Essays on Microeconometrics

by

Takuya Ura

Department of Economics
Duke University

Date: _____

Approved:

_____
Federico A. Bugni, Co-Chair

_____
V. Joseph Hotz, Co-Chair

_____
Shakeeb Khan

_____
Matthew A. Masten

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Economics
in the Graduate School of Duke University
2016

Abstract

Essays on Microeconometrics

by

Takuya Ura

Department of Economics
Duke University

Date: _____

Approved:

_____
Federico A. Bugni, Co-Chair

_____
V. Joseph Hotz, Co-Chair

_____
Shakeeb Khan

_____
Matthew A. Masten

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Economics in the Graduate School of Duke University
2016

Abstract

My dissertation has three chapters which develop and apply microeconometric techniques to empirically relevant problems. All the chapters examines the robustness issues (e.g., measurement error and model misspecification) in the econometric analysis. The first chapter studies the identifying power of an instrumental variable in the nonparametric heterogeneous treatment effect framework when a binary treatment variable is mismeasured and endogenous. I characterize the sharp identified set for the local average treatment effect under the following two assumptions: (1) the exclusion restriction of an instrument and (2) deterministic monotonicity of the true treatment variable in the instrument. The identification strategy allows for general measurement error. Notably, (i) the measurement error is nonclassical, (ii) it can be endogenous, and (iii) no assumptions are imposed on the marginal distribution of the measurement error, so that I do not need to assume the accuracy of the measurement. Based on the partial identification result, I provide a consistent confidence interval for the local average treatment effect with uniformly valid size control. I also show that the identification strategy can incorporate repeated measurements to narrow the identified set, even if the repeated measurements themselves are endogenous. Using the the National Longitudinal Study of the High School Class of 1972, I demonstrate that my new methodology can produce nontrivial bounds for the return

to college attendance when attendance is mismeasured and endogenous.

The second chapter, which is a part of a coauthored project with Federico Bugni, considers the problem of inference in dynamic discrete choice problems when the structural model is locally misspecified. We consider two popular classes of estimators for dynamic discrete choice models: K-step maximum likelihood estimators (K-ML) and K-step minimum distance estimators (K-MD), where K denotes the number of policy iterations employed in the estimation problem. These estimator classes include popular estimators such as Rust (1987)'s nested fixed point estimator, Hotz and Miller (1993)'s conditional choice probability estimator, Aguirregabiria and Mira (2002)'s nested algorithm estimator, and Pesendorfer and Schmidt-Dengler (2008)'s least squares estimator. We derive and compare the asymptotic distributions of K-ML and K-MD estimators when the model is arbitrarily locally misspecified and we obtain three main results. In the absence of misspecification, Aguirregabiria and Mira (2002) show that all K-ML estimators are asymptotically equivalent regardless of the choice of K. Our first result shows that this finding extends to a locally misspecified model, regardless of the degree of local misspecification. As a second result, we show that an analogous result holds for all K-MD estimators, i.e., all K-MD estimator are asymptotically equivalent regardless of the choice of K. Our third and final result is to compare K-MD and K-ML estimators in terms of asymptotic mean squared error. Under local misspecification, the optimally weighted K-MD estimator depends on the unknown asymptotic bias and is no longer feasible. In turn, feasible K-MD estimators could have an asymptotic mean squared error that is higher or lower than that of the K-ML estimators. To demonstrate the relevance of our asymptotic analysis, we illustrate our findings using in a simulation exercise

based on a misspecified version of Rust (1987) bus engine problem.

The last chapter investigates the causal effect of the Omnibus Budget Reconciliation Act of 1993, which caused the biggest change to the EITC in its history, on unemployment and labor force participation among single mothers. Unemployment and labor force participation are difficult to define for a few reasons, for example, because of marginally attached workers. Instead of searching for the unique definition for each of these two concepts, this chapter bounds unemployment and labor force participation by observable variables and, as a result, considers various competing definitions of these two concepts simultaneously. This bounding strategy leads to partial identification of the treatment effect. The inference results depend on the construction of the bounds, but they imply positive effect on labor force participation and negligible effect on unemployment. The results imply that the difference-in-difference result based on the BLS definition of unemployment can be misleading due to misclassification of unemployment.

Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

Symbols

$E[\cdot]$	The expectation of a random variable.
$Pr(\cdot)$	The probability of an event.
$\mathbb{R}$	The real line.
$\mathbb{N}$	The set of positive integers.

Abbreviations

AMSE	Asymptotic Mean Square Error
BLS	Bureau of Labor Statistics
CCP	Conditional Choice Probabilty
CPS	Current Population Survey
DGP	Data Generating Process
EITC	Earned Income Tax Credit
LATE	Local Average Treatment Effect
MD	Minimum Distance
MLE	Maximum Likelihood Estimation

Acknowledgements

I would like to thank my committee members, Federico A. Bugni, V. Joseph Hotz, Shakeeb Khan, and Matthew A. Masten, for their guidance and encouragement.

1

Heterogeneous treatment effects with mismeasured endogenous treatment

1.1 Introduction

Treatment effect analyses often entail a measurement error problem as well as an endogeneity problem. For example, Black et al. (2003) document a substantial measurement error in educational attainments in the 1990 U.S. Census. Educational attainments are treatment variables in a return to schooling analysis, and they are endogenous because unobserved individual ability affects both schooling decisions and wages (Card, 2001). The econometric literature, however, has offered only a few solutions for addressing the two problems at the same time. Although an instrumental variable is a standard technique for correcting both endogeneity and measurement error (e.g., Angrist and Krueger, 2001), no paper has investigated the identifying power of an instrumental variable for the heterogeneous treatment effect

when the treatment variable is both mismeasured and endogenous.

I consider a measurement error in the treatment variable in the framework of Imbens and Angrist (1994) and Angrist et al. (1996), and focus on identification and inference problems for the local average treatment effect (LATE). The LATE is the average treatment effect for the subpopulation (the compliers) whose true treatment status is strictly affected by an instrument. Focusing on LATE is meaningful for a few reasons.[1] First, LATE has been a widely used parameter to investigate the heterogeneous treatment effect with endogeneity. My analysis on LATE of a mismeasured treatment variable offers a tool for a robustness check to those who have already investigated LATE. Second, LATE can be used to extrapolate to the average treatment effect or other parameters of interest. Imbens (2010) emphasize the utility of reporting LATE even if the parameter of interest is obtained based on LATE, because the extrapolation often requires additional assumptions and the result of the extrapolation can be less credible than LATE.

I take a worst case scenario approach with respect to the measurement error and allow for arbitrary measurement error. The only assumption concerning the measurement error is its independence of the instrumental variable. The following types of measurement error are considered in my analysis. First, the measurement error is nonclassical; that is, it can be dependent on the true treatment variable. The measurement error for a binary variable is always nonclassical. It is because the measurement error cannot be negative (positive) when the true variable takes the lowest (highest) value. Second, I allow the measurement error to be endogenous; that is,

[1] Deaton (2009) and Heckman and Urzúa (2010) are cautious about interpreting LATE as a parameter of interest. See also Imbens (2010, 2014) for a discussion.

the measured treatment variable is allowed to be dependent on the outcome variable conditional on the true treatment variable. It is also called a differential measurement error. The validation study by Black et al. (2003) finds that the measurement error is likely to be correlated with individual observed and unobserved heterogeneity. The unobserved heterogeneity causes the endogeneity of the measurement error; it affects the measurement and the outcome at the same time. For example, the measurement error for educational attainment depends on the familiarity with the educational system in the U.S., and immigrants may have a higher rate of measurement error. At the same time, the familiarity with the U.S. educational system can be related to the English language skills, which can affect the labor market outcomes. Bound et al. (2001) also argue that measurement error is likely to be differential in some empirical applications. Third, there is no assumption concerning the marginal distribution of the measurement error. It is not necessary to assume anything about the accuracy of the measurement.

Even if I allow for an arbitrary measurement error, this paper demonstrates that an instrumental variable can still partially identify LATE when (a) the instrument satisfies the exclusion restriction such that the instrument affects the outcome and the measured treatment only through the true treatment, and (b) the instrument weakly increases the true treatment. These assumptions are standard in the LATE framework (Imbens and Angrist, 1994 and Angrist et al., 1996). I show that the point identification for LATE is impossible unless LATE is zero, and I characterize the sharp identified set for LATE. Based on the sharp identified set, (i) the sign of LATE is identified, (ii) there are finite upper and lower bounds on LATE even for the unbounded outcome variable, and (iii) the Wald estimand is an upper bound on

LATE in absolute value but sharp upper bound is in general smaller than the Wald estimand. I obtain an upper bound on LATE in absolute value by deriving a new implication of the exclusion restriction.

Inference for LATE in my framework does not fall directly into the existing moment inequality models particularly when the outcome variable is continuous. First, the upper bound for LATE in absolute value is not differentiable with respect to the data distribution. This non-differentiability problem precludes any estimator for the upper bound from having a uniformly valid asymptotic distribution, as is formulated in Hirano and Porter (2012) and Fang and Santos (2014). Second, the upper bound cannot be characterized as the infimum over differentiable functionals indexed by a compact subset in a finite dimensional space, unless the outcome variable has a finite support.[2] This prohibits from applying the existing methodologies in conditional moment inequalities, e.g., Andrews and Shi (2013), Kim (2009), Ponomareva (2010), Armstrong (2014, 2015), Armstrong and Chan (2014), Chetverikov (2013), and Chernozhukov et al. (2013).

I construct a confidence interval for LATE which can be applied to both discrete and continuous outcome variables. To circumvent the aforementioned problems, I approximate the sharp identified set by discretizing the support of the outcome variable where the discretization becomes finer as the sample size increases. The approximation for the sharp identified set resembles many moment inequalities in Menzel (2014) and Chernozhukov et al. (2014), who consider a finite but divergent

[2] When the outcome variable has a finite support, the identified set is characterized by a finite number of moment inequalities. Therefore I can apply the methodologies in unconditional moment inequalities, e.g., Imbens and Manski (2004), Chernozhukov et al. (2007), Romano and Shaikh (2008, 2010), Rosen (2008), Andrews and Guggenberger (2009), Stoye (2009), Andrews and Soares (2010), Bugni (2010), Canay (2010), and Andrews and Jia Barwick (2012).

number of moment inequalities. I adapt a bootstrap method in Chernozhukov et al. (2014) into my framework to construct a confidence interval with uniformly valid asymptotic size control. Moreover, I demonstrate that the confidence interval is consistent against the local alternatives in which a parameter value approaches to the sharp identified set at a certain rate.

As empirical illustrations, I apply the new methodology for evaluating the effect on wages of attending a college when the college attendance can be mismeasured. I use the National Longitudinal Survey of the High School Class of 1972 (NLS-72), as in Kane and Rouse (1995). Using the proximity to college as an instrumental variable (Card, 1995), the confidence interval developed in the present paper offers nontrivial bounds on LATE, even if I allow for measurement error in college attendance. Moreover, the empirical results confirm the theoretical result that the Wald estimator is an upper bound on LATE but is not the sharp upper bound.

As an extension, I demonstrate that my identification strategy offers a new use of repeated measurements as additional sources for identification. The existing practice of the repeated measurements exploits them as instrumental variables, as in Hausman et al. (1991) and Hausman et al. (1995). However, when the true treatment variable is endogenous, the repeated measurements are likely to be endogenous and are not good candidates for an instrumental variable. My identification strategy shows that those variables are useful for bounding LATE in the presence of measurement error, even if the repeated measurement are not valid instrumental variables. I give a necessary and sufficient condition under which the repeated measurement strictly narrows the identified set.

The remainder of the present paper is organized as follows. Subsection 1.1.1 ex-

plains several examples motivating mismeasured endogenous treatment variables and Subsection 1.1.2 reviews the related econometric literature. Section 1.2 introduces the LATE framework with mismeasured treatment variables. Section 1.3 constructs the identified set for LATE. Section 1.4 proposes an inference procedure for LATE. Section 1.5 conducts the Monte Carlo simulations. Section 1.6 implements the inference procedure in NLS-72 to estimate the return to schooling. Section 1.7 discusses how repeated measurements narrow the identified set, even if the repeated measurements themselves are not instrumental variables. Section 1.8 concludes. The Appendix collects proofs and remarks.

1.1.1 Examples for mismeasured endogenous treatment variables

I introduce several examples in which binary treatment variables can be both endogenous and mismeasured at the same time. The first example is the return to schooling, in which the outcome variable is wages and the treatment variable is educational attainment, for example, whether a person has completed college or not. It is well-known that unobserved individual ability affects both the schooling decision and wage determination, which leads to the endogeneity of educational attainment in the wage equation (see, for example, Card (2001)). Moreover, survey datasets record educational attainments based on the interviewee's answers and these self-reported educational attainments are subject to measurement error. Empirical papers by Griliches (1977), Angrist and Krueger (1999), Kane et al. (1999), Card, 2001, Black et al. (2003) have pointed out the mismeasurement. For example, Black et al. (2003) estimate that the 1990 Decennial Census has 17.7% false positive rate of reporting a doctoral degree.

The second example is labor supply response to welfare program participation, in which the outcome variable is employment status and the treatment variable is welfare program participation. Self-reported welfare program participation in survey datasets can be mismeasured (Hirano and Porter, 2012). The psychological cost for welfare program participation, welfare stigma, affects job search behavior and welfare program participation simultaneously; that is, welfare stigma may discourage individuals from participating in a welfare program, and, at the same time, affect an individual's effort in the labor market (see Moffitt (1983) and Besley and Coate (1992) for a discussion on the welfare stigma). Moreover, the welfare stigma gives welfare recipients some incentive not to reveal their participation status to the survey, which causes differential measurement error in that the unobserved individual heterogeneity affects both the measurement error and the outcome.

The third example is the effect of a job training program on wages (for example, Royalty, 1996). As it is similar to the return to schooling, the unobserved individual ability plays a key role in this example. Self-reported completion of job training program is also subject to measurement error (Bollinger, 1996). Frazis and Loewenstein (2003) develop a methodology for evaluating a homogeneous treatment effect with mismeasured endogenous treatment variable, and apply their methodology to evaluate the effect of a job training program on wages.

The last example is the effect of maternal drug use on infant birth weight. Kaestner et al. (1996) estimate that a mother tends to underreport her drug use, but, at the same time, she tends to report it correctly if she is a heavy user. When the degree of drug addiction is not observed, it becomes an individual unobserved heterogeneity variable which affects infant birth weight and the measurement in addition to the

drug use.

1.1.2 Literature review

This paper is related to a few strands of the econometric literature. First, Mahajan (2006), Lewbel (2007) and Hu (2008) use an instrumental variable to correct for measurement error in a binary treatment in the heterogeneous treatment effect framework and they achieve nonparametric point identification of the average treatment effect. This result assumes the true treatment variable is exogenous, whereas I allow it to be endogenous.

Finite mixture models are related to this paper. I consider the unobserved binary treatment, whereas finite mixture models deal with unobserved type variable. Henry et al. (2014) and Henry et al. (2015) are the most closely related to this paper. They investigate the identification problem in finite mixture models, by using the exclusion restriction in which an instrumental variable only affects the mixing distribution of a type variable without affecting the component distribution (that is, the conditional distribution given the type variable). If I applied their approach directly to my framework, their exclusion restriction would imply conditional independence between the instrumental variable and the outcome variable given the true treatment variable. In the LATE framework, this conditional independence implies that LATE does not exhibit essential heterogeneity (Heckman et al., 2010) and that LATE is equal to the mean difference between the control and treatment groups.[3] Instead of applying the

[3] This footnote uses the notation introduced in Section 1.2. The conditional independence implies $E[Y \mid T^*, Z] = E[Y \mid T^*]$. Under this assumption,

$$\begin{aligned} E[Y \mid Z] &= P(T^* = 1 \mid Z)E[Y \mid Z, T^* = 1] + P(T^* = 0 \mid Z)E[Y \mid Z, T^* = 0] \\ &= P(T^* = 1 \mid Z)E[Y \mid T^* = 1] + P(T^* = 0 \mid Z)E[Y \mid T^* = 0] \end{aligned}$$

approaches in Henry et al. (2014) and Henry et al. (2015), this paper uses a different exclusion restriction in which the instrumental variable does not affect the outcome or the measured treatment directly.

A few papers have applied an instrumental variable to a mismeasured binary regressor in the homogenous treatment effect framework. They include Aigner (1973), Kane et al. (1999), Bollinger (1996), Black et al. (2000) and Frazis and Loewenstein (2003). Frazis and Loewenstein (2003) is the most closely related to the present paper among them, since they consider an endogenous mismeasured regressor. In contrast, I allow for heterogeneous treatment effects. Therefore, I contribute to the heterogeneous treatment effect literature by investigating the consequences of the measurement errors in the treatment variable.

Kreider and Pepper (2007), Molinari (2008), Imai and Yamamoto (2010), and Kreider et al. (2012) apply a partial identification strategy for the average treatment effect to the mismeasured binary regressor problem by utilizing the knowledge of the marginal distribution for the true treatment. Those papers use auxiliary datasets to obtain the marginal distribution for the true treatment. Kreider et al. (2012) is the most closely related to the present paper, in that they allow for both treatment endogeneity and differential measurement error. My instrumental variable approach can be an an alternative strategy to deal with mismeasured endogenous treatment.

and therefore $\Delta E[Y \mid Z] = \Delta E[T^* \mid Z](E[Y \mid T^* = 1] - E[Y \mid T^* = 0])$. I obtain the equality

$$\frac{\Delta E[Y \mid Z]}{\Delta E[T^* \mid Z]} = E[Y \mid T^* = 1] - E[Y \mid T^* = 0]$$

This above equation implies that the LATE does not depend on the compliers of consideration, which is in contrast with the essential heterogeneity of the treatment effect (Heckman and Urzúa, 2010). Furthermore, since $E[Y \mid T^* = 1] - E[Y \mid T^* = 0]$ is equal to the LATE, I do not need to care about the endogeneity problem here.

It is worthwhile because, as mentioned in Schennach (2013), the availability of an auxiliary dataset is limited in empirical research. Furthermore, it is not always the case that the results from auxiliary datasets is transported into the primary dataset (Carroll et al., 2012, p.10),

Some papers investigate mismeasured endogenous continuous variables, instead of binary variables. Amemiya (1985); Hsiao (1989); Lewbel (1998); Song et al. (2015) consider nonlinear models with mismeasured continuous explanatory variables. The continuity of the treatment variable is crucial for their analysis, because they assume classical measurement error. The treatment variable in the present paper is binary and therefore the measurement error is nonclassical. Hu et al. (2015) consider mismeasured endogenous continuous variables in single index models. However, their approach depends on taking derivatives of the conditional expectations with respect to the continuous variable. It is not clear if it can be extended to binary variables. Song (2015) considers the semi-parametric model when endogenous continuous variables are subject to nonclassical measurement error. He assumes conditional independence between the instrumental variable and the outcome variable given the true treatment variable, which imposes some structure on the outcome equation (e.g., LATE does not exhibit essential heterogeneity). Instead this paper proposes an identification strategy without assuming any structure on the outcome equation.

Chalak (2013) investigates the consequences of measurement error in the instrumental variable instead of the treatment variable. He assumes that the treatment variable is perfectly observed, whereas I allow for it to be measured with error. Since I assume that the instrumental variable is perfectly observed, my analysis is not overlapped with Chalak (2013).

Manski (2003), Blundell et al. (2007), and Kitagawa (2010) have similar identification strategy to the present paper in the context of sample selection models. These papers also use the exclusion restriction of the instrumental variable for their partial identification results. Particularly, Kitagawa (2010) derives the "integrated envelope" from the exclusion restriction, which is similar to the total variation distance in the present paper because both of them are characterized as a supremum over the set of the partitions. First and the most importantly, the present paper considers mismeasurement of the treatment variable, whereas the sample selection model considers truncation of the outcome variable. It is not straightforward to apply their methodologies in sample selection models into mismeasured treatment problem. Second, the present paper offers an inference method with uniform size control, but Kitagawa (2010) derives only point-wise size control. Last, Blundell et al. (2007) and Kitagawa (2010) use their result for specification test, but I cannot use it for specification test because the sharp identified set of the present paper is always non-empty.

1.2 LATE model with misclassification

This section introduces measurement error in the treatment variable into the LATE framework (Imbens and Angrist, 1994, and Angrist et al., 1996). The objective is to evaluate the causal effect of a binary treatment variable $T^* \in \{0, 1\}$ on an outcome variable Y, where $T^* = 0$ represents the control group and $T^* = 1$ represents the treatment group. To control for endogeneity of T^*, the LATE framework requires a binary instrumental variable $Z \in \{z_0, z_1\}$ which shifts T^* exogenously without any direct effect on Y. The treatment variable T^* of interest is not directly observed, and

instead there is a binary measurement $T \in \{0, 1\}$ for T^*. I put the $*$ symbol on T^* to emphasize that the true treatment variable T^* is unobserved. Y can be discrete, continuous or mixed; Y is only required to have some dominating finite measure μ_Y on the real line. μ_Y can be the Lebesgue measure or the counting measure.

