesaran, Yasunori Fujikoshi, Toshio Honda, Eiji Kurozumi, Yohei Yamamoto, seminar participants at the University of Tokyo, Seoul National University, and Hiroshima University, and attendees at the Kansai Econometrics Meeting held in Kyoto, the Institute of Advanced Studies Workshop on Advances in Microeconometrics held in Hong Kong, the 2014 Econometric Society Australasian Meeting, the 20th International Panel Data Conference held in Tokyo, the Mini Conference in Microeconometrics held in Hakone, and the Summer Workshop on Economic Theory 2014 for helpful comments and discussion.

I gratefully appreciates Kohtaro Hitomi, Hisaki Kono, Naoya Sueishi, seminar participants at Kyoto University, and attendees of the Kansai Econometrics Meeting held in Osaka for helpful comments and discussions on the study in the third chapter.

I would like to thank Hidehiko Ichimura and attendees of the econometric seminar at Kyoto University, the Kansai Econometrics Meeting held in Kobe, the Japanese Statistical Society Spring Meeting 2012, and the Japanese Economic Association Spring Meeting 2012 for informative discussions and comments on the study in the fourth chapter.

I also acknowledge financial support from Grant-in-Aid for the Japan Society for the Promotion of Science Fellows No. 252035.

Finally, I would like to thank my family who have persistently supported and encouraged me.

Bibliography

J. G. Altonji, H. Ichimura, and T. Otsu. Estimating derivatives in nonseparable models with limited dependent variables. *Econometrica*, 80(4):1701–1719, 2012.

J. D. Angrist and V. Lavy. Using maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics*, 114(2):533–575, 1999.

Y. Arai and H. Ichimura. Simultaneous selection of optimal bandwidths for the sharp regression discontinuity estimator. *GRIPS Discussion Paper 14-03, National Graduate Institute for Policy Studies*, 2014.

M. Arellano. *Panel Data Econometrics*. Oxford University Press, 2003.

M. Arellano and S. Bonhomme. Identifying distributional characteristics in random coefficients panel data models. *Review of Economic Studies*, 79:987–1020, 2012.

B. H. Baltagi. *Econometric Analysis of Panel Data*. Wiley, 4 edition, 2008.

E. Battistin and A. Chesher. Treatment effect estimation with covariate measurement error. *Journal of Econometrics*, 178(2):707–715, 2014.

E. Battistin, A. Brugiavini, E. Rettore, and G. Weber. The retirement consumption puzzle: evidence from a regression discontinuity approach. *The American Economic Review*, 99(5): 2209–2226, 2009.

R. Blundell and J. L. Powell. Endogeneity in nonparametric and semiparametric regression models. *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, 2:312–357, 2003.

S. Bonhomme and J.-M. Robin. Generalized non-parametric deconvolution with an application to earnings dynamics. *Review of Economic Studies*, 77:491–533, 2010.

J. Bound, C. Brown, and N. Mathiowetz. Measurement error in survey data. *Handbook of Econometrics*, 5:3705–3843, 2001.

M. Browning, M. Ejrnæs, and J. Alvarez. Modelling income processes with lots of heterogeneity. *Rview of Economic Studies*, 77:1353–1381, 2010.

S. Calonico, M. D. Cattaneo, and R. Titiunik. Robust nonparametric confidence intervals for regression-discontinuity designs. *Forthcoming in Econometrica*, 2014.

D. Card and L. D. Shore-Sheppard. Using discontinuous eligibility rules to identify the effects of the federal medicaid expansions on low-income children. *Review of Economics and Statistics*, 86(3):752–766, 2004.

D. Card, C. Dobkin, and N. Maestas. The impact of nearly universal insurance coverage on health care utilization: Evidence from medicare. *American Economic Review*, 98(5):2242–2258, 2008.

G. Chamberlain. Efficiency bounds for semiparametric regression. *Econometrica*, 60(3):567–596, 1992.

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

V. Chernozhukov, G. Imbens, and W. Newey. Instrumental variable estimation of nonseparable models. *Journal of Econometrics*, 139(1):4–14, 2007.

A. Chesher. The effect of measurement error. *Biometrika*, 78(3):451–462, 1991.

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

A. Chesher. Nonparametric identification under discrete variation. *Econometrica*, 73(5):1525–1550, 2005.

A. Chesher. Identification of nonadditive structural functions. *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress*, 3:1–16, 2007a.

A. Chesher. Instrumental values. *Journal of Econometrics*, 139(1):15–34, 2007b.

A. Chesher. Instrumental variable models for discrete outcomes. *Econometrica*, 78(2):575–601, 2010.

A. Chesher and C. Schluter. Welfare measurement and measurement error. *The Review of Economic Studies*, 69(2):357–378, 2002.

J. Davidson. *Stochastic Limit Theory: An Introduction for Econometricians*. Oxford university press, 1994.

Y. A. Davydov. Convergence of distributions generated by stationary stochastic processes. *Theory of Probability & Its Applications*, 13(4):691–696, 1968.