To describe the data generating process, I consider the counterfactual variables. Let T_z^* denote the counterfactual true treatment variable when $Z = z$. Let Y_{t*} denote the counterfactual outcome when $T^* = t^*$. Let T_{t*} denote the potential measured treatment variable when $T^* = t^*$. The individual treatment effect is $Y_1 - Y_0$. It is not directly observed; Y_0 and Y_1 are not observed at the same time. Only Y_{T*} is observable. Using the notation, the observed variables (Y, T, Z) are generated by the following three equations:

$$T = T_{T*} \tag{1.1}$$

$$Y = Y_{T*} \tag{1.2}$$

$$T^* = T_Z^*. \tag{1.3}$$

Figure 1.13 describes the above three equations graphically. (1.1) is the measurement equation, which is the arrow from Z to T^* in Figure 1.13. $T - T^*$ is the measurement error; $T - T^* = 1$ (or $T_0 = 1$) represents a false positive and $T - T^* = -1$ (or $T_1 = 0$) represents a false negative. The next two equations (1.2) and (1.3) are standard in the LATE framework. (1.2) is the outcome equation, which is the arrow from T^* to Y in Figure 1.13. (1.3) is the treatment assignment equation, which is the arrow from T^* to T in Figure 1.13. . Correlation between (Y_0, Y_1) and $(T_{z_0}^*, T_{z_1}^*)$ causes an endogeneity problem.

In a return to schooling analysis, Y is wages, T^* is the true indicator for college

completion, Z is the proximity to college, and T is the measurement of T^*. The treatment effect $Y_1 - Y_0$ in the return to schooling is the effect of college attendance T^* on wages Y. The college attendance is not correctly measured in a dataset, such that only the proxy T is observed.

The only assumption for my identification analysis is as follows.

Assumption 1. *(i) Z is independent of $(T_{t*}, Y_{t*}, T^*_{z_0}, T^*_{z_1})$ for each $t^* = 0, 1$. (ii) $T^*_{z_1} \geqslant T^*_{z_0}$ almost surely.*

Part (i) is the exclusion restriction and I consider stochastic independence instead of mean independence. Although it is stronger than the minimal conditions for the identification for LATE without measurement error, a large part of the existing applied papers assume stochastic independence (Huber and Mellace, 2015, p.405). Z is also independent of T_{t*} conditional on $(Y_{t*}, T^*_{z_0}, T^*_{z_1})$, which is the only assumption on the measurement error for the identified set in Section 1.3.

Part (ii) is the monotonicity condition for the instrument, in which the instrument Z increases the value of T^* for all the individuals. de Chaisemartin (2015) relaxes the monotonicity condition, and the following analysis of the present paper only requires the complier-defiers-for-marginals condition in de Chaisemartin (2015) instead of the monotonicity condition. Moreover, Part (ii) implies that the sign of the first stage regression, which is the effect of the instrumental variable on the true treatment variable, is known. It is a reasonable assumption because most empirical applications of the LATE framework assume the sign is known. For example, Card (1995) claims that the proximity-to-college instrument weakly increases the likelihood of going to a college. Last, I do not assume a relevance condition for the instrumental variable,

such as $T^*_{z_1} \neq T^*_{z_0}$. The relevance condition is a testable assumption when $T^* = T$, but it is not testable in my analysis. I will discuss the relevance condition in my framework after Theorem 3.

As I emphasized in the introduction, the framework here does not assume anything on measurement error except for the independence from Z. I do not impose any restriction on the marginal distribution of the measurement error or on the relationship between the measurement error and $(Y_{t^*}, T^*_{z_0}, T^*_{z_1})$. Particularly, the measurement error can be differential, that is, T_{t^*} can depend on Y_{t^*}.

In this paper, I focus on the local average treatment effect (LATE), which is defined by

$$\theta \equiv E[Y_1 - Y_0 \mid T^*_{z_0} < T^*_{z_1}].$$

LATE is the average of the treatment effect $Y_1 - Y_0$ over the subpopulation (the compliers) whose treatment status depend on the instrument. Imbens and Angrist (1994, Theorem 1) show that LATE equals

$$\frac{\Delta E[Y \mid Z]}{\Delta E[T^* \mid Z]},$$

where I define $\Delta E[X \mid Z] = E[X \mid Z = z_1] - E[X \mid Z = z_0]$ for a random variable X. The present paper introduces measurement error in the treatment variable, and therefore the fraction $\Delta E[Y \mid Z]/\Delta E[T^* \mid Z]$ is not equal to the Wald estimand

$$\frac{\Delta E[Y \mid Z]}{\Delta E[T \mid Z]}.$$

Since $\Delta E[T^* \mid Z]$ is not point identified, I cannot point identify LATE. The failure for the point identification comes purely from the measurement error, because LATE would be point identified under $T = T^*$.

1.3 Sharp identified set for LATE

This section considers the partial identification problem for LATE. Before defining the sharp identified set, I express LATE as a function of the underlying distribution P^* of $(Y_0, Y_1, T_0, T_1, T^*_{z_0}, T^*_{z_1}, Z)$. I use the $*$ symbol on P^* to clarify that P^* is the distribution of the unobserved variables. In the following arguments, I denote the expectation operator E by E_{P^*} when I need to clarify the underlying distribution. The local average treatment effect is a function of the unobserved distribution P^*:

$$\theta(P^*) \equiv E_{P^*}[Y_1 - Y_0 \mid T^*_{z_0} < T^*_{z_1}].$$

The sharp identified set is the set of parameter values for LATE which is consistent with the distribution of the observed variables. I use P for the distribution of the observed variables (Y, T, Z) The equations (1.1), (1.2), and (1.3) induce the distribution of the observables (Y, T, Z) from the unobserved distribution P^*, and I denote by $P^*_{(Y,T,Z)}$ the induced distribution. When the distribution of (Y, T, Z) is P, the set of P^* which induces $P^*_{(Y,T,Z)} = P$ is

$$\{P^* \in \mathcal{P}^* : P = P^*_{(Y,T,Z)}\},$$

where $\mathcal{P}^*$ is the set of P^*'s satisfying Assumptions 1. For every distribution P of (Y, T, Z), the sharp identified set for LATE is defined as

$$\Theta_I(P) \equiv \{\theta(P^*) : P^* \in \mathcal{P}^* \text{ and } P = P^*_{(Y,T,Z)}\}.$$

The proof of Theorem 1 in Imbens and Angrist (1994) provides a relationship between $\Delta E[Y \mid Z]$ and LATE:

$$\theta(P^*)P^*(T^*_{z_0} < T^*_{z_1}) = \Delta E_{P^*}[Y \mid Z], \tag{1.4}$$

This equation gives the two pieces of information of $\theta(P^*)$. First, the sign of $\theta(P^*)$ is the same as $\Delta E_{P*}[Y \mid Z]$. Second, the absolute value of $\theta(P^*)$ is at least the absolute value of $\Delta E_{P*}[Y \mid Z]$. The following lemma summaries there two pieces.

Lemma 1.

$$\theta(P^*)\Delta E_{P*}[Y \mid Z] \geqslant 0$$

$$|\theta(P^*)| \geqslant |\Delta E_{P*}[Y \mid Z]|.$$

I derive a new implication from the exclusion restriction for the instrumental variable in order to obtain an upper bound on $\theta(P^*)$ in absolute value. To explain the new implication, I introduce the total variation distance. The total variation distance

$$TV(f_1, f_0) = \frac{1}{2}\int |f_1(x) - f_0(x)| d\mu_X(x)$$

is the distance between the distribution f_1 and f_0. In Figure 1.13, the total variation distance is the half of the area for the shaded region. I use the total variation distance to evaluate the distributional effect of a binary variable, particularly the distributional effect of Z on (Y, T). The distributional effect of Z on (Y, T) reflects the dependency of $f_{(Y,T)|Z=z}(y, t)$ on z, and I interpret the total variation distance $TV(f_{(Y,T)|Z=z_1}, f_{(Y,T)|Z=z_0})$ as the magnitude of the distributional effect. Even when the variable X is discrete, I use the density f for X to represent the probability function for X.

The new implication is based on the exclusion restriction imposes that the instrumental variable has direct effect on the true treatment variable T^* and has indirect effect on the outcome variable Y and on the measured treatment variable T. The

new implication formalizes the idea that the magnitude of the direct effect of Z on T^* is no smaller than the magnitude of the indirect effect of Z on (Y,T).

Lemma 2. *Under Assumption 1, then*

$$\begin{aligned}&TV(f_{(Y,T)|Z=z_1}, f_{(Y,T)|Z=z_0})\\&= TV(f_{(Y_1,T_1)|T^*_{z_0}<T^*_{z_1}}, f_{(Y_0,T_0)|T^*_{z_0}<T^*_{z_1}})TV(f_{T^*|Z=z_1}, f_{T^*|Z=z_0})\end{aligned}$$

and therefore

$$TV(f_{(Y,T)|Z=z_1}, f_{(Y,T)|Z=z_0}) \leqslant TV(f_{T^*|Z=z_1}, f_{T^*|Z=z_0}) = P^*(T^*_{z_0} < T^*_{z_1}).$$

The new implication in Lemma 2 gives a lower bound on $P^*(T^*_{z_0} < T^*_{z_1})$ and therefore yields an upper bound on LATE in absolute value, combined with Eq. (1.4). Therefore, I use these relationships to derive an upper bound on LATE in absolute value, that is,

$$|\theta(P^*)| = \frac{|\Delta E_{P^*}[Y \mid Z]|}{P^*(T^*_{z_0} < T^*_{z_1})} \leqslant \frac{|\Delta E_{P^*}[Y \mid Z]|}{TV(f_{(Y,T)|Z=z_1}, f_{(Y,T)|Z=z_0})}$$

as long as $TV(f_{(Y,T)|Z=z_1}, f_{(Y,T)|Z=z_0}) > 0$.

The next theorem shows that the above observations characterize the sharp identified set for LATE.

Theorem 3. *Suppose that Assumption 1 holds, and consider an arbitrary data distribution P of (Y,T,Z). (i) The sharp identified set $\Theta_I(P)$ for LATE is included in $\Theta_0(P)$, where $\Theta_0(P)$ is the set of θ's which satisfies the following three inequalities.*

$$\begin{aligned}&\theta\Delta E_P[Y \mid Z] \geqslant 0\\&|\theta| \geqslant |\Delta E_P[Y \mid Z]|\\&|\theta|TV(f_{(Y,T)|Z=z_1}, f_{(Y,T)|Z=z_0}) \leqslant |\Delta E_P[Y \mid Z]|.\end{aligned}$$

(ii) If Y is unbounded, then $\Theta_I(P)$ is equal to $\Theta_0(P)$.

Corollary 4. *Consider an arbitrary data distribution P of (Y,T,Z). If $f_{(Y,T)|Z=z_1}$ and $f_{(Y,T)|Z=z_0}$ are different such that $TV(f_{(Y,T)|Z=z_1}, f_{(Y,T)|Z=z_0}) > 0$,*

$$\Theta_0(P) = \begin{cases} \left[\Delta E_P[Y \mid Z], \dfrac{\Delta E_P[Y \mid Z]}{TV(f_{(Y,T)|Z=z_1}, f_{(Y,T)|Z=z_0})}\right] & \text{if } \Delta E_P[Y \mid Z] > 0 \\ \{0\} & \text{if } \Delta E_P[Y \mid Z] = 0 \\ \left[\dfrac{\Delta E_P[Y \mid Z]}{TV(f_{(Y,T)|Z=z_1}, f_{(Y,T)|Z=z_0})}, \Delta E_P[Y \mid Z]\right] & \text{if } \Delta E_P[Y \mid Z] < 0. \end{cases}$$

If $TV(f_{(Y,T)|Z=z_1}, f_{(Y,T)|Z=z_0}) = 0$, then $\Theta_0(P) = \mathbb{R}$.

The total variation distance $TV(f_{(Y,T)|Z=z_1}, f_{(Y,T)|Z=z_0})$ measures the strength for the instrumental variable in my analysis, that is, $TV(f_{(Y,T)|Z=z_1}, f_{(Y,T)|Z=z_0}) > 0$ is the relevance condition in my identification analysis. $TV(f_{(Y,T)|Z=z_1}, f_{(Y,T)|Z=z_0}) = 0$ means that the instrumental variable Z does not affect Y and T, in which case Z has no identifying power for the local average treatment effect. When $f_{(Y,T)|Z=z_1}$ and $f_{(Y,T)|Z=z_0}$ are different, the interval in the above theorem is always nonempty and bounded, which implies that Z has some identifying power for the local average treatment effect.

The Wald estimand $\Delta E_P[Y \mid Z]/\Delta E_P[T \mid Z]$ can be outside the identified set. The inequality

$$\left|\frac{\Delta E_P[Y \mid Z]}{TV(f_{(Y,T)|Z=z_1}, f_{(Y,T)|Z=z_0})}\right| \leqslant \left|\frac{\Delta E_P[Y \mid Z]}{\Delta E_P[T \mid Z]}\right|$$

holds and a strict inequality holds unless the sign of $f_{(Y,T)|Z=z_1}(y,t) - f_{(Y,T)|Z=z_0}(y,t)$ is constant in y for every t. It might seem counter-intuitive that the Wald estimand equals to LATE without measurement error but that it is not necessarily in the

identified set $\Theta_0(P)$ when measurement error is allowed. Recall that my framework includes no measurement error as a special case. As in Balke and Pearl (1997) and Heckman and Vytlacil (2005), the the LATE framework has the testable implications:

$$f_{(Y,T)|Z=z_1}(y,1) \geqslant f_{(Y,T)|Z=z_0}(y,1) \text{ and } f_{(Y,T)|Z=z_1}(y,0) \leqslant f_{(Y,T)|Z=z_0}(y,0).$$

When the data distribution does not satisfy the testable implications and there is no measurement error on the treatment variable, the identified set for LATE becomes empty and, therefore, the Wald estimand is no longer equal to LATE anymore. My framework has no testable implications, because the identified set is always non-empty. The recent papers by Huber and Mellace (2015), Kitagawa (2014) and Mourifié and Wan (2014) propose the testing procedures for the testable implications.

1.4 Inference

Having derived what can be identified about LATE under Assumption 1, this section considers statistical inference about LATE. I construct a confidence interval for LATE based on $\Theta_0(P)$ in Theorem 3. There are two difficulties with directly using Theorem 3 for statistical inference. First, as is often the case in the partially identified models, the length of the interval is unknown ex ante, which causes uncertainty of how many moment inequalities are binding for a given value of the parameter. Second, the identified set depends on the total variation distance which involves absolute values of the data distribution. I cannot apply the delta method to derive the asymptotic distribution for the total variation distance, because of the failure of differentiability (Hirano and Porter, 2012, Fang and Santos, 2014). This non-differentiability problem remains even if the support of Y is finite.

I will take three steps to construct a confidence interval. In the first step, I use the supremum representation of the total variation distance to characterize the identified set $\Theta_0(P)$ by the moment inequalities with differentiable moment functions. When the outcome variable Y has a finite support, I can apply methodologies developed for a finite number of moment inequalities (e.g., Andrews and Soares, 2010) to construct a confidence interval for LATE. When the outcome variable Y has an infinite support, however, none of the existing methods can be directly applied because the moment inequalities are not continuously indexed by a compact subset of the finite dimensional space. In the second step, therefore, I discretize the support of Y to make the number of the moment inequalities to be finite in the finite sample. I let the discretization finer as the sample size goes to the infinity, such that eventually the approximation error from the discretization vanishes. The number of the moment inequalities become finite but growing, particularly diverging to the infinity when Y has an infinite support. This structure resembles many moment inequalities in Chernozhukov et al. (2014). The third step is to adapt a bootstrapped critical value construction in Chernozhukov et al. (2014) to my framework.

1.4.1 Supremum representation of the total variation distance

In order to avoid the non-differentiability problem, I characterize the identified set by the moment inequalities which are differentiable with respect to the data distribution.

Lemma 5. *Let P be an arbitrary data distribution of (Y, T, Z).*

1. *Let $\mathbf{Y}$ be the support for the random variable Y and $\mathbf{T} \equiv \{0, 1\}$ be the support for T. Denote by $\mathbf{H}$ the set of measurable functions on $\mathbf{Y} \times \mathbf{T}$ taking a value*

in $\{-1, 1\}$. *Then*

$$TV(f_{(Y,T)|Z=z_1}, f_{(Y,T)|Z=z_0}) = \sup_{h \in \mathbf{H}} \Delta E_P[h(Y,T) \mid Z]/2$$

2. *The identified set* $\Theta_0(P)$ *for LATE is the set of* θ*'s which satisfy the following conditions*

$$-(1\{\theta \geqslant 0\} - 1\{\theta < 0\})\Delta E_P[Y \mid Z] \leqslant 0$$
$$(1\{\theta \geqslant 0\} - 1\{\theta < 0\})\Delta E_P[Y \mid Z] - |\theta| \leqslant 0$$
$$|\theta|\Delta E_P[h(Y,T) \mid Z]/2 - (1\{\theta \geqslant 0\} - 1\{\theta < 0\})\Delta E_P[Y \mid Z] \leqslant 0$$
$$\textit{for every } h \in \mathbf{H}.$$

The number of elements $\mathbf{H}$ can be large; $\mathbf{H}$ is an infinite set when Y is continuous, and $\mathbf{H}$ has the same elements of the power set of $\mathbf{Y} \times \mathbf{T}$ when Y takes only finite values.

1.4.2 Discretizing the outcome variable

To make the inference problem statistically and computationally feasible, I discretize the support for Y and make the number of the moment inequalities finite. Consider a partition $\mathbf{I}_n = \{I_{n,1}, \ldots, I_{n,K_n}\}$ over $\mathbf{Y}$, in which $I_{n,k}$ depends on n, and K_n can grow with sample size. Let $h_{n,j}$ be a generic function of $\mathbf{Y} \times \mathbf{T}$ into $\{-1, 1\}$ that is constant over $I_{n,k} \times \{t\}$ for every $k = 1, \ldots, K_n$ and every $t = 0, 1$. Let $\{h_{n,1}, \ldots, h_{n,4^{K_n}}\}$ be the set of all such functions. Note that $\{h_{n,1}, \ldots, h_{n,4^{K_n}}\}$ is a subset of $\mathbf{H}$. Using these $h_{n,j}$'s, I consider the following set $\Theta_n(P)$ characterized by the moment inequalities

$$\Theta_n(P) = \{\theta \in \Theta : \Delta E_P[g_{Z,j}((Y,T),\theta) \mid Z] \leqslant 0 \text{ for every } 1 \leqslant j \leqslant p_n\}$$

where $p_n = 4^{K_n} + 2$ is the number of the moment inequalities, and

$$\begin{aligned} g_{z,j}((y,t),\theta) &= |\theta| h_{n,j}(y,t)/2 - (1\{\theta \geqslant 0\} - 1\{\theta < 0\})y \\ &\qquad\qquad\qquad\qquad \text{for every } j = 1, \ldots, 4^{K_n} \\ g_{z,4^{K_n}+1}((y,t),\theta) &= (1\{\theta \geqslant 0\} - 1\{\theta < 0\})y - 1\{z = z_1\}|\theta| \\ g_{z,4^{K_n}+2}((y,t),\theta) &= -(1\{\theta \geqslant 0\} - 1\{\theta < 0\})y \end{aligned}$$

for every $z = z_0, z_1$. That the set $\Theta_n(P)$ is an outer identified set. That is, it is a superset of the identified set $\Theta_0(P)$. The next subsections consider consistency with respect to the identified set $\Theta_0(P)$ by letting $\Theta_n(P)$ converge to $\Theta_0(P)$. This point is different than the usual use of the outer identified set and is similar to sieve estimation. To clarify the convergence of $\Theta_n(P)$ to $\Theta_0(P)$, I call $\Theta_n(P)$ the *approximated identified set*.

1.4.3 Confidence interval for LATE

The approximated identified set $\Theta_n(P)$ consists of a finite number of moments inequalities, but the number of moment inequalities depends on the sample size. As Section 1.4.4 requires, the number of moment inequalities $p_n = 4^{K_n} + 2$ needs to diverge in order to obtain the consistency for the confidence interval when Y is continuous. The approximated identified set is defined to converge to the sharp identified set, so that the confidence interval in the present paper exhausts all the information in the large sample. As in Subsection 1.4.4, the confidence interval has asymptotic power 1 against all the fixed alternatives outside the sharp identified set.

This divergent number of the moment inequalities in the approximated identified set resembles the identified set in Chernozhukov et al. (2014), who considers testing a

growing number of moment inequalities in which each moment inequalities are based on different random variables. I modify their methodology into the two sample framework where one sample is the group with $Z = z_1$ and the other sample is the group with $Z = z_0$. For simplicity, I assume that Z is deterministic, which makes the notation in the following analysis simpler. The assumption of deterministic Z yields two independent samples: $(Y_{z_0,1}, T_{z_0,1}), \ldots, (Y_{z_0,n_0}, T_{z_0,n_0})$ are the observations with $Z = z_0$ and $(Y_{z_1,1}, T_{z_1,1}), \ldots, (Y_{z_1,n_1}, T_{z_1,n_1})$ are the observations with $Z = z_1$. n_0 is the sample size for the observations with $Z = z_0$ and n_1 is for $Z = z_1$. The total sample size is $n = n_0 + n_1$. I assume $n_1/n_0, n_0/n_1 = o(1)$.

In order to discuss a test statistic and a critical value, I introduce estimators for the moment functions and the estimated standard deviations for the moment functions. For an estimator for the moment function,

$$\hat{m}_j(\theta) = \hat{m}_{1,j}(\theta) - \hat{m}_{0,j}(\theta)$$

estimates the jth moment function $m_j(\theta) = \Delta E_P[g_{Z,j}((Y,T),\theta) \mid Z]$, where

$$\hat{m}_{z,j}(\theta) = n_z^{-1} \sum_{i=1}^{n_z} g_{z,j}((Y_{z,i}, T_{z,i}), \theta).$$

Denote by $\sigma_{z,j}(\theta)$ the standard deviation of $n_z^{-1/2} \sum_{i=1}^{n_z} g_{z,j}((Y_{z,i}, T_{z,i}), \theta)$. The standard deviation $\sigma_j(\theta)$ for $\sqrt{n}\hat{m}_j(\theta)$ is $\sigma_j^2(\theta) = n^{-1}(n_1\sigma_{1,j}^2(\theta) + n_0\sigma_{0,j}^2(\theta))$. Denote by $\hat{\sigma}_{z,j}(\theta)$ the estimated standard deviation of $n_z^{-1/2} \sum_{i=1}^{n_z} g_{z,j}((Y_{z,i}, T_{z,i}), \theta)$, that is,

$$\hat{\sigma}_{z,j}(\theta) = n_z^{-1} \sum_{i=1}^{n_z} (g_{z,j}((Y_{z,i}, T_{z,i}), \theta) - \hat{m}_{z,j}(\theta))^2.$$

Algorithm 1 Two-step multiplier bootstrap (Chernozhukov et al., 2014)

1: For each $z = 0, 1$, generate independent random variables $\epsilon_{z,1}, \ldots, \epsilon_{z,n_z}$ from $N(0, 1)$.
2: Construct the bootstrap test statistics for the moment inequality selection by

$$W(\theta) = \max_{1 \leqslant j \leqslant p_n} \frac{\sqrt{n}\hat{m}_j^B(\theta)}{\max\{\hat{\sigma}_j(\theta), \xi\}},$$

where we use $\hat{m}_{z,j}^B(\theta) = n_z^{-1} \sum_{i=1}^{n_z} \epsilon_{z,i} \left(g_{z,j}((Y_{z,i}, T_{z,i}), \theta) - \hat{m}_{z,j}(\theta)\right)$
3: and $\hat{m}_j^B(\theta) = \hat{m}_{1,j}^B(\theta) - \hat{m}_{0,j}^B(\theta)$.
4: Construct the bootstrap critical value $c(\beta, \theta)$ for the moment inequality selection as the conditional $(1 - \beta)$-quantile of $W(\theta)$ given $\{(Y_{z,i}, T_{z,i})\}$.
5: Select the moment inequalities and save

$$\hat{J} = \left\{ j = 1, \ldots, p_n : \frac{\sqrt{n}\hat{m}_j(\theta)}{\max\{\hat{\sigma}_j(\theta), \xi\}} > -2c(\beta, \theta) \right\}.$$

6: Construct the bootstrap test statistics by

$$W_{\hat{J}}(\theta) = \max_{j \in \hat{J}} \frac{\sqrt{n}\hat{m}_j^B(\theta)}{\max\{\hat{\sigma}_j(\theta), \xi\}}$$

where $W_{\hat{J}}(\theta) = 0$ if $\hat{J}$ is empty.
7: Construct the bootstrap critical value $c^{2S}(\alpha, \theta)$ as the conditional $(1 - \alpha + 2\beta)$-quantile of $W_{\hat{J}}(\theta)$ given $\{(Y_{z,i}, T_{z,i})\}$.

$\hat{\sigma}_j(\theta)$ estimates the standard deviation $\sigma_j(\theta)$, that is,

$$\hat{\sigma}_j^2(\theta) = n^{-1}(n_1\hat{\sigma}_{1,j}^2(\theta) + n_0\hat{\sigma}_{0,j}^2(\theta)).$$

The test statistics for $\theta_0 = \theta$ is

$$T(\theta) = \max_{1 \leqslant j \leqslant p_n} \frac{\sqrt{n}\hat{m}_j(\theta)}{\max\{\hat{\sigma}_j(\theta), \xi\}}$$

where ξ is a small positive number which prohibits the fraction from becoming too large when the estimated standard deviation is near zero. The truncation via ξ

controls the effect of the approximation error from the approximated identified set on the power against local alternatives, as in Subsection 1.4.4. The size is $\alpha \in (0, 1/2)$ and the pretest size for the moment inequality selection is $\beta \in (0, \alpha/2)$. The critical value $c^{2S}(\alpha, \theta)$ for $T(\theta)$ is based on the two-step multiplier bootstrap (Chernozhukov et al., 2014), described in Algorithm 1. The $(1-\alpha)$-confidence interval for LATE is

$$\{\theta \in \Theta : T(\theta) \leqslant c^{2S}(\alpha, \theta)\}.$$

Under the following three assumptions, I show that this confidence uniformly valid asymptotic size control for the confidence interval, by adapting Theorem 4.4 in Chernozhukov et al. (2014) into the two independent samples.

Assumption 2. $\Theta \subset \mathbb{R}$ *is bounded.*

Assumption 3. *There are constants* $c_1 \in (0, 1/2)$ *and* $C_1 > 0$ *such that* $\log^{7/2}(p_n n) \leqslant C_1 n^{1/2-c_1}$ *with* $p_n = 4^{K_n} + 2$.

Assumption 4. *(i) There is a constant* $C_0 > 0$ *such that* $\max\{E[Y^3]^{2/3}, E[Y^4]^{1/2}\} < C_0$. *(ii)* $0 < \sigma_{z,j}(\theta) < \infty$.

Theorem 6. *Under Assumptions 2 and 3,*

$$\liminf_{n\to\infty} \inf_{(\theta,P)\in\mathcal{H}_0} P(T(\theta) \leqslant c^{2S}(\alpha, \theta)) \geqslant 1 - \alpha,$$

where $\mathcal{H}_0$ *is the set of* (θ, P) *such that* P *satisfies Assumption 4 and* $\theta \in \Theta_0(P)$.

1.4.4 Power against fixed and local alternatives

This section discusses the power properties of the confidence interval. First, I assume that the density function $f_{(Y,T)|Z=z}$ satisfies the Hölder continuity. This assumption

justifies the approximation of the total variation distance via step functions, which is similar to the sieve estimation.

Assumption 5. *The density function $f_{(Y,T)|Z=z}$ is Hölder continuous in (y,t) with the Hölder constant D_0 and exponent d.*

I restrict the number of the moment inequalities and, in turn, restrict the magnitude of the critical value. Note that the number of the moment inequalities is the tuning parameter in this framework. The tradeoff is as follows: the approximation error is large if $p_n \to \infty$ slow, and the sampling error is large if $p_n \to \infty$ fast.

Assumption 6. $\log^{1/2}(p_n) \leqslant C_1 n^{1/2-c_1}$.

The last condition is that the grids in $\mathbf{I}_n$ becomes finer as the sample size goes to infinity.

Assumption 7. *There is a positive constant D_1 such that $I_{n,k}$ is a subset of some open ball with radius D_1/K_n in $\mathbf{Y} \times \mathbf{T}$.*

I obtain the following power property against local alternatives, based on Corollary 5.1 in Chernozhukov et al. (2014).

Theorem 7. *Fix $\delta, \epsilon > 0$ and $\tau_n \to \infty$ with $\tau_n = o(n)$. Denote by $\mathcal{H}_{1,n}$ the set of local alternatives (θ, P)'s satisfying Assumptions 4, 5 and at least one of the following inequalities:*

$$-\theta \Delta E_P[Y \mid Z] \geqslant \kappa_n \tag{1.5}$$

$$|\Delta E_P[Y \mid Z]| - |\theta| \geqslant \kappa_n \tag{1.6}$$

$$|\theta| TV(f_{(Y,T)|Z=z_1}, f_{(Y,T)|Z=z_0}) - |\Delta E_P[Y \mid Z]| \geqslant \kappa_n + \sup_{\theta \in \Theta} |\theta| 2^{d+2} D_0 D_1^d K_n^{-d} \mu_Y(\mathbf{Y}), \tag{1.7}$$

where $\kappa_n = \xi(1+\delta)(1+\epsilon)\sqrt{2\log(\max\{p_n, \tau_n\}/(\alpha - 2\beta))/n}$. *Under Assumptions 2 and 3, 6 and 7,*

$$\lim_{n\to\infty} \inf_{(\theta,P)\in\mathcal{H}_{1,n}} P(T(\theta) > c^{2S}(\alpha, \theta)) = 1.$$

The violation of the moment inequalities includes local alternatives in the sense that K_n^{-d} and κ_n go to zero in the large sample.

1.5 Monte Carlo simulations

This section illustrates the theoretical properties for the confidence interval in Section 1.7, using simulated datasets. Consider four independent random variables U_1, U_2, U_3, U_4 from $U(0,1)$. Using the $N(0,1)$ cumulative distribution function Φ, I generate (Y, T, Z) in the following way:

$$\begin{aligned}
Z &= 1\{U_1 \leqslant 0.5\} \\
T^* &= 1\{U_2 \leqslant 0.5 + \gamma_1(Z - 0.5)\} \\
Y &= \Phi\left(\gamma_2 T^* + \frac{\Phi^{-1}(U_4) + 0.5\Phi^{-1}(U_2)}{1 + 0.5^2}\right) \\
T &= \begin{cases} T^* & \text{if } U_3 \leqslant \gamma_3 \\ 1 - T^* & \text{otherwise.} \end{cases}
\end{aligned}$$

I have the three parameters in the model: γ_1 represents the strength of the instrumental variable, γ_2 represents the magnitude of treatment effect, and γ_3 represents the degree of the measurement error. This is the heterogeneous treatment effect model, because Φ is nonlinear. I select several values for $(\gamma_1, \gamma_2, \gamma_3)$ as in Table 1.1. The treatment effect is small ($\gamma_2 = 1$) in Designs 1-4 and large ($\gamma_1 = 3$) in Designs

5-8. The measurement error is small ($1 - \gamma_2 = 20\%$) in Designs 3,4,7,8 and large ($1 - \gamma_2 = 40\%$) in Designs 1,2,5,6. The instrumental variable is strong ($\gamma_1 = 0.5$) in Designs 2,4,6,8 and weak ($\gamma_1 = 0.1$) in Design 1,3,5,7.

In Table 1.1, I compute the three population objects: LATE, the Wald estimand, and the sharp identified set for LATE. As expected, LATE is included in the sharp identified set in all the designs. The comparison between the Wald estimand and the sharp identified set hints that the Wald estimand is relatively large compared to the upper bound of the identified set $\Theta_0(P)$ when the measurement error has a large degree in Design 1,2,5,6. For those designs, the Wald estimand is too large to be interpreted as an upper bound on LATE, because the upper bound of the identified set $\Theta_0(P)$ is much smaller.

I choose the sample size $n = 500, 1000, 5000$ for the Monte Carlo simulations. Note that the numbers covers the sample size (2,909) in NLS-72. I simulate 2,000 datasets of three sample sizes. For each dataset, I construct the different confidence set with confidence size $1 - \alpha = 95\%$, as in Section 1.4. I use the partition of equally spaced grids over $\mathbf{Y}$ with the number the partitions $K_n = 1, 3, 5$. In all the confidence intervals, I use 5,000 bootstraps repetitions. Figures 1.3-1.10 describe the coverage probabilities of the confidence intervals for each parameter value.

For each design, two figures are displayed. First, the left figures demonstrate the coverage probabilities according to $n = 500, 100, 5000$ given $K_n = 3$. The left figures of all the designs support the consistency results in the previous section; as the sample size increases, the coverage probabilities of the confidence intervals accumulate over $\Theta_0(P)$. Second, the right figures demonstrate the coverage probabilities according to $K_n = 1, 3, 5$ given $n = 1000$. When the Wald estimand is close to the upper bound

of $\Theta_0(P)$(Design 3,4,7,8), it seems advantageous to use $K_n = 1$. It is presumably because $K_n = 3, 5$ uses more inequalities for inference compared to $K_n = 1$ but these inequalities are not informative for LATE. When the Wald estimand is significantly larger than the upper bound of $\Theta_0(P)$ (Design 1,2,5,6), it seems advantageous to use $K_n = 3, 5$, particularly for the coverage probabilities near the upper bound of $\Theta_0(P)$. In these designs, the coverage probabilities are not sensitive the choice of $K_n = 3$ or $K_n = 5$.

1.6 Empirical illustrations

To illustrate the theoretical results on identification and inference, this section uses the National Longitudinal Survey of the High School Class of 1972 (NLS-72) to investigate the effect on wages of attending a college when the college attendance can be mismeasured. Kane and Rouse (1995) and Kane et al. (1999) use the same dataset to investigate the educational effect on wages in the presence of the endogeneity and the measurement error in the educational attainments. However, they do not consider the two problems and their results are dependent on the constant return to schooling. For an instrument, I follow the strategy in Card (1995) and Kane and Rouse (1995) closely and use the proximity to college as an instrumental variable for the college attendance.

NLS-72 was conducted by the National Center for Education Statistics with the U.S. Department of Education, and it contains 22,652 seniors (as of 1972) from 1,200 schools across the U.S. The sampled individuals were asked to participate in multiple surveys from 1972 through 1986. The survey collects labor market experiences, schooling information and demographic characteristics. I drop the individuals with

college degree or more, to focus on the comparison between high school graduates and the individuals with some college education. I also drop those who have missing values for wages in 1986 or educational attainments. The resulting size is 2,909.

I consider the effect of the college attendance T^* on Y (the log of wages in 1986). The treatment group with $T^* = 1$ is the individuals who have attended a college without a degree, and the control group is the individuals who have never been to a college. Some summary statistics are on Tables 1.2 and 1.3. I allows for the possibility that T^* is mismeasured, that is, the college attendance T in the dataset can be different from the truth T^*. I define the instrumental variable Z as an indicator for whether an individual grew up near 4 year college. I use 10 miles as a threshold for the proximity-to-college to similar to the strategy in Card (1995).

I present inference results in Table 1.4. The first row is the Wald estimate and the 95% confidence interval for the Wald estimand. The second row is the 95% confidence interval for LATE based on the identified set $\Theta_0(P)$ in Theorem 3. For the calculation of this confidence interval, I use the partition of equally spaced grids over $\mathbf{Y}$ with the number of the partitions equal to $K_n = 3$. In all the confidence intervals, I use 5000 bootstrap repetitions. The results are consistent with my identification analysis in the following two points. First, the Wald estimate is too large for the effect of attending a college. For example, Card (1999) documents the existing estimates for the return to schooling and most of them fall in the range of 5-15% as the percentage increases for one additional year of education. According to my analysis, the large value of the Wald estimate can result from the mismeasurement of the college attendance. Second, when I compare the upper bounds of the confidence intervals for the Wald estimand and LATE, the upper bound (1.04) based on $\Theta_0(P)$

is strictly lower than that (1.61) of the Wald estimand. This implies that the Wald estimator is an upper bound for LATE but it does not offer the sharp upper bound for LATE. These two findings are still valid when I consider six subgroups (Table 1.5).

1.7 Identifying power of repeated measurements

This section explores the identifying power of repeated measurements. Repeated measurements (for example, Hausman et al., 1991) is a popular approach in the literature on measurement error, but they cannot be instrumental variables in this framework. This is because the true treatment variable T^* is endogenous and it is natural to suspect that a measurement of T^* is also endogenous. The more accurate the measurement is, the more likely it is to be endogenous. Nevertheless, the identification strategy of the present paper incorporates repeated measurements as an additional information to narrow the identified set for LATE, when they are coupled with the instrumental variable Z. Unlike the other paper on repeated measurements, I do not need to assume the independence of measurement errors among multiple measurements. The strategy of the present paper also benefits from having more than two measurements unlike Hausman et al. (1991) who achieve the point identification with two measurements.

Consider a second measurement R for T^*. I do not require that R is binary, so R can be discrete or continuous. Like $T = T_{T^*}$, I model R using the counterfactual outcome notations. R_1 is a counterfactual second measurement when the true variable T^* is 1, and R_0 is a counterfactual second measurement when the true variable

T^* is 0. Then the data generation of R is

$$R = R_{T^*}.$$

I assume that the instrumental variable Z is independent of R_{t^*} conditional on $(Y_{t^*}, T_{t^*}, T^*_{z_0}, T^*_{z_1})$.

Assumption 8. *(i) Z is independent of $(R_{t^*}, T_{t^*}, Y_{t^*}, T^*_{z_0}, T^*_{z_1})$ for each $t^* = 0, 1$. (ii) $T^*_{z_1} \geqslant T^*_{z_0}$ almost surely.*

Note that I do not assume the independence between R_{t^*} and T_{t^*}, where the independence between the measurement errors is a key assumption when the repeated measurement is an instrumental variable.

Under this assumption, I refine the identified set for LATE as follows.

Theorem 8. *Suppose that Assumption 8 holds, and consider an arbitrary data distribution P of (R, Y, T, Z). (i) The sharp identified set $\Theta_I(P)$ for LATE is included in $\Theta_0(P)$, where $\Theta_0(P)$ is the set of θ's which satisfies the following three inequalities.*

$$\theta \Delta E_P[Y \mid Z] \geqslant 0$$
$$|\theta| \geqslant |\Delta E_P[Y \mid Z]|$$
$$|\theta| TV(f_{(R,Y,T)|Z=z_1}, f_{(R,Y,T)|Z=z_0}) \leqslant |\Delta E_P[Y \mid Z]|.$$

(ii) If Y is unbounded, then $\Theta_I(P)$ is equal to $\Theta_0(P)$.

The total variation distance $TV(f_{(R,Y,T)|Z=z_1}, f_{(R,Y,T)|Z=z_0})$ in Theorem 8 is weakly larger than that in Theorem 3, which implies that the identified set in Theorem 8 is

weakly smaller than the identified set in Theorem 3:

$$
\begin{aligned}
&TV(f_{(R,Y,T)|Z=z_1}, f_{(R,Y,T)|Z=z_0}) \\
&\qquad = \frac{1}{2}\sum_{t=0,1}\iint |(f_{(R,Y,T)|Z=z_1} - f_{(R,Y,T)|Z=z_0})(r,y,t)|d\mu_R(r)d\mu_Y(y) \\
&\qquad \geqslant \frac{1}{2}\sum_{t=0,1}\int |\int (f_{(R,Y,T)|Z=z_1} - f_{(R,Y,T)|Z=z_0})(r,y,t)d\mu_R(r)|d\mu_Y(y) \\
&\qquad = \frac{1}{2}\sum_{t=0,1}\int |(f_{(Y,T)|Z=z_1} - f_{(Y,T)|Z=z_0})(y,t)|d\mu_Y(y) \\
&\qquad = TV(f_{(Y,T)|Z=z_1}, f_{(Y,T)|Z=z_0})
\end{aligned}
$$

and the strict inequality holds unless the sign of $(f_{(R,Y,T)|Z=z_1} - f_{(R,Y,T)|Z=z_0})(r,y,t)$ is constant in r for every (y,t). Therefore, it is possible to test whether the repeated measurement R has additional information, by testing whether the sign of $(f_{(R,Y,T)|Z=z_1} - f_{(R,Y,T)|Z=z_0})(r,y,t)$ is constant in r.

1.8 Conclusion

This paper studies the identifying power of instrumental variable in the heterogeneous treatment effect framework when a binary treatment variable is mismeasured and endogenous. The assumptions in this framework are the monotonicity of the instrumental variable Z on the true treatment variable T^* and the exogeneity of Z. I use the total variation distance to characterize the identified set for LATE parameter $E[T_1 - T_0 \mid T^*_{z_0} < T^*_{z_1}]$. I also provide an inference procedure for LATE. Unlike the existing literature on measurement error, the identification strategy does not reply on a specific assumption on the measurement error; the only assumption

on the measurement error is its independence of the instrumental variable. I apply the new methodology to study the return to schooling in the proximity-to-college instrumental variable regression using the NLS-72 dataset.

There are several directions for future research. First, the choice of the partition $\mathbf{I}_n$ in Section 1.4, particularly the choice of K_n, is an interesting direction. To the best of my knowledge, the literature on many moment inequalities has not investigated how econometricians choose the numbers of the many moment inequalities. Second, it is worthwhile to investigating the other parameter for the treatment effect. This paper has focused on the local average treatment effect (LATE) for the reasons mentioned in the introduction, but the literature on heterogeneous treatment effect has emphasized the choice of treatment effect parameter as an answer to relevant policy questions.