G. Dhaene and K. Jochmans. Split-panel jackknife estimation of fixed effects models. mimeo, 2014.

Y. Dong. Regression discontinuity applications with rounding errors in the running variable. *Journal of Applied Econometrics*, 2014.

K. Evdokimov. Identification and estimation of a nonparametric panel data model with unobserved heterogeneity. mimeo, 2009.

J. Fan and I. Gijbels. Variable bandwidth and local linear regression smoothers. *The Annals of Statistics*, pages 2008–2036, 1992.

J. Fan and I. Gijbels. *Local polynomial modelling and its applications*. CRC Press, 1996.

I. Fernández-Val and J. Lee. Panel data models with nonadditive unobserved heterogeneity: Estimation and inference. *Quantitative Economics*, 4:453–481, 2013.

B. R. Frandsen, M. Frölich, and B. Melly. Quantile treatment effects in the regression discontinuity design. *Journal of Econometrics*, 168(2):382–395, 2012.

A. F. Galvao and K. Kato. Estimation and inference for linear panel data models under misspecification when both n and T are large. *Journal of Business & Economic Statistics*, 32 (2):285–309, 2014.

S. Gonçalves and M. Kaffo. Bootstrap inference for linear dynamic panel data models with individual fixed effects. mimeo, 2014.

B. S. Graham and J. L. Powell. Identification and estimation of average partial effects in "irregular" correlated random coefficient panel data models. *Econometrica*, 80(5):2105–2152, 2012.

F. Guvenen. Learning your earning: Are labor income shocks really very persistent? *American Economic Review*, 97(3):687–712, 2007.

J. Hahn, P. Todd, and W. van der Klaauw. Evaluating the effect of an antidiscrimination law using a regression-discontinuity design. *NBER Working Paper*, 7131:1–32, 1999.

J. Hahn, P. Todd, and W. van der Klaauw. Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1):201–209, 2001.

J. J. Heckman. Dummy endogenous variables in a simultaneous equation system. *Econometrica*, 46(6):931–959, 1978.

S. Hoderlein and E. Mammen. Identification of marginal effects in nonseparable models without monotonicity. *Econometrica*, 75(5):1513–1518, 2007.

S. Hoderlein and E. Mammen. Identification and estimation of local average derivatives in non-separable models without monotonicity. *The Econometrics Journal*, 12(1):1–25, 2009.

J. L. Horowitz. *Semiparametric and nonparametric methods in econometrics*, volume 692. Springer, 2010.

J. L. Horowitz and M. Markatou. Semiparametric estimation of regression models for panel data. *Review of Economic Studies*, 63(1):145–168, 1996.

C. Hsiao, M. H. Pesaran, and A. K. Tahmiscioglu. Bayes estimation of short-run coefficients in dynamic panel data models ", in analysis of panels and limited dependent variables models. In K. L. C. Hsiao, L.F. Lee and M. Pesaran, editors, *Analysis of Panels and Limited Dependent Variables Models*, pages 268–296. Cambridge University press, 1999.

T.-C. Hu, F. Móricz, and R. Taylor. Strong laws of large numbers for arrays of rowwise independent random variables. *Acta Mathematica Hungarica*, 54(1):153–162, 1989.

P. Hullegie and T. J. Klein. The effect of private health insurance on medical care utilization and self-assessed health in germany. *Health Economics*, 19(9):1048–1062, 2010.

H. Ichimura and P. E. Todd. Implementing nonparametric and semiparametric estimators. *Handbook of Econometrics*, 6:5369–5468, 2007.

G. Imbens and K. Kalyanaraman. Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, 79:933–959, 2012.

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

G. W. Imbens and T. Lemieux. Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142(2):615–635, 2008.

M. Jones. On kernel density derivative estimation. *Communications in Statistics-Theory and Methods*, 23(8):2133–2139, 1994.

S. J. Jun, J. Pinkse, and H. Xu. Tighter bounds in triangular systems. *Journal of Econometrics*, 161(2):122–128, 2011.

S. J. Jun, J. Pinkse, and H. Xu. Discrete endogenous variables in weakly separable models. *The Econometrics Journal*, 15(2):288–303, 2012.

M. Kaffo. Bootstrap inference for nonlinear dynamic panel data models with individual fixed effects. mimeo, 2014.

T. G. Koch. Using rd design to understand heterogeneity in health insurance crowd-out. *Journal of Health Economics*, 32(3):599–611, 2013.

A. N. Kolmogorov. *Sulla determinazione empirica di una legge di distribuzione*. 1933.

D. S. Lee. Randomized experiments from non-random selection in us house elections. *Journal of Econometrics*, 142(2):675–697, 2008.

D. S. Lee and D. Card. Regression discontinuity inference with specification error. *Journal of Econometrics*, 142(2):655–674, 2008.

D. S. Lee and T. Lemieux. Regression discontinuity designs in economics. *Journal of Economic Literature*, 48(2):281–355, 2010.

Y. Lee. Bias in dynamic panel models under time series misspecification. *Journal of Econometrics*, 169:54–60, 2012.

Y.-J. Lee, R. Okui, and M. Shintani. Asymptotic inference for dynamic panel estimators of infinite order autoregressive processes. mimeo, 2013.