1.9 Proofs of Lemmas 1, 2, and 5

Proof of Lemma 1. Eq. (1.4) implies

$$\begin{aligned} \theta(P^*)\Delta E_{P^*}[Y \mid Z] &= \theta(P^*)^2 P^*(T^*_{z_0} < T^*_{z_1}) \\ &\geqslant 0 \end{aligned}$$

and

$$\begin{aligned} |\Delta E_{P^*}[Y \mid Z]| &= |\theta(P^*)P^*(T^*_{z_0} < T^*_{z_1})| \\ &\leqslant |\theta(P^*)|, \end{aligned}$$

because $0 \leqslant P^*(T^*_{z_0} < T^*_{z_1}) \geqslant 1$. □

Proof of Lemma 2. I obtain

$$f_{(Y,T)|Z=z_1} - f_{(Y,T)|Z=z_0} = P^*(T^*_{z_0} < T^*_{z_1})(f_{(Y_1,T_1)|T^*_{z_0}<T^*_{z_1}} - f_{(Y_0,T_0)|T^*_{z_0}<T^*_{z_1}})$$

by applying the same logic as Theorem 1 in Imbens and Angrist (1994):

$$\begin{aligned}
f_{(Y,T)|Z=z_0} &= P^*(T^*_{z_0} = T^*_{z_1} = 1 \mid Z = z_0) f_{(Y,T)|Z=z_0, T^*_{z_0}=T^*_{z_1}=1} \\
&\quad + P^*(T^*_{z_0} < T^*_{z_1} \mid Z = z_0) f_{(Y,T)|Z=z_0, T^*_{z_0}<T^*_{z_1}} \\
&\quad + P^*(T^*_{z_0} = T^*_{z_1} = 0 \mid Z = z_0) f_{(Y,T)|Z=z_0, T^*_{z_0}=T^*_{z_1}=0} \\
&= P^*(T^*_{z_0} = T^*_{z_1} = 1) f_{(Y_1,T_1)|T^*_{z_0}=T^*_{z_1}=1} \\
&\quad + P^*(T^*_{z_0} < T^*_{z_1}) f_{(Y_0,T_0)|T^*_{z_0}<T^*_{z_1}} \\
&\quad + P^*(T^*_{z_0} = T^*_{z_1} = 0) f_{(Y_0,T_0)|T^*_{z_0}=T^*_{z_1}=0} \\
f_{(Y,T)|Z=z_1} &= P^*(T^*_{z_0} = T^*_{z_1} = 1) f_{(Y_1,T_1)|T^*_{z_0}=T^*_{z_1}=1} \\
&\quad + P^*(T^*_{z_0} < T^*_{z_1}) f_{(Y_1,T_1)|T^*_{z_0}<T^*_{z_1}} \\
&\quad + P^*(T^*_{z_0} = T^*_{z_1} = 0) f_{(Y_0,T_0)|T^*_{z_0}=T^*_{z_1}=0}.
\end{aligned}$$

This implies

$$
\begin{aligned}
&TV(f_{(Y,T)|Z=z_1}, f_{(Y,T)|Z=z_0}) \\
&\quad = \frac{1}{2}\sum_{t=0,1}\int |f_{(Y,T)|Z=z_1}(y,t) - f_{(Y,T)|Z=z_0}(y,t)| d\mu_Y(y) \\
&\quad = \frac{1}{2}\sum_{t=0,1}\int |P^*(T^*_{z_0} < T^*_{z_1})(f_{(Y_1,T_1)|T^*_{z_0}<T^*_{z_1}}(y,t) - f_{(Y_0,T_0)|T^*_{z_0}<T^*_{z_1}}(y,t))| d\mu_Y(y) \\
&\quad = P^*(T^*_{z_0} < T^*_{z_1})\frac{1}{2}\sum_{t=0,1}\int |(f_{(Y_1,T_1)|T^*_{z_0}<T^*_{z_1}}(y,t) - f_{(Y_0,T_0)|T^*_{z_0}<T^*_{z_1}}(y,t))| d\mu_Y(y) \\
&\quad = P^*(T^*_{z_0} < T^*_{z_1}) TV(f_{(Y_1,T_1)|T^*_{z_0}<T^*_{z_1}}, f_{(Y_0,T_0)|T^*_{z_0}<T^*_{z_1}}) \\
&\quad \leqslant P^*(T^*_{z_0} < T^*_{z_1}),
\end{aligned}
$$

where the last inequality follows because the total variation distance is at most one. Moreover, since $T^*_{z_0} \leqslant T^*_{z_1}$ almost surely,

$$
\begin{aligned}
P^*(T^*_{z_0} < T^*_{z_1}), &= |P^*(T^*_{z_0} = 1) - P^*(T^*_{z_1} = 1)| \\
&= \frac{1}{2}|P^*(T^*_{z_0} = 1) - P^*(T^*_{z_1} = 1)| + \frac{1}{2}|P^*(T^*_{z_0} = 0) - P^*(T^*_{z_1} = 0)| \\
&= \frac{1}{2}\sum_{t^*=0,1} |f_{T^*|Z=z_1}(t^*) - f_{T^*|Z=z_0}(t^*)| \\
&= TV(f_{T^*|Z=z_1}, f_{T^*|Z=z_0}).
\end{aligned}
$$

□

Proof of Lemma 5. The lemma follows from

$$\begin{aligned}
\Delta E_P[h(Y,T) \mid Z] &= E_P[h(Y,T) \mid Z = z_1] - E_P[h(Y,T) \mid Z = z_0] \\
&= \sum_{t=0,1} \int h(y,t)(f_{(Y,T)|Z=z_1}(y,t) - f_{(Y,T)|Z=z_0}(y,t))d\mu_Y(y) \\
&= \sum_{t=0,1} \int h(y,t)\Delta f_{(Y,T)|Z}(y,t)d\mu_Y(y) \\
&\leqslant \sum_{t=0,1} \int |\Delta f_{(Y,T)|Z}(y,t)|d\mu_Y(y) \\
&= 2 \times TV(f_{(Y,T)|Z=z_1}, f_{(Y,T)|Z=z_0})
\end{aligned}$$

where the maximization is achieved if $h(y,t) = 1$ if $\Delta f_{(Y,T)|Z}(y,t) > 0$ and $h(y,t) = -1$ if $\Delta f_{(Y,T)|Z}(y,t) < 0$. □

1.10 Proofs of Theorems 3 and 8

Theorem 3 is a special case of Theorem 8 with R being constant, and therefore I demonstrate the proof only for Theorem 8. Lemma 2 is modified into the following lemma in the framework of Theorem 8.

Lemma 9. *Under Assumption 8, then*

$$TV(f_{(R,Y,T)|Z=z_1}, f_{(R,Y,T)|Z=z_0}) \leqslant P^*(T^*_{z_0} < T^*_{z_1}).$$

Proof. The proof is the same as Lemma 2 and this lemma follows from

$$f_{(R,Y,T)|Z=z_1} - f_{(R,Y,T)|Z=z_0} = P^*(T^*_{z_0} < T^*_{z_1})(f_{(R_1,Y_1,T_1)|T^*_{z_0}<T^*_{z_1}} - f_{(R_0,Y_0,T_0)|T^*_{z_0}<T^*_{z_1}}).$$

□

From Lemmas 1 and 9, all the three inequalities in Theorem 8 are satisfied when θ is the true value for LATE, which is the first part of Theorem 8. To prove Theorem 8, I am going to show the sharpness of the three inequalities, that is, that any point satisfying the three inequalities is generated by some data generating process P^* which is consistent with the data distribution P. I will consider two cases based on the value of $TV(f_{(R,Y,T)|Z=z_1}, f_{(R,Y,T)|Z=z_0})$.

1.10.1 Case 1: Zero total variation distance

Consider $TV(f_{(R,Y,T)|Z=z_1}, f_{(R,Y,T)|Z=z_0}) = 0$. In this case, $f_{(R,Y,T)|Z=z_1} = f_{(R,Y,T)|Z=z_0}$ almost everywhere over (r, y, t) and particularly $\Delta E_P[Y \mid Z] = 0$. Note that all the three inequalities in Theorem 8 have no restriction on θ in this case. For every $y \in \mathbb{R}$, consider the following two data generating processes. First, $P^*_{y,L}$ is defined by

$$
\begin{aligned}
Z &\sim P(Z = z) \\
(T^*_{z_0}, T^*_{z_1}) \mid Z &= \begin{cases} (0,1) & \text{with probability } P(Y > y) \\ (1,1) & \text{with probability } P(Y \leqslant y) \end{cases} \\
(R_1, Y_1, T_1) \mid (T^*_{z_0}, T^*_{z_1}, Z) &\sim f_{(R,Y,T)}(r, y, t) \\
(R_0, Y_0, T_0) \mid (T^*_{z_0}, T^*_{z_1}, Z) &\sim \begin{cases} f_{(R,Y,T)|Y>y}(r, y, t) & \text{if } T^*_{z_0} < T^*_{z_1} \\ f_{(R,Y,T)|Y\leqslant y}(r, y, t) & \text{if } T^*_{z_0} = T^*_{z_1} \end{cases}
\end{aligned}
$$

where (R_1, Y_1, T_1) and (R_0, Y_0, T_0) are conditionally independent of $(T^*_{z_0}, T^*_{z_1}, Z)$. Second, $P^*_{y,U}$ is defined by

$$\begin{aligned} Z &\sim P(Z = z) \\ (T^*_{z_0}, T^*_{z_1}) \mid Z &= \begin{cases} (0,1) & \text{with probability } P(Y < y) \\ (1,1) & \text{with probability } P(Y \geqslant y) \end{cases} \\ (R_1, Y_1, T_1) \mid (T^*_{z_0}, T^*_{z_1}, Z) &\sim f_{(R,Y,T)}(r, y, t) \\ (R_0, Y_0, T_0) \mid (T^*_{z_0}, T^*_{z_1}, Z) &\sim \begin{cases} f_{(R,Y,T)|Y<y}(r, y, t) & \text{if } T^*_{z_0} < T^*_{z_1} \\ f_{(R,Y,T)|Y\geqslant y}(r, y, t) & \text{if } T^*_{z_0} = T^*_{z_1} \end{cases} \end{aligned}$$

where (R_1, Y_1, T_1) and (R_0, Y_0, T_0) are conditionally independent of $(T^*_{z_0}, T^*_{z_1}, Z)$.

Lemma 10. *Consider the assumptions in Theorem 8. If $f_{(R,Y,T)|Z=z_1}$ and $f_{(R,Y,T)|Z=z_0}$ are different such that $TV(f_{(R,Y,T)|Z=z_1}, f_{(R,Y,T)|Z=z_0}) = 0$, then, for every $y \in \mathbb{R}$ and every $\pi \in [0,1]$,*

1. *the mixture distribution $\pi P^*_{y,L} + (1-\pi)P^*_{y,U}$ satisfies Assumption 8;*
2. *the mixture distribution $\pi P^*_{y,L} + (1-\pi)P^*_{y,U}$ generates the data distribution P;*
3. *under the mixture distribution $\pi P^*_{y,L} + (1-\pi)P^*_{y,U}$, LATE is equal to $E_P[Y] - \pi E_P[Y \mid Y > y] - (1-\pi)E_P[Y \mid Y < y]$.*

Proof. (1) Both $P^*_{y,L}$ and $P^*_{y,U}$ satisfy the independence between the instrumental variable Z and the variables $(R_{t^*}, T_{t^*}, Y_{t^*}, T^*_{z_0}, T^*_{z_1})$ for each $t^* = 0, 1$. Furthermore, $P^*_{y,L}$ and $P^*_{y,U}$ have the same marginal distribution for Z: $P(Z = z)$. Therefore, the mixture of $P^*_{y,L}$ and $P^*_{y,U}$ also satisfies the independence. Since the mixture of $P^*_{y,L}$ and $P^*_{y,U}$ satisfies $T^*_{z_1} \geqslant T^*_{z_0}$ almost surely, the first part of this lemma is established.

(2) The second part follows from the fact that both $P^*_{y,L}$ and $P^*_{y,U}$ generate the data distribution P. Since the proof is essentially the same for $P^*_{y,L}$ and $P^*_{y,U}$, I demonstrate it only for $P^*_{y,L}$. Denote by f^* the density function of $P^*_{y,L}$. Then

$$\begin{aligned}
f^*_{(R,Y,T)|Z=z_0}(r,y,t) &= P^*_{y,L}(T^*_{z_0} < T^*_{z_1}) f^*_{(R,Y,T)|T^*_{z_0}<T^*_{z_1},Z=z_0}(r,y,t) \\
&\quad + P^*_{y,L}(T^*_{z_0} = T^*_{z_1} = 1) f^*_{(R,Y,T)|T^*_{z_0}=T^*_{z_1}=1,Z=z_0}(r,y,t) \\
&= P^*_{y,L}(T^*_{z_0} < T^*_{z_1}) f^*_{(R_0,Y_0,T_0)|T^*_{z_0}<T^*_{z_1}}(r,y,t) \\
&\quad + P^*_{y,L}(T^*_{z_0} = T^*_{z_1} = 1) f^*_{(R_1,Y_1,T_1)|T^*_{z_0}=T^*_{z_1}=1}(r,y,t) \\
&= P(Y > y) f_{(R,Y,T)|Y>y}(r,y,t) \\
&\quad + P(Y \leqslant y) f_{(R,Y,T)|Y\leqslant y}(r,y,t) \\
&= f_{(R,Y,T)}(r,y,t) \\
f^*_{(R,Y,T)|Z=z_1}(r,y,t) &= f_{(R,Y,T)}(r,y,t),
\end{aligned}$$

where the last equality uses $Z = z_1$ implies $T^* = 1$.

(3) LATE under $P^*_{y,L}$ is

$$\begin{aligned}
E_{P^*_{y,L}}[Y_1 - Y_0 \mid T^*_{z_0} < T^*_{z_1}] &= E_{P^*_{y,L}}[Y_1 \mid T^*_{z_0} < T^*_{z_1}] - E_{P^*_{y,L}}[Y_0 \mid T^*_{z_0} < T^*_{z_1}] \\
&= E_P[Y] - E_P[Y \mid Y > y]
\end{aligned}$$

and LATE under $P^*_{y,U}$ is

$$\begin{aligned}
E_{P^*_{y,L}}[Y_1 - Y_0 \mid T^*_{z_0} < T^*_{z_1}] &= E_{P^*_{y,L}}[Y_1 \mid T^*_{z_0} < T^*_{z_1}] - E_{P^*_{y,L}}[Y_0 \mid T^*_{z_0} < T^*_{z_1}] \\
&= E_P[Y] - E_P[Y \mid Y < y].
\end{aligned}$$

They imply that LATE under the mixture distribution is equal to $E_P[Y] - \pi E_P[Y \mid Y > y] - (1-\pi) E_P[Y \mid Y < y]$. □

Now I will prove Theorem 8 for Case 1. Let θ be any real number. Since Y is unbounded, there are y_L and y_U with $y_L \leqslant E_P[Y]-\theta \leqslant y_U$ such that $P(Y < y_L) > 0$ and $P(Y > y_L) > 0$. Since

$$E_P[Y \mid Y < y_L] \leqslant E_P[Y] - \theta \leqslant E_P[Y \mid Y > y_U],$$

there is $\pi \in [0,1]$ such that

$$\theta = E_P[Y] - \pi E_P[Y \mid Y > y_U] - (1-\pi)E_P[Y \mid Y < y_L].$$

Using Lemma 10, the right hand side of the above equation is LATE under the mixture distribution $\pi P^*_{y_L,L} + (1-\pi)P^*_{y_U,U}$. This proves that θ is LATE under some data generating process which is consistent with the observed distribution P.

1.10.2 Case 2: Positive total variation distance

Consider $TV(f_{(R,Y,T)|Z=z_1}, f_{(R,Y,T)|Z=z_0}) > 0$. This means that the instrumental variable Z has non-zero indirect effect on (R,Y,T). Consider the following two data generating processes. First, P^*_L is defined by

$$\begin{aligned} Z &\sim P(Z=z) \\ (T^*_{z_0}, T^*_{z_1}) \mid Z &= (0,1) \\ (R_0,Y_0,T_0) \mid (T^*_{z_0}, T^*_{z_1}, Z) &\sim f_{(R,Y,T)|Z=z_0} \\ (R_1,Y_1,T_1) \mid (T^*_{z_0}, T^*_{z_1}, Z) &\sim f_{(R,Y,T)|Z=z_1} \end{aligned}$$

where (R_1,Y_1,T_1) and (R_0,Y_0,T_0) are conditionally independent of $(T^*_{z_0}, T^*_{z_1}, Z)$. Second, P^*_U defined as follows. Define

$$H = 1\{\Delta f_{(R,Y,T)|Z}(R,Y,T) \geqslant 0\} - 1\{\Delta f_{(R,Y,T)|Z}(R,Y,T) < 0\}.$$

and define P_U^* as

$$
\begin{aligned}
Z &\sim P(Z=z) \\
(T_{z_0}^*, T_{z_1}^*) \mid Z &= \begin{cases} (0,1) & \text{with probability } \Delta E_P[H \mid Z]/2 \\ (0,0) & \text{with probability } P(H=-1 \mid Z=z_1) \\ (1,1) & \text{with probability } P(H=1 \mid Z=z_0) \end{cases} \\
(R_1, Y_1, T_1) \mid (T_{z_0}^*, T_{z_1}^*, Z) &\sim \begin{cases} \frac{\Delta f_{(R,Y,T,H)|Z}(r,y,t,1)}{\Delta E_P[H|Z]/2} & \text{if } T_{z_0}^* < T_{z_1}^* \\ \text{any distribution} & \text{if } T_{z_0}^* = T_{z_1}^* = 0 \\ f_{(R,Y,T)|H=1,Z=z_0}(r,y,t) & \text{if } T_{z_0}^* = T_{z_1}^* = 1 \end{cases} \\
(R_0, Y_0, T_0) \mid (T_{z_0}^*, T_{z_1}^*, Z) &\sim \begin{cases} -\frac{\Delta f_{(R,Y,T,H)|Z}(r,y,t,-1)}{\Delta E_P[H|Z]/2} & \text{if } T_{z_0}^* < T_{z_1}^* \\ f_{(R,Y,T)|H=0,Z=z_1}(y,t) & \text{if } T_{z_0}^* = T_{z_1}^* = 0 \\ \text{any distribution} & \text{if } T_{z_0}^* = T_{z_1}^* = 1 \end{cases}
\end{aligned}
$$

where (R_1, Y_1, T_1) and (R_0, Y_0, T_0) are conditionally independent of $(T_{z_0}^*, T_{z_1}^*, Z)$.

Lemma 11. *Consider the assumptions in Theorem 8. If $f_{(R,Y,T)|Z=z_1}$ and $f_{(R,Y,T)|Z=z_0}$ are different such that $TV(f_{(R,Y,T)|Z=z_1}, f_{(R,Y,T)|Z=z_0}) > 0$, then*

1. *P_L^* generates the data distribution P and LATE under P_L^* is equal to $\Delta E_P[Y \mid Z]$; and*

2. *P_U^* generates the data distribution P and LATE under P_U^* is equal to*

$$\Delta E_P[Y \mid Z]/TV(f_{(R,Y,T)|Z=z_1}, f_{(R,Y,T)|Z=z_0}).$$

Proof. (1) Denote by f^* the density function of P_L^*. P_L^* generates the data distribu-

tion P:

$$\begin{aligned}
f^*_{(R,Y,T)|Z=z_0}(r,y,t) &= f^*_{(R,Y,T)|T^*=0,Z=z_0}(r,y,t) \\
&= f^*_{(R_0,Y_0,T_0)|T^*=0,Z=z_0}(r,y,t) \\
&= f_{(R,Y,T)|Z=z_0}(r,y,t) \\
f^*_{(R,Y,T)|Z=z_1}(r,y,t) &= f^*_{(R,Y,T)|T^*=1,Z=z_1}(r,y,t) \\
&= f^*_{(R_1,Y_1,T_1)|T^*=1,Z=z_1}(r,y,t) \\
&= f_{(R,Y,T)|Z=z_1}(r,y,t)
\end{aligned}$$

where the first equality uses $T^* = 1\{Z = z_1\}$. Under P^*_L, LATE is equal to $\Delta E_P[Y \mid Z]$:

$$\begin{aligned}
E_{P^*_L}[Y_1 - Y_0 \mid T^*_{z_0} < T^*_{z_1}] &= E_{P^*_L}[Y_1] - E_{P^*_L}[Y_0] \\
&= E_P[Y \mid Z = z_1] - E_P[Y \mid Z = z_0] \\
&= \Delta E_P[Y \mid Z].
\end{aligned}$$

(2) Denote by f^* the density function of P^*_U. When $\Delta E_P[Y \mid Z] \neq 0$, the density function $f^*_{(R_{t^*},Y_{t^*},T_{t^*})|(T^*_{z_0},T^*_{z_1})}$ is positive on :

$$\begin{aligned}
\Delta f_{(R,Y,T,H)|Z}(r,y,t,1) &= \Delta f_{(R,Y,T)|Z}(r,y,t) 1\{\Delta f_{(R,Y,T)|Z}(r,y,t) \geqslant 0\} \\
&\geqslant 0 \\
\Delta f_{(R,Y,T,H)|Z}(r,y,t,-1) &= \Delta f_{(R,Y,T)|Z}(r,y,t) 1\{\Delta f_{(R,Y,T)|Z}(r,y,t) < 0\} \\
&< 0.
\end{aligned}$$

P_U^* generates the data distribution P:

$$\begin{aligned}
f^*_{(R,Y,T)|Z=z_0}(r,y,t) &= P^*_U(T^*_{z_0} < T^*_{z_1} \mid Z = z_0) f^*_{(R,Y,T)|T^*_{z_0}<T^*_{z_1},Z=z_0}(r,y,t) \\
&\quad + P^*_U(T^*_{z_1} = T^*_{z_0} = 0 \mid Z = z_0) f^*_{(R,Y,T)|T^*_{z_1}=T^*_{z_0}=0,Z=z_0}(r,y,t) \\
&\quad + P^*_U(T^*_{z_1} = T^*_{z_0} = 1 \mid Z = z_0) f^*_{(R,Y,T)|T^*_{z_1}=T^*_{z_0}=1,Z=z_0}(r,y,t) \\
&= P^*_U(T^*_{z_0} < T^*_{z_1}) f^*_{(R_0,Y_0,T_0)|T^*_{z_0}<T^*_{z_1}}(r,y,t) \\
&\quad + P^*_U(T^*_{z_1} = T^*_{z_0} = 0) f^*_{(R_0,Y_0,T_0)|T^*_{z_1}=T^*_{z_0}=0}(r,y,t) \\
&\quad + P^*_U(T^*_{z_1} = T^*_{z_0} = 1) f^*_{(R_1,Y_1,T_1)|T^*_{z_1}=T^*_{z_0}=1}(r,y,t) \\
&= -\frac{\Delta E_P[H \mid Z]}{2} \frac{\Delta f_{(R,Y,T,H)|Z}(r,y,t,-1)}{\Delta E_P[H \mid Z]/2} \\
&\quad + P(H = -1 \mid Z = z_1) f_{(R,Y,T)|H=-1,Z=z_1}(r,y,t) \\
&\quad + P(H = 1 \mid Z = z_0) f_{(R,Y,T)|H=1,Z=z_0}(r,y,t) \\
&= f_{(R,Y,T)|Z=z_0}(r,y,t).
\end{aligned}$$

$$
\begin{aligned}
f^*_{(R,Y,T)|Z=z_1}(r,y,t) &= P^*_U(T^*_{z_0} < T^*_{z_1} \mid Z = z_1) f^*_{(R,Y,T)|T^*_{z_0}<T^*_{z_1},Z=z_1}(r,y,t) \\
&\quad + P^*_U(T^*_{z_1} = T^*_{z_0} = 0 \mid Z = z_1) f^*_{(R,Y,T)|T^*_{z_1}=T^*_{z_0}=0,Z=z_1}(r,y,t) \\
&\quad + P^*_U(T^*_{z_1} = T^*_{z_0} = 1 \mid Z = z_1) f^*_{(R,Y,T)|T^*_{z_1}=T^*_{z_0}=1,Z=z_1}(r,y,t) \\
&= P^*_U(T^*_{z_0} < T^*_{z_1}) f^*_{(R_1,Y_1,T_1)|T^*_{z_0}<T^*_{z_1}}(r,y,t) \\
&\quad + P^*_U(T^*_{z_1} = T^*_{z_0} = 0) f^*_{(R_0,Y_0,T_0)|T^*_{z_1}=T^*_{z_0}=0}(r,y,t) \\
&\quad + P^*_U(T^*_{z_1} = T^*_{z_0} = 1) f^*_{(R_1,Y_1,T_1)|T^*_{z_1}=T^*_{z_0}=1}(r,y,t) \\
&= \Delta E_P[H \mid Z]/2 \frac{\Delta f_{(R,Y,T,H)|Z}(r,y,t,1)}{\Delta E_P[H \mid Z]/2} \\
&\quad + P(H = -1 \mid Z = z_1) f_{(R,Y,T)|H=-1,Z=z_1}(r,y,t) \\
&\quad + P(H = 1 \mid Z = z_0) f_{(R,Y,T)|H=1,Z=z_0}(r,y,t) \\
&= f_{(R,Y,T)|Z=z_1}(r,y,t).
\end{aligned}
$$

Under P^*_U, LATE is equal to $\Delta E_P[Y \mid Z]/TV(f_{(R,Y,T)|Z=z_1}, f_{(R,Y,T)|Z=z_0})$:

$$
\begin{aligned}
&\Delta E_{P^*_U}[H \mid Z]/2 \\
&\quad = \frac{1}{2} \sum_{t=0,1} \iint (1\{\Delta f_{(R,Y,T)|Z}(r,y,t) \geqslant 0\} - 1\{\Delta f_{(R,Y,T)|Z}(r,y,t) < 0\}) \\
&\qquad \times \Delta f_{(R,Y,T)|Z}(r,y,t) d\mu_Y(y) d\mu_R(r) \\
&\quad = \frac{1}{2} \sum_{t=0,1} \iint |\Delta f_{(R,Y,T)|Z}(r,y,t)| d\mu_Y(y) d\mu_R(r) \\
&\quad = TV(f_{(R,Y,T)|Z=z_1}, f_{(R,Y,T)|Z=z_0})
\end{aligned}
$$

$$
\begin{aligned}
&E_{P_U^*}[Y_1 - Y_0 \mid T_{z_0}^* < T_{z_1}^*] \\
&\quad = \sum_{t=0,1} \iint y \left(\frac{\Delta f_{(R,Y,T,H)|Z}(r,y,t,1)}{\Delta E_P[H \mid Z]/2} + \frac{\Delta f_{(R,Y,T,H)}(y,t,-1)}{\Delta E_P[H \mid Z]/2} \right) d\mu_Y(y) d\mu_R(r) \\
&\quad = \sum_{t=0,1} \iint y \frac{\Delta f_{(R,Y,T)|Z}(r,y,t)}{\Delta E_P[H \mid Z]/2} d\mu_Y(y) d\mu_R(r) \\
&\quad = \frac{\Delta E_P[Y \mid Z]}{\Delta E_P[H \mid Z]/2} \\
&\quad = \frac{\Delta E_P[Y \mid Z]}{TV(f_{(R,Y,T)|Z=z_1}, f_{(R,Y,T)|Z=z_0})}.
\end{aligned}
$$

□

Lemma 12. *Consider the assumptions in Theorem 8.* $f_{(R,Y,T)|Z=z_1}$ *and* $f_{(R,Y,T)|Z=z_0}$ *are different such that* $TV(f_{(R,Y,T)|Z=z_1}, f_{(R,Y,T)|Z=z_0}) > 0$*, then, for every* $\pi \in [0,1]$*,*

1. *the mixture distribution* $\pi P_L^* + (1-\pi) P_U^*$ *satisfies Assumption 1;*
2. *the mixture distribution* $\pi P_L^* + (1-\pi) P_U^*$ *generates the data distribution* P*;*
3. *under the mixture distribution* $\pi P_L^* + (1-\pi) P_U^*$*, LATE is equal to*

$$
\pi \Delta E_P[Y \mid Z] + (1-\pi) \frac{\Delta E_P[Y \mid Z]}{TV(f_{(R,Y,T)|Z=z_1}, f_{(R,Y,T)|Z=z_0})}.
$$

Proof. The proof for this lemma is the same as the proof in Lemma 9. □

Now I will prove Theorem 8 for Case 2. Let θ be any real number satisfying all the three inequalities in Theorem 8. Then there is $\pi \in [0,1]$ such that

$$
\theta = \pi \Delta E_P[Y \mid Z] + (1-\pi) \frac{\Delta E_P[Y \mid Z]}{TV(f_{(R,Y,T)|Z=z_1}, f_{(R,Y,T)|Z=z_0})}.
$$

Using Lemma 12, the right hand side of the above equation is LATE under the mixture distribution $\pi P_L^* + (1-\pi)P_U^*$. This proves that θ is LATE under some data generating process which is consistent with the observed distribution P.

1.11 Proofs of Theorems 6 and 7

These results are obtained by applying the results in Chernozhukov et al. (2014) for two independent samples. I use the same notation as them and modify their proofs to the two samples. First, I introduce the following lemma.

Lemma 13. *(1) If x_1, x_2 are p-dimensional random variables and $0 < x_{2,j} \leqslant 1$, then*

$$\left|\max_{1\leqslant j\leqslant p} x_{1,j}x_{2,j}\right| \leqslant \left|\max_{1\leqslant j\leqslant p} x_{1,j}\right|.$$

(2) If $X_1, \bar{X}_1, X_2, \bar{X}_2$ are p-dimensional random variables and if the pair of random variables $(X_1, \bar{X}_1)$ and $(X_2, \bar{X}_2)$ are independent, then

$$\begin{aligned}\sup_{\mathbf{t}\in\mathbb{R}^p} |P(X_1 + X_2 \leqslant \mathbf{t}) - P(\bar{X}_1 + \bar{X}_2 \leqslant \mathbf{t})| \quad &\leqslant \quad \sup_{\mathbf{t}\in\mathbb{R}^p} |P(X_1 \leqslant \mathbf{t}) - P(\bar{X}_1 \leqslant \mathbf{t})| \\ &\quad + \sup_{\mathbf{t}\in\mathbb{R}^p} |P(X_2 \leqslant \mathbf{t}) - P(\bar{X}_2 \leqslant \mathbf{t})|.\end{aligned}$$

(3) If x_1, x_2 are p-dimensional random variables and $x_{2,j} > 0$, then

$$|\max_{1\leqslant j\leqslant p} x_{1,j} - \max_{1\leqslant j\leqslant p} x_{1,j}x_{2,j}| \leqslant \left|\max_{1\leqslant j\leqslant p} x_{1,j}\right| \max_{1\leqslant j\leqslant p} |x_{2,j} - 1|.$$

Proof. The first statement is as follows. If $\max_{1\leqslant j\leqslant p} x_{1,j} \geqslant 0$, then

$$\left|\max_{1\leqslant j\leqslant p} x_{1,j}x_{2,j}\right| = \max_{1\leqslant j\leqslant p} x_{1,j}1\{x_{1,j} > 0\}x_{2,j} \leqslant \max_{1\leqslant j\leqslant p} x_{1,j}1\{x_{1,j} > 0\} = \max_{1\leqslant j\leqslant p} x_{1,j}.$$

If $\max_{1\leqslant j\leqslant p} x_{1,j} < 0$, then

$$\left|\max_{1\leqslant j\leqslant p} x_{1,j}x_{2,j}\right| = \min_{1\leqslant j\leqslant p}(-x_{1,j})x_{2,j} \leqslant \min_{1\leqslant j\leqslant p}(-x_{1,j}) = -\max_{1\leqslant j\leqslant p} x_{1,j}.$$

The second statement is as follows.

$$\begin{aligned}
&\sup_{\mathbf{t}\in\mathbb{R}^p} |P(X_1 + X_2 \leqslant \mathbf{t}) - P(\bar{X}_1 + \bar{X}_2 \leqslant \mathbf{t})| \\
&\quad\leqslant \sup_{\mathbf{t}\in\mathbb{R}^p} |P(X_1 + X_2 \leqslant \mathbf{t}) - P(\bar{X}_1 + X_2 \leqslant \mathbf{t})| \\
&\qquad + \sup_{\mathbf{t}\in\mathbb{R}^p} |P(\bar{X}_1 + X_2 \leqslant \mathbf{t}) - P(\bar{X}_1 + \bar{X}_2 \leqslant \mathbf{t})| \\
&\quad= \sup_{\mathbf{t}\in\mathbb{R}^p} |E[P(X_1 + X_2 \leqslant \mathbf{t} \mid X_2) - P(\bar{X}_1 + X_2 \leqslant \mathbf{t} \mid X_2)]| \\
&\qquad + \sup_{\mathbf{t}\in\mathbb{R}^p} |E[P(\bar{X}_1 + X_2 \leqslant \mathbf{t} \mid \bar{X}_1) - P(\bar{X}_1 + \bar{X}_2 \leqslant \mathbf{t} \mid \bar{X}_1)]| \\
&\quad\leqslant \sup_{\mathbf{t}\in\mathbb{R}^p} |P(X_1 \leqslant \mathbf{t}) - P(\bar{X}_1 \leqslant \mathbf{t})| \\
&\qquad + \sup_{\mathbf{t}\in\mathbb{R}^p} |P(X_2 \leqslant \mathbf{t}) - P(\bar{X}_2 \leqslant \mathbf{t})|.
\end{aligned}$$

The third statement is as follows. If $\max_{1\leqslant j\leqslant p} x_{1,j}x_{2,j} \geqslant \max_{1\leqslant j\leqslant p} x_{1,j} \geqslant 0$, then

$$\begin{aligned}
\left|\max_{1\leqslant j\leqslant p} x_{1,j} - \max_{1\leqslant j\leqslant p} x_{1,j}x_{2,j}\right| &= \left|\max_{1\leqslant j\leqslant p} x_{1,j}\right| \left(\max_{1\leqslant j\leqslant p} \frac{x_{1,j}1\{x_{1,j} > 0\}}{\max_{1\leqslant j\leqslant p} x_{1,j}1\{x_{1,j} > 0\}} x_{2,j} - 1\right) \\
&\leqslant \left|\max_{1\leqslant j\leqslant p} x_{1,j}\right| \left(\max_{1\leqslant j\leqslant p} x_{2,j} - 1\right).
\end{aligned}$$

If $0 > \max_{1\leqslant j\leqslant p} x_{1,j}x_{2,j} \geqslant \max_{1\leqslant j\leqslant p} x_{1,j}$, then

$$\begin{aligned}
\left|\max_{1\leqslant j\leqslant p} x_{1,j} - \max_{1\leqslant j\leqslant p} x_{1,j}x_{2,j}\right| &= \left|\max_{1\leqslant j\leqslant p} x_{1,j}\right| \left(1 - \min_{1\leqslant j\leqslant p} \frac{(-x_{1,j})}{\min_{1\leqslant j\leqslant p}(-x_{1,j})} x_{2,j}\right) \\
&\leqslant \left|\max_{1\leqslant j\leqslant p} x_{1,j}\right| \left(1 - \min_{1\leqslant j\leqslant p} x_{2,j}\right).
\end{aligned}$$

If $\max_{1\leqslant j\leqslant p} x_{1,j} \geqslant \max_{1\leqslant j\leqslant p} x_{1,j}x_{2,j} \geqslant 0$, then

$$\begin{aligned}
&\left|\max_{1\leqslant j\leqslant p} x_{1,j} - \max_{1\leqslant j\leqslant p} x_{1,j}x_{2,j}\right| \\
&= \left|\max_{1\leqslant j\leqslant p} x_{1,j}\right| \left(1 - \frac{\max_{1\leqslant j\leqslant p} x_{1,j}1\{x_{1,j}>0\}x_{2,j}}{\max_{1\leqslant j\leqslant p} x_{1,j}1\{x_{1,j}>0\}}\right) \\
&\leqslant \left|\max_{1\leqslant j\leqslant p} x_{1,j}\right| \left(1 - \frac{\max_{1\leqslant j\leqslant p} x_{1,j}1\{x_{1,j}>0\}\min_{1\leqslant j\leqslant p} x_{2,j}}{\max_{1\leqslant j\leqslant p} x_{1,j}1\{x_{1,j}>0\}}\right) \\
&= \left|\max_{1\leqslant j\leqslant p} x_{1,j}\right| \left(1 - \min_{1\leqslant j\leqslant p} x_{2,j}\right).
\end{aligned}$$

If $0 > \max_{1\leqslant j\leqslant p} x_{1,j} \geqslant \max_{1\leqslant j\leqslant p} x_{1,j}x_{2,j}$, then

$$\begin{aligned}
\left|\max_{1\leqslant j\leqslant p} x_{1,j} - \max_{1\leqslant j\leqslant p} x_{1,j}x_{2,j}\right| &= \left|\max_{1\leqslant j\leqslant p} x_{1,j}\right| \left(\frac{\max_{1\leqslant j\leqslant p}(-x_{1,j})x_{2,j}}{\max_{1\leqslant j\leqslant p}(-x_{1,j})} - 1\right) \\
&\leqslant \left|\max_{1\leqslant j\leqslant p} x_{1,j}\right| \left(\frac{\max_{1\leqslant j\leqslant p}(-x_{1,j})\max_{1\leqslant j\leqslant p} x_{2,j}}{\max_{1\leqslant j\leqslant p}(-x_{1,j})} - 1\right) \\
&= \left|\max_{1\leqslant j\leqslant p} x_{1,j}\right| \left(\max_{1\leqslant j\leqslant p} x_{2,j} - 1\right).
\end{aligned}$$

□

Now I modify Chernozhukov et al. (2014) for two independent sample. In order to simplify the notations, this section focuses on a fixed value of θ and omits θ in the following discussion. All the following results are uniformly valid in θ. Denote $X_{z,i,j} = g_{z,j}((Y_{z,i}, T_{z,i}), \theta)$, $\hat{\mu}_{z,j} = n_z^{-1}\sum_{i=1}^{n_z} X_{z,i,j}$ and $\hat{\mu}_j = \hat{\mu}_{1,j} - \hat{\mu}_{0,j}$. For every

$J \subset \{1, \ldots, p_n\}$, define

$$\begin{aligned}
T(J) &= \max_{j \in J} \frac{\sqrt{n}\hat{\mu}_j}{\max\{\hat{\sigma}_j, \xi\}} \\
W(J) &= \max_{j \in J} \frac{\sqrt{n}\hat{\mu}_j^B}{\max\{\hat{\sigma}_j, \xi\}} \\
\bar{T}(J) &= \max_{j \in J} \frac{\sqrt{n}(\hat{\mu}_j - E[\hat{\mu}_j])}{\max\{\hat{\sigma}_j, \xi\}} \\
T_0(J) &= \max_{j \in J} \frac{\sqrt{n}(\hat{\mu}_j - E[\hat{\mu}_j])}{\max\{\sigma_j, \xi\}} \\
\bar{W}^B(J) &= \max_{j \in J} \frac{\sqrt{n}\hat{\mu}_j^B}{\max\{\sigma_j, \xi\}}.
\end{aligned}$$

Denote by $c(\gamma, J)$ the conditional $(1-\gamma)$-quantile of $W(J)$ given $\{(Y_{z,i}, T_{z,i})\}$.

Denote by $\Sigma_{z,J}$ the variance-covariance matrix of $\{X_{z,j}\}_{j \in J}$. $\Sigma_J = n^{-1}(n_1\Sigma_{1,J} + n_0\Sigma_{0,J})$ is the variance-covariance matrix of $\{\sqrt{n}\hat{\mu}_j\}_{j \in J}$. Denote by

$$\Omega_{z,J} = \frac{n_z}{n} \max\{diag(\Sigma_J), \xi^2 I_{|J|}\}^{-1} \Sigma_{z,J}$$

where max is the element-wise maximum. Let U_1 and U_0 be $|J|$-dimensional independent normal random variable with

$$U_1 \sim N(0, \Omega_{1,J}) \text{ and } U_0 \sim N(0, \Omega_{0,J}).$$

Denote by $c_0(\gamma, J)$ the $(1-\gamma)$ quantile of $\max_{j\in J}(U_{1,j} - U_{0,j})$. Define

$$\begin{aligned}
\rho_{n,J} &= \sup_{t\in\mathbb{R}} \left| P(T_0(J) \leqslant t) - P(\max_{j\in J}(U_{1,j} - U_{0,j}) \leqslant t) \right| \\
\rho^B_{n,J} &= \sup_{t\in\mathbb{R}} \left| P(\bar{W}^B(J) \leqslant t \mid Data) - P(\max_{j\in J}(U_{1,j} - U_{0,j}) \leqslant t) \right| \\
\rho_{z,n,J} &= \sup_{\mathbf{t}\in\mathbb{R}^{|J|}} \left| P\left(\frac{\sqrt{n_z}(\hat{\mu}_{z,j} - E[\hat{\mu}_{z,j}])}{\max\{\sigma_j, \xi\}} \leqslant \mathbf{t}_j, \forall j \in J \right) - P(U_{z,j} \leqslant \mathbf{t}_j, \forall j \in J) \right| \\
\rho^B_{z,n,J} &= \sup_{\mathbf{t}\in\mathbb{R}^{|J|}} \left| P\left(\frac{\sqrt{n_z}\hat{\mu}^B_{z,j}}{\max\{\sigma_j, \xi\}} \leqslant \mathbf{t}_j, \forall j \in J \mid \{(Y_{z,i}, T_{z,i})\} \right) - P(U_{z,j} \leqslant \mathbf{t}_j, \forall j \in J) \right|.
\end{aligned}$$

Note that $\mathbf{t}$ is a $|J|$-dimensional vector in the definitions of $\rho_{z,n,J}$, and therefore I need to use the central limit and bootstrap theorems for hyper-rectangles, which is slightly different from Chernozhukov et al. (2014).

Note the following statements are taken from Chernozhukov et al. (2014) and Chernozhukov et al. (2015).

Lemma 14. *(1) There are positive numbers $c \in (0, 1/2)$ and $C > 0$ such that*

$$\rho_{z,n,J} \leqslant C n_z^{-c} \tag{1.8}$$

$$P(\rho^B_{z,n,J} < C\nu_{z,n}) \geqslant 1 - C n_z^{-c} \text{ with some } \nu_{z,n} = C n_z^{-c} \tag{1.9}$$

$$P\left(\max_{j\in J} \left| \frac{\sigma^2_{z,j}}{\hat{\sigma}^2_{z,j}} - 1 \right| > n_z^{-1/2+c_1/4} B^2_{z,n_z} \log(|J|) \right) \leqslant C n_z^{-c} \tag{1.10}$$

$$P\left(\max_{j\in J} \left| \frac{\sqrt{n_z}(\hat{\mu}_{z,j} - E[\hat{\mu}_{z,j}])}{\sigma_{z,j}} \right| > n_z^{c_1/4} \sqrt{\log(|J|)} \right) \leqslant C n_z^{-c}. \tag{1.11}$$

(2) The following inequalities hold:

$$E_P\left[\left|\max_{j\in J}\frac{\sqrt{n_z}\hat{\mu}^B_{z,j}}{\hat{\sigma}_{z,j}}\right| \mid Data\right] \leqslant \sqrt{2\log(2|J|)} \tag{1.12}$$

$$c_0(\gamma, J) \leqslant \sqrt{2\log(|J|)} + \sqrt{2\log(1/\gamma)} \tag{1.13}$$

$$P\left(\left|\max_{j\in J}(U_{1,j} - U_{0,j}) - c_0(\gamma, J)\right| \leqslant \epsilon\right) \leqslant 4\epsilon(\sqrt{\log(|J|)} + 1) \tag{1.14}$$

$$c(\gamma, J) \leqslant \sqrt{2\log(|J|)} + \sqrt{2\log(1/\gamma)}. \tag{1.15}$$

Proof. Note that

$$|g_{z,j}((Y_{z,i}, T_{z,i}), \theta) - E_P[g_{z,j}((Y_{z,i}, T_{z,i}), \theta)]| \leqslant |Y_{z,i} - E_P[Y_{z,i}]| + \max_{\theta\in\Theta}|\theta|.$$

By Assumptions 2, 3, and 4,

$$(M_{z,n_z,3}(\theta, P)^3 \vee M_{z,n_z,4}(\theta, P)^2 \vee B_{z,n_z}(\theta, P))^2 \log^{7/2}(p_n/n_z) \leqslant \frac{C_0C_1}{\xi^2} n_z^{1/2-c_1},$$

where

$$M_{z,n_z,k}(\theta, P) = \max_{1\leqslant j\leqslant p_n} E_P\left[\left|\frac{g_{z,j}((Y_{z,i}, T_{z,i}), \theta) - E_P[g_{z,j}((Y_{z,i}, T_{z,i}), \theta)]}{\max\{\sigma_{z,j}(\theta), \xi\}}\right|^k\right]^{1/k}$$

$$B_{z,n_z}(\theta, P) = \left(E_P\left[\max_{1\leqslant j\leqslant p_n}\left(\frac{g_{z,j}((Y_{z,i}, T_{z,i}), \theta) - E_P[g_{z,j}((Y_{z,i}, T_{z,i}), \theta)]}{\max\{\sigma_{z,j}(\theta), \xi\}}\right)^4\right]\right)^{1/4}.$$

This is the key assumption in Chernozhukov et al. (2014, Eq.(49)) and therefore we can borrow their results. The first statement is Theorem 2.1 in Chernozhukov et al. (2015). The second is Corollary 4.3 in Chernozhukov et al. (2015). The third and fourth are Step 3 of Theorem 4.3. The last two statements are Lemma A.4 in Chernozhukov et al. (2014). □

Lemma 14 yields the following lemma in the two sample setting.