Q. Li and J. S. Racine. *Nonparametric econometrics: Theory and practice*. Princeton University Press, 2007.

L. A. Lillard and R. J. Willis. Dynamic aspects of earning mobility. *Econometrica*, 46(5): 985–1012, 1978.

C. F. Manski. *Identification for prediction and decision*. Harvard University Press, 2009.

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

E. Masry. Multivariate regression estimation local polynomial fitting for time series. *Stochastic Processes and their Applications*, 65(1):81–101, 1996b.

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

R. L. Matzkin. Nonparametric identification. *Handbook of Econometrics*, 6:5307–5368, 2007.

R. L. Matzkin. Identification in nonparametric simultaneous equations models. *Econometrica*, 76(5):945–978, 2008.

S. Mavroeidis, Y. Sasaki, and I. Welch. Estimation of heterogenous autoregressive parameters using short panel data. mimeo, 2014.

J. McCrary. Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, 142(2):698–714, 2008.

C. Meghir and L. Pistaferri. Income variance dynamics and heterogeneity. *Econometrics*, 72 (1):1–32, 2004.

J. Neyman and E. L. Scott. Consistent estimates based on partially consistent observations. *Econometrica*, 16:1–32, 1948.

S. Nickell. Biases in dynamic models with fixed effects. *Econometrica*, pages 1417–1426, 1981.

R. Okui. Panel AR(1) estimators under misspecification. *Economics Letters*, 101:210–213, 2008.

R. Okui. Asymptotically unbiased estimation of autocovariances and autocorrelations with long panel data. *Econometric Theory*, 26:1263–1304, 2010.

R. Okui. Asymptotically unbiased estimation of autocovariances and autocorrelations for panel data with incidental trends. *Economics Letters*, 112:49–52, 2011.

R. Okui. Asymptotically unbiased estimation of autocovariances and autocorrelations with panel data in the presence of individual and time effects. *Journal of Time Series Econometrics*, 6 (2):129–181, 2014.

A. Pagan and A. Ullah. *Nonparametric econometrics*. Cambridge university press, 1999.

Z. Pei. Regression discontinuity design with measurement error in the assignment variable. mimeo, 2011.

M. H. Pesaran. Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica*, 74(4):967–1002, 2006.

M. H. Pesaran and R. Smith. Estimating long-run relationships from dynamic heterogeneous panels. *Journal of Econometrics*, 68(1):79–113, 1995.

M. H. Pesaran, Y. Shin, and R. P. Smith. Pooled mean group estimation of dynamic heterogeneous panels. *Journal of the American Statistical Association*, 94(446):621–634, 1999.

P. C. B. Phillips and H. R. Moon. Linear regression limit theory for nonstationary panel data. *Econometrica*, 67(5):1057–1111, 1999.

J. Porter. Estimation in the regression discontinuity model. mimeo, 2003.

M. H. Quenouille. Approximate tests of correlation in time-series 3. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 45, pages 483–484. Cambridge Univ Press, 1949.

M. H. Quenouille. Notes on bias in estimation. *Biometrika*, pages 353–360, 1956.

P. C. Reiss and F. A. Wolak. Structural econometric modeling: Rationales and examples from industrial organization. *Handbook of econometrics*, 6:4277–4415, 2007.

D. Ruppert and M. P. Wand. Multivariate locally weighted least squares regression. *The Annals of Statistics*, pages 1346–1370, 1994.

Y. Sasaki. What do quantile regressions identify for general structural functions? *forthcoming in Econometric Theory*, 2014.

D. W. Schanzenbach. Do school lunches contribute to childhood obesity? *Journal of Human Resources*, 44(3):684–709, 2009.

S. M. Schennach. Measurement error in nonlinear models - a review. *Advances in Economics and Econometrics Tenth World Congress*, 3(51):296–337, 2013.

R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley Series in Probability and Statistics. Wiley, 2002.

A. M. Shaikh and E. J. Vytlacil. Partial identification in triangular systems of equations with binary dependent variables. *Econometrica*, 79(3):949–955, 2011.

J. Shao. *Mathematical Statistics*. Springer, 2003.

B. W. Silverman. Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *The Annals of Statistics*, 6(1):177–184, 1978.

N. V. Smirnov. Approximate laws of distribution of random variables from empirical data. *Uspekhi Matematicheskikh Nauk*, (10):179–206, 1944.

G. J. Székely and C. R. Rao. Identifiability of distributions of independent random variables by linear combinations and moments. *Sankhyä*, 62:193–202, 2000.

D. L. Thistlethwaite and D. T. Campbell. Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51(6):309, 1960.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Process*. Springer, 1996.

E. J. Vytlacil and N. Yildiz. Dummy endogenous variables in weakly separable models. *Econometrica*, 75(3):757–779, 2007.

R. Yokoyama. Moment bounds for stationary mixing sequences. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 52(1):45–57, 1980.

P. Yu. Identification of treatment effects in regression discontinuity designs with measurement error. mimeo, 2012.