Lemma 15.

$$\rho_{n,J} \leqslant C(n_1^{-c} + n_0^{-c})$$

$$P(\rho_{n,J}^B < \nu_{1,n} + \nu_{0,n})) \geqslant 1 - C(n_1^{-c} + n_0^{-c})$$

$$P\left(\max_{j\in J}\left|\frac{\max\{\sigma_j,\xi\}}{\max\{\hat{\sigma}_j,\xi\}} - 1\right| > (n_1^{-1/2+c_1/4} + n_0^{-1/2+c_1/4})B^2\log(p_n)\right) \leqslant C(n_1^{-c} + n_0^{-c})$$

$$P\left(\max_{j\in J}\left|\frac{\sqrt{n}(\hat{\mu}_j - E[\hat{\mu}_j])}{\sigma_j}\right| > \left(\frac{n}{n_1}n_1^{c_1/4} + \frac{n}{n_0}n_0^{c_1/4}\right)\sqrt{\log(|J|)}\right) \leqslant C(n_1^{-c} + n_0^{-c})$$

$$E_P\left[|W(J)| \mid Data\right] \leqslant 2\left(\frac{n}{n_1} + \frac{n}{n_0}\right)\sqrt{2\log(2|J|)}.$$

where $B^2 = C_0/\xi^2$.

Proof. Using the inequality in Lemma 13 (2), $\rho_{n,J} \leqslant \rho_{1,n} + \rho_{0,n}$ and $\rho_{n,J}^B \leqslant \rho_{1,n}^B + \rho_{0,n}^B$, which implies $\rho_{n,J} \leqslant Cn^{-c}$ and $P(\rho_{n,J}^B < Cn^{-c}) \geqslant 1 - Cn^{-c}$.

Using the inequality $|\sqrt{a}-1| \leqslant |a-1|$ for $a>0$,

$$
\begin{aligned}
&\max_{j\in J}\left|\frac{\max\{\sigma_j,\xi\}}{\max\{\hat\sigma_j,\xi\}}-1\right| \\
&\quad=\max_{j\in J}\frac{\hat\sigma_j}{\max\{\hat\sigma_j,\xi\}}\left|\max\left\{\frac{\sigma_j}{\hat\sigma_j},\frac{\xi}{\hat\sigma_j}\right\}-\max\left\{1,\frac{\xi}{\hat\sigma_j}\right\}\right| \\
&\quad\leqslant\max_{j\in J}\left|\frac{\sigma_j}{\hat\sigma_j}-1\right| \\
&\quad\leqslant\max_{j\in J}\left|\frac{\sigma_j^2}{\hat\sigma_j^2}-1\right| \\
&\quad=\max_{j\in J}\left|\frac{\frac{n_1}{n}\sigma_{1,j}^2+\frac{n_0}{n}\sigma_{0,j}^2}{\frac{n_1}{n}\hat\sigma_{1,j}^2+\frac{n_0}{n}\hat\sigma_{0,j}^2}-1\right| \\
&\quad\leqslant\max_{j\in J}\left|\frac{\frac{n_1}{n}\hat\sigma_{1,j}^2}{\frac{n_1}{n}\hat\sigma_{1,j}^2+\frac{n_0}{n}\hat\sigma_{0,j}^2}\left(\frac{\sigma_{1,j}^2}{\hat\sigma_{1,j}^2}-1\right)\right|+\max_{j\in J}\left|\frac{\frac{n_0}{n}\hat\sigma_{0,j}^2}{\frac{n_1}{n}\hat\sigma_{1,j}^2+\frac{n_0}{n}\hat\sigma_{0,j}^2}\left(\frac{\sigma_{0,j}^2}{\hat\sigma_{0,j}^2}-1\right)\right| \\
&\quad\leqslant\max_{j\in J}\left|\frac{\sigma_{1,j}^2}{\hat\sigma_{1,j}^2}-1\right|+\max_{j\in J}\left|\frac{\sigma_{0,j}^2}{\hat\sigma_{0,j}^2}-1\right|.
\end{aligned}
$$

Applying the triangle inequality and $\sigma_j \geqslant \frac{\sqrt{n_z}}{\sqrt{n}}\sigma_{z,j}$

$$
\begin{aligned}
&\max_{j\in J}\left|\frac{\sqrt{n}(\hat\mu_j-E[\hat\mu_j])}{\max\{\sigma_j,\xi\}}\right| \\
&\quad\leqslant\max_{j\in J}\left|\frac{\frac{\sqrt{n}}{\sqrt{n_1}}\sqrt{n_1}(\hat\mu_{1,j}-E[\hat\mu_{1,j}])}{\max\{\sigma_j,\xi\}}\right|+\max_{j\in J}\left|\frac{\frac{\sqrt{n}}{\sqrt{n_0}}\sqrt{n_0}(\hat\mu_{0,j}-E[\hat\mu_{0,j}])}{\max\{\sigma_j,\xi\}}\right| \\
&\quad\leqslant\max_{j\in J}\left|\frac{\frac{\sqrt{n}}{\sqrt{n_1}}\sqrt{n_1}(\hat\mu_{1,j}-E[\hat\mu_{1,j}])}{\frac{\sqrt{n_1}}{\sqrt{n}}\sigma_{1,j}}\right|+\max_{j\in J}\left|\frac{\frac{\sqrt{n}}{\sqrt{n_0}}\sqrt{n_0}(\hat\mu_{0,j}-E[\hat\mu_{0,j}])}{\frac{\sqrt{n_0}}{\sqrt{n}}\sigma_{0,j}}\right| \\
&\quad=\frac{n}{n_1}\max_{j\in J}\left|\frac{\sqrt{n_1}(\hat\mu_{1,j}-E[\hat\mu_{1,j}])}{\sigma_{1,j}}\right|+\frac{n}{n_0}\max_{j\in J}\left|\frac{\sqrt{n_0}(\hat\mu_{0,j}-E[\hat\mu_{0,j}])}{\sigma_{0,j}}\right|.
\end{aligned}
$$

Applying the triangle inequality

$$\begin{aligned}
|W(J)| &= \left| \max_{j \in J} \frac{\frac{\sqrt{n}}{\sqrt{n_1}} \sqrt{n_1} \hat{\mu}^B_{1,j} - \frac{\sqrt{n}}{\sqrt{n_0}} \sqrt{n_0} \hat{\mu}^B_{0,j}}{\max\{\hat{\sigma}_j, \xi\}} \right| \\
&\leqslant \left| \max_{j \in J} \frac{\frac{\sqrt{n}}{\sqrt{n_1}} \sqrt{n_1} \hat{\mu}^B_{1,j}}{\max\{\hat{\sigma}_j, \xi\}} \right| + \left| \max_{j \in J} \frac{\frac{\sqrt{n}}{\sqrt{n_0}} \sqrt{n_0} \hat{\mu}^B_{0,j}}{\max\{\hat{\sigma}_j, \xi\}} \right| \\
&= \frac{n}{n_1} \left| \max_{j \in J} \frac{\sqrt{n_1} \hat{\mu}^B_{1,j}}{\hat{\sigma}_{1,n_1,j}} \frac{\frac{\sqrt{n_1}}{\sqrt{n}} \hat{\sigma}_{1,n_1,j}}{\max\{\hat{\sigma}_j, \xi\}} \right| + \frac{n}{n_0} \left| \max_{j \in J} \frac{\sqrt{n_0} \hat{\mu}^B_{0,j}}{\hat{\sigma}_{0,n_0,j}} \frac{\frac{\sqrt{n_0}}{\sqrt{n}} \hat{\sigma}_{0,n_0,j}}{\max\{\hat{\sigma}_j, \xi\}} \right| \\
&\leqslant 2\frac{n}{n_1} \left| \max_{j \in J} \frac{\sqrt{n_1} \hat{\mu}^B_{1,j}}{\hat{\sigma}_{1,n_1,j}} \right| + 2\frac{n}{n_0} \left| \max_{j \in J} \frac{\sqrt{n_0} \hat{\mu}^B_{0,j}}{\hat{\sigma}_{0,n_0,j}} \right|.
\end{aligned}$$

□

Lemma 16. *If (θ, P) satisfies at least one of the inequalities in Theorem 7, then*

$$\max_{1 \leqslant j \leqslant p_n} \frac{\sqrt{n}\mu_j}{\max\{\sigma_j, \xi\}} \geqslant (1+\delta)(1+\epsilon)\sqrt{2\log(\max\{p_n, \tau_n\}/(\alpha - 2\beta))}.$$

Proof. When one of the first two inequalities in Theorem 7 holds, this lemma is immediate. I will show the case when the last inequality holds. By Assumptions 5 and 7,

$$\max_{(y,t) \in I_{n,k}} \Delta f_{(Y,T)|Z}(y,t) - \min_{(y,t) \in I_{n,k}} \Delta f_{(Y,T)|Z}(y,t) \leqslant 2D_0 \left(2\frac{D_1}{K_n}\right)^d = 2^{d+1} D_0 D_1^d K_n^{-d}.$$

Define $D = 2^{d+1} D_0 D_1^d$ and then

$$\max_{(y,t) \in I_{n,k}} \Delta f_{(Y,T)|Z}(y,t) - \min_{(y,t) \in I_{n,k}} \Delta f_{(Y,T)|Z}(y,t) \leqslant D K_n^{-d}.$$

for every $P \in \mathcal{P}$ and $k = 1, \ldots, K_n$. Define

$$\begin{aligned} h^* &= \arg\max_{h \in \mathbf{H}} \Delta E[h(Y,T) \mid Z] \\ h_{n,j*} &= \arg\max_{h_{n,j}: 1 \leqslant j \leqslant 4^{K_n}} \Delta E[h(Y,T) \mid Z]. \end{aligned}$$

If $|\Delta f_{(Y,T)|Z}(y,t)| > DK_n^{-d}$ for some (y,t) with $y \in I_{n,k}$, then $h^*(y,t)$ is constant on $I_{n,k}$, and therefore $h^* = h_{n,j*}$ is constant on $I_{n,k}$. Then, on every $I_{n,k}$, either $h^* = h_{n,j*}$ or $|\Delta f_{(Y,T)|Z}(y,t)| \leqslant DK_n^{-d}$. It implies

$$\begin{aligned} &\Delta E_P[h^*(Y,T) - h_{n,j}(Y,T) \mid Z] \\ &\quad = \sum_{t=0,1} \int (h^*(y,t) - h_{n,j}(y,t)) \Delta f_{(Y,T)|Z}(y,t) d\mu_Y(t) \\ &\quad = \sum_{k=1}^{K_n} \sum_{t=0,1} \int_{I_{n,k}} (h^*(y,t) - h_{n,j}(y,t)) \Delta f_{(Y,T)|Z}(y,t) d\mu_Y(t) \\ &\quad \leqslant \sum_{k=1}^{K_n} \sum_{t=0,1} \int_{I_{n,k}} DK_n^{-d} d\mu_Y(t) \\ &\quad = 2DK_n^{-d} \mu_Y(\mathbf{Y}) \end{aligned}$$

and

$$\begin{aligned} &TV(f_{(Y,T)|Z=z_1}, f_{(Y,T)|Z=z_0}) \\ &\quad = \Delta E_P[h^*(Y,T) \mid Z] \\ &\quad = \Delta E_P[h_{n,j*}(Y,T) \mid Z] + \Delta E_P[h^*(Y,T) - h_{n,j*}(Y,T) \mid Z] \\ &\quad = \Delta E_P[h_{n,j*}(Y,T) \mid Z] + 2DK_n^{-d} \mu_Y(\mathbf{Y}). \end{aligned}$$

Then

$$\begin{aligned}
\sqrt{n}\mu_{j*} &= \sqrt{n}(|\theta|\Delta E_P[h_{n,j*}(Y,T) \mid Z]/2 - |\Delta E_P[Y \mid Z]|) \\
&= \sqrt{n}(|\theta| TV(f_{(Y,T)|Z=z_1}, f_{(Y,T)|Z=z_0}) - |\Delta E_P[Y \mid Z]|) \\
&\quad +|\theta|\sqrt{n}(\Delta E_P[h_{n,j*}(Y,T) \mid Z]/2 - TV(f_{(Y,T)|Z=z_1}, f_{(Y,T)|Z=z_0})) \\
&\geqslant \sqrt{n}(|\theta| TV(f_{(Y,T)|Z=z_1}, f_{(Y,T)|Z=z_0}) - |\Delta E_P[Y \mid Z]|) \\
&\quad -|\theta|\sqrt{n}2DK_n^{-d}\mu_Y(\mathbf{Y}) \\
&\geqslant \sqrt{n}(|\theta| TV(f_{(Y,T)|Z=z_1}, f_{(Y,T)|Z=z_0}) - |\Delta E_P[Y \mid Z]|) \\
&\quad -\sqrt{n}\sup_{\theta\in\Theta}|\theta| 2^{d+2}D_0 D_1^d K_n^{-d}\mu_Y(\mathbf{Y}).
\end{aligned}$$

By the last inequality (1.7) in Theorem 7,

$$\frac{\sqrt{n}\mu_{j*}}{\max\{\sigma_{j*}, \xi\}} \geqslant (1+\delta)(1+\epsilon)\sqrt{2\log(\max\{p_n, \tau_n\}/(\alpha - 2\beta))}.$$

□

Define

$$\begin{aligned}
\zeta_{n1} &= \left(\frac{n}{n_1}n_1^{c_1/4} + \frac{n}{n_0}n_0^{c_1/4}\right)(n_1^{-1/2+c_1/4} + n_0^{-1/2+c_1/4})B^2\log^{3/2}(p_n) \\
\zeta_{n2} &= 8\frac{\left(\frac{n}{n_1} + \frac{n}{n_0}\right)}{\left(\frac{n}{n_1}n_1^{c_1/4} + \frac{n}{n_0}n_0^{c_1/4}\right)}.
\end{aligned}$$

Lemma 17.

$$\begin{aligned}
&P(|\bar{T}(J) - T_0(J)| > \zeta_{n1}) \leqslant Cn^{-c} \\
&P(P(|W(J) - \bar{W}^B(J)| > \zeta_{n1} \mid Data) > \zeta_{n2}) \leqslant 2C(n_1^{-c} + n_0^{-c}).
\end{aligned}$$

Proof. Since

$$\begin{aligned}
\bar{T}(J) &= \max_{j\in J} \frac{\sqrt{n}(\hat{\mu}_j - E[\hat{\mu}_j])}{\max\{\hat{\sigma}_j, \xi\}} \\
&= \max_{j\in J} \frac{\sqrt{n}(\hat{\mu}_j - E[\hat{\mu}_j])}{\max\{\sigma_j, \xi\}} \frac{\max\{\sigma_j, \xi\}}{\max\{\hat{\sigma}_j, \xi\}} \\
T_0(J) &= \max_{j\in J} \frac{\sqrt{n}(\hat{\mu}_j - E[\hat{\mu}_j])}{\max\{\sigma_j, \xi\}},
\end{aligned}$$

Lemma 13 implies

$$|\bar{T}(J) - T_0(J)| \leqslant \max_{j\in J} \left| \frac{\sqrt{n}(\hat{\mu}_j - E[\hat{\mu}_j])}{\max\{\sigma_j, \xi\}} \right| \max_{j\in J} \left| \frac{\max\{\sigma_j, \xi\}}{\max\{\hat{\sigma}_j, \xi\}} - 1 \right|$$

By Lemma 15, I have

$$P(|\bar{T}(J) - T_0(J)| > \zeta_{n1}) \leqslant 2C(n_1^{-c} + n_0^{-c}).$$

Since Lemma 13 implies

$$\begin{aligned}
|W(J) - \bar{W}^B(J)| &\leqslant \left| \max_{j\in J} \frac{\sqrt{n}\hat{\mu}_j^B}{\max\{\hat{\sigma}_j, \xi\}} \right| \max_{j\in J} \left| 1 - \frac{\max\{\hat{\sigma}_j, \xi\}}{\max\{\sigma_j, \xi\}} \right| \\
&= |W(J)| \max_{j\in J} \left| 1 - \frac{\max\{\hat{\sigma}_j, \xi\}}{\max\{\sigma_j, \xi\}} \right|,
\end{aligned}$$

Markov inequality implies

$$\begin{aligned}
P(|W(J) - \bar{W}^B(J)| > \zeta_{n1} \mid Data) &\leqslant \frac{1}{\zeta_{n1}} E[|W(J) - \bar{W}^B(J)| \mid Data] \\
&\leqslant \frac{1}{\zeta_{n1}} E[|W(J)| \mid Data] \max_{j \in J} \left| 1 - \frac{\max\{\hat{\sigma}_j, \xi\}}{\max\{\sigma_j, \xi\}} \right| \\
&\leqslant \frac{1}{\zeta_{n1}} 2 \left(\frac{n}{n_1} + \frac{n}{n_0} \right) \sqrt{2 \log(2|J|)} \\
&\qquad \times \max_{j \in J} \left| 1 - \frac{\max\{\hat{\sigma}_j, \xi\}}{\max\{\sigma_j, \xi\}} \right| \\
&\leqslant \zeta_{n2} \frac{\max_{j \in J} \left| 1 - \frac{\max\{\hat{\sigma}_j, \xi\}}{\max\{\sigma_j, \xi\}} \right|}{(n_1^{-1/2+c_1/4} + n_0^{-1/2+c_1/4}) B^2 \log(|J|)}.
\end{aligned}$$

Lemma 15 implies

$$P(P(|W(J) - \bar{W}^B(J)| > \zeta_{n1} \mid Data) > \zeta_{n2}) \leqslant 2C(n_1^{-c} + n_0^{-c}).$$

□

Define

$$\varphi_n = \zeta_{n2} + \nu_{1,n} + \nu_{0,n} + 4\zeta_{n1}(\sqrt{\log(p_n)} + 1).$$

Lemma 18.

$$\begin{aligned}
&c_0(\gamma + 4\zeta_{n1}(\sqrt{\log(|J|)} + 1), J) + \zeta_{n1} \leqslant c_0(\gamma, J) \\
&P(c(\gamma, J) \geqslant c_0(\gamma + \varphi_n, J)) \geqslant 1 - 3C(n_1^{-c} + n_0^{-c}).
\end{aligned}$$

Proof. By Lemma 14, I have

$$P\left(\max_{j\in J}(U_{1,j}-U_{0,j})\leqslant t+\zeta_{n1}\right)$$
$$\leqslant P\left(\max_{j\in J}(U_{1,j}-U_{0,j})\leqslant t\right)+4\zeta_{n1}(\sqrt{\log(|J|)}+1)$$

and therefore

$$c_0(\gamma+4\zeta_{n1}(\sqrt{\log(|J|)}+1),J)+\zeta_{n1}\leqslant c_0(\gamma,J).$$

If $\rho^B_{n,J}<\nu_{1,n}+\nu_{0,n}$ and $P(|W-\bar{W}|>\zeta_{n1}\mid Data)<\zeta_{n2}$, then

$$\begin{aligned}
P(W(J)\leqslant t\mid Data) &\leqslant P(\bar{W}\leqslant t+\zeta_{n1}\mid Data)+P(|W-\bar{W}|>\zeta_{n1}\mid Data)\\
&\leqslant P\left(\max_{j\in J}(U_{1,j}-U_{0,j})\leqslant t+\zeta_{n1}\mid Data\right)\\
&\quad +\rho^B_{n,J}+P(|W-\bar{W}|>\zeta_{n1}\mid Data)\\
&\leqslant P\left(\max_{j\in J}(U_{1,j}-U_{0,j})\leqslant t\right)+4\zeta_{n1}(\sqrt{\log(|J|)}+1)\\
&\quad +\rho^B_{n,J}+P(|W-\bar{W}|>\zeta_{n1}\mid Data)\\
&\leqslant P\left(\max_{j\in J}(U_{1,j}-U_{0,j})\leqslant t\right)\\
&\quad +4\zeta_{n1}(\sqrt{\log(|J|)}+1)+\nu_{1,n}+\nu_{0,n}+\zeta_{n2}\\
&\leqslant P\left(\max_{j\in J}(U_{1,j}-U_{0,j})\leqslant t\right)+\varphi_n
\end{aligned}$$

and, at $t=c_0(\gamma+\varphi_n,J)$,

$$P(W(J)\leqslant c_0(\gamma+\varphi_n,J)\mid Data)\leqslant 1-\gamma.$$

Therefore

$$c(\gamma, J) \geqslant c_0(\gamma + \varphi_n, J)$$

if $\rho^B_{n,J} < \nu_{1,n} + \nu_{0,n}$ and $P(|W - \bar{W}| > \zeta_{n1} \mid Data) < \zeta_{n2}$. By Lemmas 15 and 17, this lemma is established. □

Define $J_2 = \{j \in \{1, \ldots, p_n\} : \sqrt{n}\mu_j / \max\{\sigma_j, \xi\} > -c_0(\beta + \varphi_n, \{1, \ldots, p_n\})\}$ and I obtain the following lemma.

Lemma 19.

$$\begin{aligned} P(J_2 \nsubseteq \hat{J}) \quad &< \quad \beta + \varphi_n + 4r_n \left(\sqrt{2\log(p_n)} + \sqrt{-2\log(\beta + \varphi_n)}\right) \sqrt{\log(p_n) + 1} \\ &\quad + \rho_{n,J} + 4C(n_1^{-c} + n_0^{-c}). \end{aligned}$$

Proof. Define $r_n = 2(n_1^{-1/2+c_1/4} + n_0^{-1/2+c_1/4})B^2 \log(p_n)$ and consider $J = \{1, \ldots, p_n\}$. Note that $J_2 \nsubseteq \hat{J}$ implies

$$\sqrt{n}\frac{\mu_j}{\max\{\sigma_j, \xi\}} > -c_0(\beta + \varphi_n, J) \text{ and } \frac{\sqrt{n}\mu_j}{\max\{\hat{\sigma}_j(\theta), \xi\}} > -2c(\beta, J) \text{ for some } j \in J.$$

If $c(\beta, J) \geqslant c_0(\beta + \varphi_n, J)$ and $\max_{j \in J} |(\max\{\hat{\sigma}_j, \xi\} / \max\{\sigma_j, \xi\}) - 1| \leqslant r_n/2$, then $J_2 \nsubseteq \hat{J}$ implies that

$$\sqrt{n}(\hat{\mu}_j - \mu_j) - (1 - r_n)\max\{\sigma_j, \xi\}c_0(\beta + \varphi_n, J) > 0 \text{ for some } j \in J.$$

Therefore

$$
\begin{aligned}
&P(J_2 \nsubseteq \hat{J}) \\
&\quad \leqslant P\left(\max_{1\leqslant j\leqslant p_n} \sqrt{n}(\hat{\mu}_j - \mu_j)/\max\{\sigma_j, \xi\} > (1-r_n)c_0(\beta+\varphi_n, J)\right) \\
&\qquad + P(c(\beta, J) < c_0(\beta+\varphi_n, J)) \\
&\qquad + P\left(\max_{1\leqslant j\leqslant p_n} |(\max\{\hat{\sigma}_j, \xi\}/\max\{\sigma_j, \xi\}) - 1| > r_n/2\right) \\
&\quad = P\left(\max_{j\in J}(U_{1,j} - U_{0,j}) > (1-r_n)c_0(\beta+\varphi_n, J))\right) \\
&\qquad + \rho_{n,J} + P(c(\beta, J) < c_0(\beta+\varphi_n, J)) \\
&\qquad + P\left(\max_{1\leqslant j\leqslant p_n} |(\max\{\hat{\sigma}_j, \xi\}/\max\{\sigma_j, \xi\}) - 1| > r_n/2\right) \\
&\quad \leqslant P\left(\max_{j\in J}(U_{1,j} - U_{0,j}) > c_0(\beta+\varphi_n, J))\right) \\
&\qquad + P\left(\left|\max_{j\in J}(U_{1,j} - U_{0,j}) - c_0(\beta+\varphi_n, J))\right| < r_n c_0(\beta+\varphi_n, J)\right) \\
&\qquad + \rho_{n,J} + P(c(\beta, J) < c_0(\beta+\varphi_n, J)) \\
&\qquad + P\left(\max_{1\leqslant j\leqslant p_n} |(\max\{\hat{\sigma}_j, \xi\}/\max\{\sigma_j, \xi\}) - 1| > r_n/2\right) \\
&\quad = \beta + \varphi_n + P\left(\left|\max_{j\in J}(U_{1,j} - U_{0,j}) - c_0(\beta+\varphi_n, J))\right| < r_n c_0(\beta+\varphi_n, J)\right) \\
&\qquad + \rho_{n,J} + P\left(c(\beta, J) < c_0(\beta+\varphi_n, J)\right) \\
&\qquad + P\left(\max_{1\leqslant j\leqslant p_n} |(\max\{\hat{\sigma}_j, \xi\}/\max\{\sigma_j, \xi\}) - 1| > r_n/2\right).
\end{aligned}
$$

By Lemma 14,

$$
\begin{aligned}
P\left(\left|\max_{j\in J}(U_{1,j}-U_{0,j})-c_0(\beta+\varphi_n,J)\right|<r_nc_0(\beta+\varphi_n,J)\right)&\\
\leqslant 4r_nc_0(\beta+\varphi_n,J)\sqrt{\log(p_n)+1}&\\
\leqslant 4r_n\left(\sqrt{2\log(p_n)}+\sqrt{-2\log(\beta+\varphi_n)}\right)\sqrt{\log(p_n)+1}&
\end{aligned}
$$

and then

$$
\begin{aligned}
P(J_2\nsubseteq\hat{J}) \;\leqslant\; & \beta+\varphi_n+4r_n\left(\sqrt{2\log(p_n)}+\sqrt{-2\log(\beta+\varphi_n)}\right)\sqrt{\log(p_n)+1}+\rho_{n,J}\\
&+P(c(\beta,J)<c_0(\beta+\varphi_n,J))\\
&+P\left(\max_{1\leqslant j\leqslant p_n}|(\max\{\hat{\sigma}_j,\xi\}/\max\{\sigma_j,\xi\})-1|>r_n/2\right).
\end{aligned}
$$

By Lemmas 15 and 18, this lemma is established. □

Proof of Theorem 6

First, I show

$$
\begin{aligned}
P(T(J)>c(\gamma,J)) \;\leqslant\; & \gamma+\zeta_{n2}+\nu_n+8\zeta_{n1}(\sqrt{\log(|J|)}+1)+\rho_{n,J}\\
&+P(c_0(\gamma+\varphi_n,J)>c(\gamma,J))\\
&+P(|\bar{T}(J)-T_0(J)|>\zeta_{n1}) \qquad (1.16)
\end{aligned}
$$

for every $J\subset\{1,\ldots,p_n\}$. Since $T(J)\leqslant\bar{T}(J)$,

$$
\begin{aligned}
P(T(J)>c(\gamma,J)) &\leqslant P(\bar{T}(J)>c(\gamma,J))\\
&\leqslant P(T_0(J)>c(\gamma,J)-\zeta_{n1})+P(|\bar{T}(J)-T_0(J)|>\zeta_{n1})\\
&\leqslant P(T_0(J)>c_0(\gamma+\varphi_n,J)-\zeta_{n1})\\
&\quad +P(c_0(\gamma+\varphi_n,J)>c(\gamma,J))+P(|\bar{T}(J)-T_0(J)|>\zeta_{n1}).
\end{aligned}
$$

Using Lemma 18,

$$
\begin{aligned}
&P(T_0(J) > c_0(\gamma + \varphi_n, J) - \zeta_{n1}) \\
&\quad \leqslant P(T_0(J) > c_0(\gamma + \varphi_n + 4\zeta_{n1}(\sqrt{\log(|J|)} + 1), J)) \\
&\quad = P(T_0(J) > c_0(\gamma + \zeta_{n2} + \nu_n + 8\zeta_{n1}(\sqrt{\log(|J|)} + 1), J)) \\
&\quad \leqslant P\left(\max_{j \in J}(U_{1,j} - U_{0,j}\right) > c_0(\gamma + \zeta_{n2} + \nu_n + 8\zeta_{n1}(\sqrt{\log(|J|)} + 1), J)) + \rho_{n,J} \\
&\quad = \gamma + \zeta_{n2} + \nu_n + 8\zeta_{n1}(\sqrt{\log(|J|)} + 1) + \rho_{n,J}.
\end{aligned}
$$

Next, I show

$$
P\left(\max_{j \notin J_2} \hat{\mu}_j \leqslant 0\right) > 1 - \beta - \varphi_n - \rho_{n,J}. \tag{1.17}
$$

Since $\max_{j \notin J_2} \hat{\mu}_j > 0$ implies

$$
\max_{1 \leqslant j \leqslant p_n} \sqrt{n}(\hat{\mu}_j - \mu_j)/\max\{\sigma_j, \xi\} > c_0(\beta + \varphi_n, \{1, \ldots, p_n\}),
$$

it follows that

$$
\begin{aligned}
P\left(\max_{j \notin J_2} \hat{\mu}_j > 0\right) &\leqslant P\left(\max_{1 \leqslant j \leqslant p_n} \frac{\sqrt{n}(\hat{\mu}_j - \mu_j)}{\max\{\sigma_j, \xi\}} > c_0(\beta + \varphi_n, \{1, \ldots, p_n\})\right) \\
&\leqslant P\left(\max_{1 \leqslant j \leqslant p_n} (U_{1,j} - U_{0,j}) > c_0(\beta + \varphi_n, \{1, \ldots, p_n\})\right) + \rho_{n,J} \\
&= \beta + \varphi_n + \rho_{n,J}.
\end{aligned}
$$

Last, I show that the statement of this theorem $P(T \leqslant c^{2S}(\alpha)) \geqslant 1 - (\alpha - 2\beta) -$

Cn^{-c}. If the following three statements are true

$$T(J_2) \leqslant c(\alpha - 2\beta, J_2)$$

$$\max_{j \notin J_2} \sqrt{n}\hat{\mu}_j / \max\{\hat{\sigma}_j, \xi\} \leqslant 0$$

$$J_2 \subset \hat{J},$$

then I have

$$\begin{aligned} T &= \max_{1 \leqslant j \leqslant p_n} \sqrt{n}\hat{\mu}_j / \max\{\hat{\sigma}_j, \xi\} \\ &= \max_{j \in J_2} \sqrt{n}\hat{\mu}_j / \max\{\hat{\sigma}_j, \xi\} \\ &= T(J_2) \\ &\leqslant c(\alpha - 2\beta, J_2) \\ &\leqslant c(\alpha - 2\beta, \hat{J}) \\ &= c^{2S}(\alpha). \end{aligned}$$

By Lemmas 19 and Eq. (1.16) and (1.17), I have

$$\begin{aligned} P(T > c^{2S}(\alpha)) \leqslant\ & \alpha - 2\beta + \varphi_n + \rho_{n,J} \\ & + \zeta_{n2} + \nu_n + 8\zeta_{n1}(\sqrt{\log(|J|)} + 1) + \rho_{n,J} \\ & + P(c_0(\gamma + \varphi_n, J) > c(\gamma, J)) \\ & + P(|\bar{T}(J) - T_0(J)| > \zeta_{n1}) \\ & + \varphi_n + 4r_n \left(\sqrt{2\log(p_n)} + \sqrt{-2\log(\beta + \varphi_n)} \right) \sqrt{\log(p_n) + 1} \\ & + \rho_{n,J} + 4C(n_1^{-c} + n_0^{-c}). \end{aligned}$$

Except for $\alpha - 2\beta$, all the terms on the right-hand side converges to 0 uniformly over P.

Proof of Theorem 7

Denote by $j^* = \arg\max_{1 \leqslant j \leqslant p_n} \mu_j / \max\{\sigma_j, \xi\}$. If the following four statements are true

$$\left| \frac{\max\{\sigma_{j^*}, \xi\}}{\max\{\hat{\sigma}_{j^*}, \xi\}} - 1 \right| < \delta \tag{1.18}$$

$$\left| \frac{\sqrt{n}(\hat{\mu}_{j^*} - \mu_{j^*})}{\sigma_{j^*}} \right| < (1-\delta)\epsilon\sqrt{2\log(\max\{p_n, \tau_n\}/(\alpha - 2\beta))} \tag{1.19}$$

$$\frac{\sqrt{n}\mu_{j^*}}{\max\{\sigma_{j^*}, \xi\}} \geqslant (1+\delta)(1+\epsilon)\sqrt{2\log(\max\{p_n, \tau_n\}/(\alpha - 2\beta))} \tag{1.20}$$

$$c(\alpha - 2\beta, \{1, \ldots, p_n\}) \leqslant \sqrt{2\log(p_n)} + \sqrt{-2\log(\alpha - 2\beta)}, \tag{1.21}$$

then $T > c^{2S}(\alpha)$, because

$$
\begin{aligned}
T &\geqslant \frac{\sqrt{n}\hat{\mu}_{j*}}{\max\{\hat{\sigma}_{j*}, \xi\}} \\
&= \frac{\sqrt{n}\mu_{j*}}{\max\{\hat{\sigma}_{j*}, \xi\}} + \frac{\sqrt{n}(\hat{\mu}_{j*} - \mu_{j*})}{\max\{\hat{\sigma}_{j*}, \xi\}} \\
&\geqslant \frac{1}{1+\delta}\frac{\sqrt{n}\mu_{j*}}{\max\{\sigma_{j*}, \xi\}} - \frac{1}{1-\delta}\left|\frac{\sqrt{n}(\hat{\mu}_{j*} - \mu_{j*})}{\max\{\sigma_{j*}, \xi\}}\right| \\
&> \frac{1}{1+\delta}(1+\delta)(1+\epsilon)\sqrt{2\log(\max\{p_n, \tau_n\}/(\alpha - 2\beta))} - \frac{1}{1-\delta}\left|\frac{\sqrt{n}(\hat{\mu}_{j*} - \mu_{j*})}{\max\{\sigma_{j*}, \xi\}}\right| \\
&= (1+\epsilon)\sqrt{2\log(\max\{p_n, \tau_n\}/(\alpha - 2\beta))} - \frac{1}{1-\delta}\frac{\sigma_{j*}}{\max\{\sigma_{j*}, \xi\}}\left|\frac{\sqrt{n}(\hat{\mu}_{j*} - \mu_{j*})}{\sigma_{j*}}\right| \\
&\geqslant (1+\epsilon)\sqrt{2\log(\max\{p_n, \tau_n\}/(\alpha - 2\beta))} - \epsilon\sqrt{2\log(\max\{p_n, \tau_n\}/(\alpha - 2\beta))} \\
&= \sqrt{2\log(\max\{p_n, \tau_n\}/(\alpha - 2\beta))} \\
&\geqslant \sqrt{2\log(p_n)} + \sqrt{-2\log(\alpha - 2\beta)} \\
&\geqslant c(\alpha - 2\beta, \{1, \ldots, p_n\}) \\
&\geqslant c(\alpha - 2\beta, \hat{J}) \\
&= c^{2S}(\alpha).
\end{aligned}
$$

(1.18) holds at least with probability $1 - C(n_1^{-c} + n_0^{-c})$ from Lemma 15. (1.19) hold at least with probability approaching to one because, using the Markov inequality,

$$\begin{aligned}
&P\left(\frac{\sqrt{n}(\hat{\mu}_{j*} - \mu_{j*})}{\sigma_{j*}} > -(1-\delta)\epsilon\sqrt{\log(\max\{p_n, \tau_n\}/(\alpha - 2\beta))}\right) \\
&\quad = 1 - P\left(-\frac{\sqrt{n}(\hat{\mu}_{j*} - \mu_{j*})}{\sigma_{j*}} \geqslant (1-\delta)\epsilon\sqrt{\log(\max\{p_n, \tau_n\}/(\alpha - 2\beta))}\right) \\
&\quad \geqslant 1 - \frac{1}{(1-\delta)^2\epsilon^2 \log(\max\{p_n, \tau_n\}/(\alpha - 2\beta))} E\left[\left(\frac{\sqrt{n}(\hat{\mu}_{j*} - \mu_{j*})}{\sigma_{j*}}\right)^2\right] \\
&\quad = 1 - \frac{1}{(1-\delta)^2\epsilon^2 \log(\max\{p_n, \tau_n\}/(\alpha - 2\beta))}.
\end{aligned}$$

(1.20) hold by Lemma 16 and (1.21) holds by Lemma 14. Therefore

$$P(T > c^{2S}(\alpha)) \geqslant 1 - \frac{1}{(1-\delta)^2\epsilon^2 \log(\max\{p_n, \tau_n\}/(\alpha - 2\beta))} - C(n_1^{-c} + n_0^{-c}).$$

Since the right-hand side of the above equation does not depend on P, the uniform convergence in Theorem 7 follows.

1.12 Tables

Table 1.1: Parameter values for Monte Carlo simulations. I numerically calculate LATE $\theta(P^*)$, the identified set $\Theta_0(P)$ and the Wald estimand $\Delta E_P[Y \mid Z]/\Delta E_P[T \mid Z]$ for each parameter value.

γ_1	γ_2	γ_3	LATE	Identified Set		Wald Estimand
0.1	1	0.6	0.28	[0.03,	0.58]	1.43
0.5	1	0.6	0.28	[0.14,	0.58]	1.40
0.1	1	0.8	0.28	[0.03,	0.43]	0.46
0.5	1	0.8	0.28	[0.14,	0.43]	0.47
0.1	3	0.6	0.49	[0.05,	0.52]	2.52
0.5	3	0.6	0.49	[0.24,	0.52]	2.42
0.1	3	0.8	0.49	[0.05,	0.52]	0.84
0.5	3	0.8	0.49	[0.24,	0.52]	0.81

Table 1.2: Summary Statistics for (Y, T, Z)

All sample	Mean	Std. Dev.
Y (log hourly wage)	2.08	0.47
T (college attendance)	0.30	0.46
Z (lived near college)	0.51	0.50

Table 1.3: Demographic groups

	white	black	male	female	non-south	south
observations	2,249	329	1,267	1,642	1,884	1,025
mean of Y	2.09	1.97	2.28	1.91	2.10	2.03
mean of T	0.31	0.31	0.31	0.31	0.32	0.28
mean of Z	0.50	0.64	0.48	0.53	0.53	0.47

Table 1.4: 95% confidence intervals for the Wald estimand and LATE

All sample	Estimates	CI
Wald estimand	0.95	[0.56, 1.61]
LATE	N.A.[4]	[0.05, 1.04]

Table 1.5: 95% confidence intervals for various subpopulations

White	Estimates	CI
Wald	1.20	[0.68, 2.25]
LATE	N.A.	[0.06, 1.16]

Black	Estimates	CI
Wald	0.95	[0.54, 1.66]
LATE	N.A.	[0.05, 1.06]

Male	Estimates	CI
Wald	0.93	[0.40, 2.07]
LATE	N.A.	[0.04, 1.44]

Female	Estimates	CI
Wald	1.31	[0.76, 2.72]
LATE	N.A.	[0.07, 1.28]

Non-south	Estimates	CI
Wald	1.15	[0.60, 2.30]
LATE	N.A.	[0.06, 1.47]

South	Estimates	CI
Wald	0.58	[0.00, 1.63]
LATE	N.A.	[-0.02, 1.04]

1.13 Figures

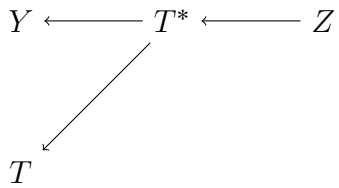

FIGURE 1.1: Three equations in the model

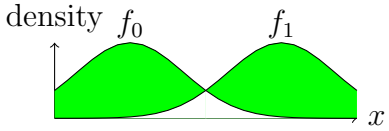

FIGURE 1.2: Definition of the total variation distance

[4] Estimates for θ are not available. In this paper, I focus on an inference about the true parameter value θ, instead of the identified set $\Theta_0(P)$. I do not have a consistent set estimator for $\Theta_0(P)$.

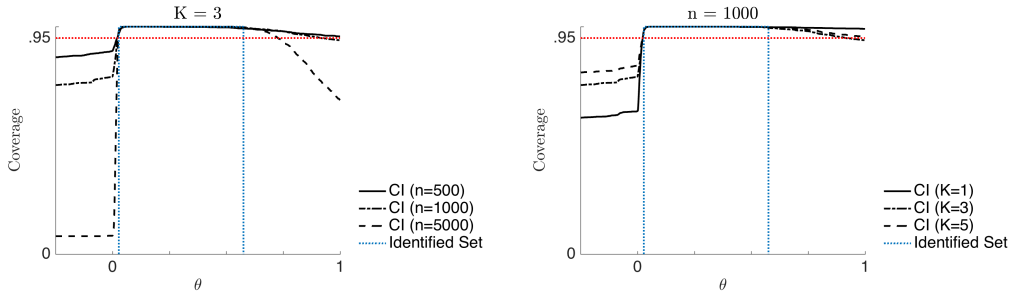

FIGURE 1.3: Coverage of the confidence interval for $(\gamma_1, \gamma_2, \gamma_3) = (0.1, 1, 0.6)$

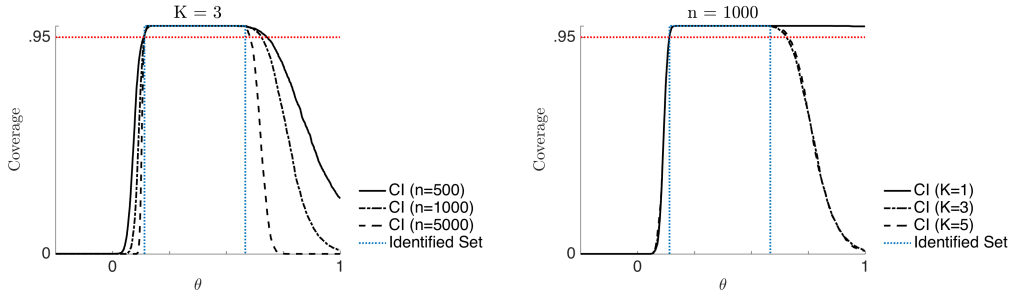

FIGURE 1.4: Coverage of the confidence interval for $(\gamma_1, \gamma_2, \gamma_3) = (0.5, 1, 0.6)$

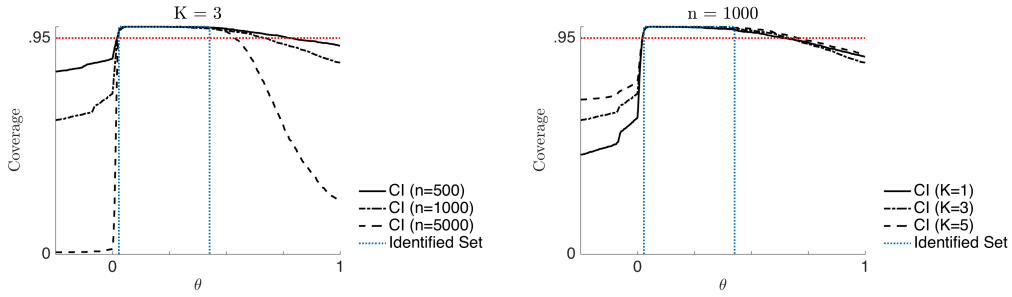

FIGURE 1.5: Coverage of the confidence interval for $(\gamma_1, \gamma_2, \gamma_3) = (0.1, 1, 0.8)$

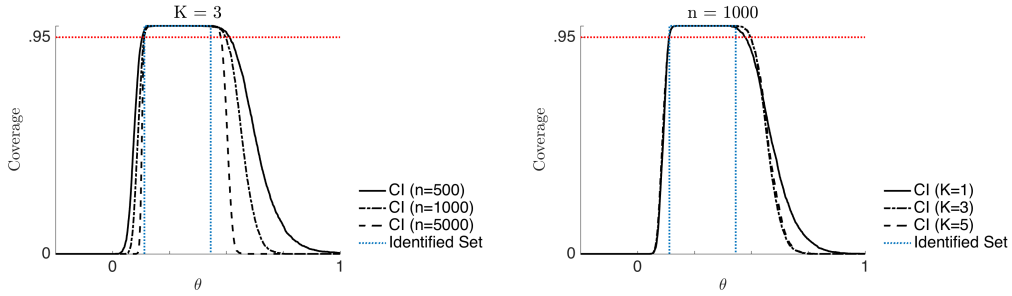

FIGURE 1.6: Coverage of the confidence interval for $(\gamma_1, \gamma_2, \gamma_3) = (0.5, 1, 0.8)$

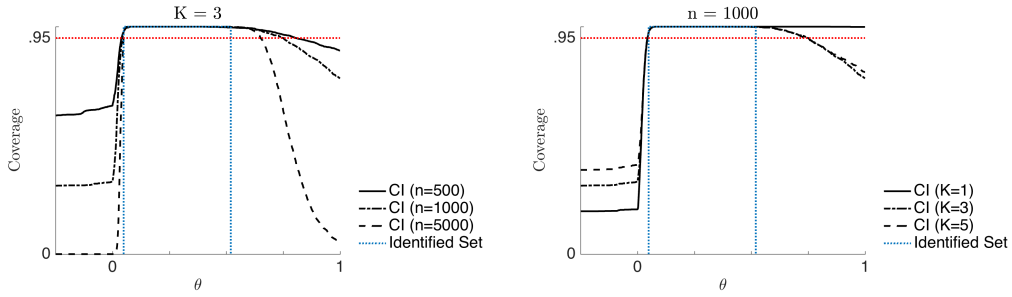

FIGURE 1.7: Coverage of the confidence interval for $(\gamma_1, \gamma_2, \gamma_3) = (0.1, 3, 0.6)$

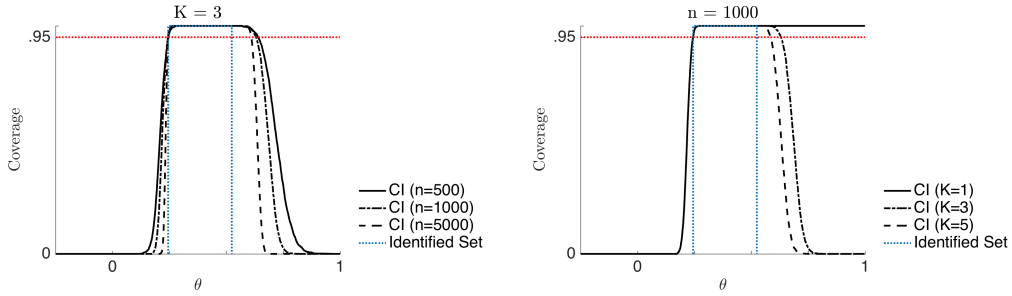

FIGURE 1.8: Coverage of the confidence interval for $(\gamma_1, \gamma_2, \gamma_3) = (0.5, 3, 0.6)$

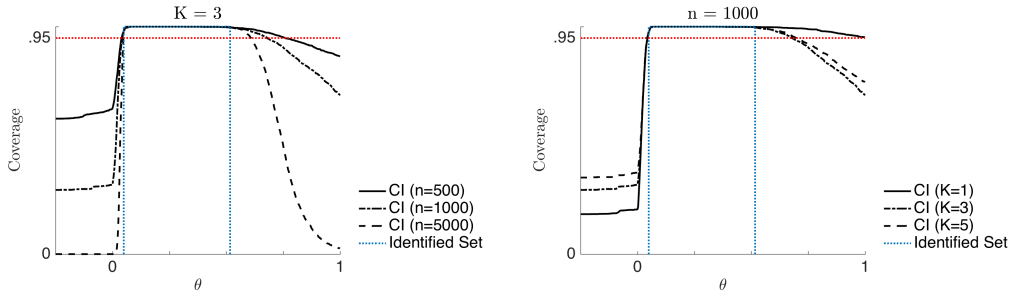

FIGURE 1.9: Coverage of the confidence interval for $(\gamma_1, \gamma_2, \gamma_3) = (0.1, 3, 0.8)$

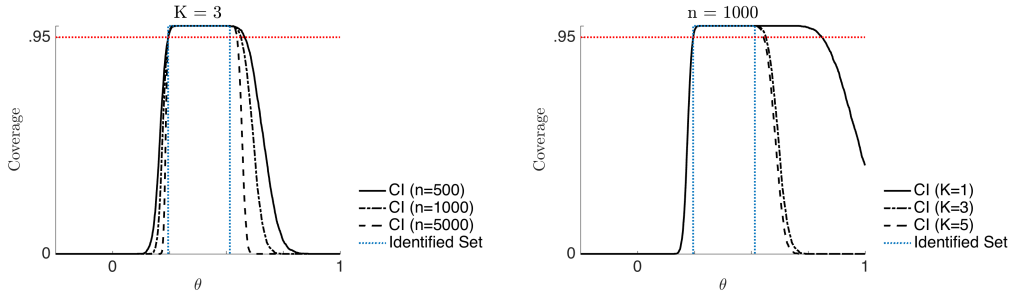

FIGURE 1.10: Coverage of the confidence interval for $(\gamma_1, \gamma_2, \gamma_3) = (0.5, 3, 0.8)$

2

Inference in dynamic discrete choice problems under local misspecification

2.1 Introduction

This paper studies the problem of inference of dynamic discrete choice models under misspecification. This project is motivated by two observations regarding the existing literature. First, econometric frameworks used to model these problems are typically heavily parametrized and are therefore prone to misspecification. Second, there are several methods that can be used to estimate these models, including Rust (1987)'s nested fixed point estimator, Hotz and Miller (1993)'s conditional choice probability estimator, Aguirregabiria and Mira (2002)'s nested algorithm estimator, and Pesendorfer and Schmidt-Dengler (2008)'s least squares estimator. While researchers have investigated the behavior of these estimators under correct specification, their properties under misspecification are less understood. To the best of our knowledge,

our paper is the first attempt to investigate the effect of misspecification in dynamic discrete choice models.

We introduce local misspecification in a parametric dynamic discrete choice model considered by Aguirregabiria and Mira (2002). In other words, in this framework, we assume that the researcher proposes an econometric model that is incorrect (i.e. it is misspecified), but the amount of misspecification vanishes as the sample size increases (i.e. it is local to zero). Throughout this paper, we allow the rate at which the misspecification problem disappears to be completely arbitrary. In particular, it can be faster, equal, or slower than the regular rate for sampling error, i.e., $\sqrt{n}$.

We consider the class of two stage estimators based on the K-step sequential policy iteration algorithm developed by Aguirregabiria and Mira (2002), where K denotes the number of iterations employed in the estimation. By appropriate choice of the criterion function, this class captures the K-step maximum likelihood estimators (K-ML) and the K-step minimum distance estimators (K-MD). This is a very general class which includes all the previously mentioned estimators as special cases.

We use this general framework to derive and compare the asymptotic distributions of K-ML and K-MD estimators when the model is arbitrarily locally misspecified. We obtain three main results. In the absence of misspecification, Aguirregabiria and Mira (2002) show that all K-ML estimators are asymptotically equivalent regardless of the choice of K. Our first result shows that this finding extends to a locally misspecified model, regardless of the degree of local misspecification. As a second result, we show that an analogous result holds for all K-MD estimators, i.e., all K-MD estimator are asymptotically equivalent regardless of the choice of K. Our third and final result is to compare K-MD and K-ML estimators in terms of asymptotic

mean squared error (AMSE). Under local misspecification, the optimally weighted K-MD estimator depends on the unknown asymptotic bias and is no longer feasible. In turn, feasible K-MD estimators could have an AMSE that is higher or lower than that of the K-ML estimators. To demonstrate the relevance of our asymptotic analysis, we illustrate our findings using in a simulation exercise based on a misspecified version of Rust (1987) bus engine problem.

Local misspecification is an asymptotic device that can help provide some concrete conclusions in the context of misspecification. First, if its definition is taken literally, local misspecification could constitute a reasonable approximation to the asymptotic behavior when there are small mistakes in the model specification. Second, as we have already explained, there are multiple available estimation methods and their performance under misspecification is not well understood. The relative performance of the various estimators under local misspecification should be a very relevant comparison criterion.

In practice, researchers specify econometric models that may contain non vanishing specification errors. In this sense, one might prefer to analyze the effect of global misspecification, rather than local misspecification. Relative to the global misspecification analysis, our local misspecification approach has two related advantages. First, regular misspecification analysis is typically intractable and general results are almost impossible to obtain. In contrast, we will be able to derive general results under local misspecification. Second, as we have already mentioned, there are multiple available estimation methods for the structural parameter of interest. Under correct specification, all of these estimators are (typically) consistent and they only differ in asymptotic distribution. Under global misspecification, the different

estimation methods typically converge to different pseudo-true parameters, making results hard to interpret and compare. In contrast, under local misspecification, all of these estimators will be shown to be consistent to the same true parameter value. Furthermore, local misspecification will produce asymptotic distributions that are relatively easy to compare analytically.

The local misspecification approach has been used in the econometrics literature is a wide array of settings. White (1982); Newey (1985a,b) investigate the power properties of the specification tests based on the point identified models under local misspecification. See also White (1996) for an excellent review of inference results under misspecification. Schorfheide (2005) considers a local misspecified vector autoregression process and proposes an information criterion for the lag length in the autoregression model. Finally, Bugni et al. (2012) uses compare inference methods in partially identified moment (in)equality models that are locally misspecified.

The remainder of the paper is structured as follows. Section 2.2 describes the setup of the econometric framework. Section 2.2.1 introduces the econometric model and provides an example that illustrates the assumptions. Section 2.2.2 introduces the possibility of local misspecification in the econometric model. Section 2.3 studies the problem of econometric inference in the locally misspecified model. The estimation procedure is occurs in two stages. In the first stage, the researcher estimates the parameters of the transition probability and, in the second stage, the researcher estimates the remainder of the parameters taking the first stage as given. The main result of the paper describes the asymptotic distribution of the second stage estimators under certain high-level conditions. Section 2.4 applies the high-level result to several estimators. In particular, Section 2.4.2 applies the result to a class of maxi-

mum likelihood estimators and Section 2.4.3 applies the result to a class of minimum distance estimators. Section 2.5 presents results of Monte Carlo simulations and Section 2.6 concludes the paper. The appendix of the paper collects all the proofs and several intermediate results.

Throughout the paper, we use the following notation. For any $s \in \mathbb{N}$, $\mathbf{0}_s$ and $\mathbf{1}_s$ denote a column vector of size $s \times 1$ composed of zeros and ones, respectively, and $\mathbf{I}_s$ denotes the identity matrix of size $s \times s$. We use $||\cdot||$ to denote the euclidean norm. Also, for any $S \subset \Theta$, $Int(S)$ and ∂S denote the interior and boundary of S relative to the topology defined by Θ. For sets of finite indices $S_1 = \{1, \ldots, |S_1|\}$ and $S_2 = \{1, \ldots, |S_2|\}$, the expression $\{M(s_1, s_2)\}_{(s_1,s_2)\in S_1 \times S_2}$ denotes the column vector equal to the vectorization of $\{\{M(s_1, s_2)\}_{s_1=1}^{|S_1|}\}_{s_2=1}^{|S_2|}$.

2.2 Setup

The researcher is interested in modeling the behavior of an agent solving a dynamic discrete choice problem. In order to study this problem, he consider an econometric model that is described in Section 2.2.1. In this paper, we are interested in investigating the consequences of imposing an incorrect econometric model. The nature of these modeling mistakes (misspecification) are characterized in Section 2.2.2.

2.2.1 The econometric model

The researcher assumes that the agent behaves according to the discrete Markov decision framework in Aguirregabiria and Mira (2002), which we now review in detail.

In each period $t = 1, \ldots, T \equiv \infty$, the agent is assumed to observe a vector of state variables s_t and to choose an action $a_t \in A \equiv \{1, \ldots, |A|\}$ with the objective

of maximizing the expected sum of current and future discounted utilities. The vector of state variables $s_t = (x_t, \epsilon_t)$ is composed by two subvectors. The subvector $x_t \in X \equiv \{1, \ldots, |X|\}$ represents a scalar state variables that is observed by both the agent and the researcher, whereas the subvector $\epsilon_t \in \mathbb{R}^{|A|}$ represents an action-specific state vector that is only observed by the agent.

The uncertainty about the agent's future state variables $(x_{t+1}, \epsilon_{t+1})$ are modeled by a Markov transition probability density $d\Pr(x_{t+1}, \epsilon_{t+1}|x_t, \epsilon_t, a_t)$ that can factors in the following manner:

$$d\Pr(x_{t+1}, \epsilon_{t+1}|x_t, \epsilon_t, a_t) = g_{\theta_g}(\epsilon_{t+1}|x_{t+1}) f_{\theta_f}(x_{t+1}|x_t, a_t),$$

where $g_{\theta_g}(\cdot)$ is the (conditional) distribution of the unobserved state variable and $f_{\theta_f}(\cdot)$ is the transition probability of the observed state variable, with parameters θ_g and θ_f, respectively.

The utility is assumed to be time separable and the agent discounts future utility by a known discount factor $\beta \in (0, 1)$.[1] The current utility function of choosing action a_t under state variables (x_t, ϵ_t) is given by:

$$u_{\theta_u}(x_t, a_t) + \epsilon_t(a_t),$$

where $u_{\theta_u}(\cdot)$ is non-stochastic component of the current utility with parameter θ_u, and $\epsilon_t(a_t)$ denotes the a_t-th coordinate of ϵ_t.

The researcher's goal is to estimate the unknown parameters in the model, $\theta = (\theta_g, \theta_u, \theta_f) \in \Theta$, where Θ is the compact parameter space. For the sake of notation, we use $\theta = (\alpha, \theta_f) \in \Theta = \Theta_\alpha \times \Theta_f$ with $\alpha \equiv (\theta_u, \theta_g) \in \Theta_\alpha$ and $\theta_f \in \Theta_f$.

[1] This follows Aguirregabiria and Mira (2002, Footnote 12) and the identification analysis in Magnac and Thesmar (2002).

Following Aguirregabiria and Mira (2002), we impose the following additional regularity conditions.

Assumption 9. *For every $\theta \in \Theta$, assume that:*

(a) For every x', $g_{\theta_g}(\epsilon'|x')$ has finite first moments and is continuous and twice differentiable in ϵ',

(b) $\epsilon = \{\epsilon(a)\}_{a \in A}$ has full support.

(c) $g_{\theta_g}(\epsilon'|x')$, $f_{\theta_f}(x'|x,a)$, and $u_{\theta_u}(x,a)$ are twice continuously differentiable with respect to θ.

By Blackwell (1965)'s theorem (and its further generalization by Rust (1988) to allow for unbounded period utilities), the model proposed by the researcher implies that the agent has a stationary and Markovian optimal decision rule. This implies that the researcher can drop the time subscript from the model and use prime to denote future periods. Furthermore, the agent's optimal value function V_θ as the unique solution to the following Bellman equation:

$$V_\theta(x,\epsilon) = \max_{a \in A}\{u_{\theta_u}(x,a) + \epsilon(a) + \beta \int_{(x',\epsilon')} V_\theta(x',\epsilon') g_{\theta_g}(\epsilon'|x') f_{\theta_f}(x'|x,a) d(x',\epsilon')\}, \quad (2.1)$$

where the subscript is now explicitly recognizing the dependence of the value function on the parameters of the econometric model. By integrating out the unobserved error, we obtain the smooth value function:

$$V_\theta(x) \equiv \int_\epsilon V_\theta(x,\epsilon) g_{\theta_g}(\epsilon|x) d\epsilon.$$

Under our assumptions, the smooth value function can be shown to be the unique solution to the smooth Bellman equation, given by:

$$V_\theta(x) = \int_\epsilon \max_{a \in A} \{u_{\theta_u}(x,a) + \epsilon(a) + \beta \sum_{x' \in X} V_\theta(x') f_{\theta_f}(x'|x,a)\} g_{\theta_g}(\epsilon|x) d\epsilon. \tag{2.2}$$

We now turn to the description of the model to the conditional choice probability (CCP), denoted by $P_\theta(a|x)$, which is the model implied probability that an agent chooses action a when the observed state is x. Since the agent necessarily chooses among the actions in A, if follows that $P_\theta(|A||x) = 1 - \sum_{a \in \tilde{A}} P_\theta(a|x)$ for all $x \in X$. As a consequence, the vector of model implied CCPs can be completely characterized by $\{P_\theta(a|x)\}_{(a,x) \in \tilde{A}X}$ with $\tilde{A} \equiv \{1, \ldots, |A| - 1\}$. For the remainder of the paper, we use $\Theta_P \subset [0,1]^{|(A-1)X|}$ to denote the parameter space for the CCPs.

The CCPs is a equilibrium object of the model that is central to all derivations in this paper. As we will show in Lemma 20, it is the unique fixed point of the policy operator mapping. Given its relevance for the present study, we devote the remainder of this section to describe the CCPs and the policy operator that characterizes it.

According to the econometric model, the vector of CCPs are determined by the following equation:

$$P_\theta(a|x) \equiv \int_\epsilon 1 \left[a = \arg\max_{\tilde{a} \in A} [u_{\theta_u}(x,\tilde{a}) + \beta \sum_{x' \in X} V_\theta(x') f_{\theta_f}(x'|x,\tilde{a}) + \epsilon_{\tilde{a}}] \right] dg_{\theta_g}(\epsilon|x). \tag{2.3}$$

Notice that Eq. 2.3 can be succinctly represented as follows:

$$\{P_\theta(a|x)\}_{(a,x) \in \tilde{A}X} \;=\; \Lambda_\theta(\{V_\theta(x)\}_{x \in X}). \tag{2.4}$$

Also, notice that Eq. 2.2 can be re-written as:

$$V_\theta(x) = \sum_{a \in A} P_\theta(a|x) \left\{ u_{\theta_u}(x,a) + E_\theta[\epsilon(a)|x,a] + \beta \sum_{x' \in X} V_\theta(x') f_{\theta_f}(x'|x,a) \right\}, \quad (2.5)$$

where $E_\theta[\epsilon(a)|x,a]$ denotes the expectation of the unobservable $\epsilon(a)$ conditional on the state being x and on the optimal action being a. Under our assumptions, Hotz and Miller (1993) show that there is a one-to-one mapping that relates the CCPs and the (normalized) smooth value functions. The inverse of this mapping allows us to re-express $\{E_\theta[\epsilon(a)|x,a]\}_{(a,x) \in AX}$ as a function of the vector of CCPs. By combining this re-expression and Eq. 2.5, we can express the vector $\{V_\theta(x)\}_{x \in X}$ as a function of the vector $\{P_\theta(a|x)\}_{(a,x) \in \tilde{A}X}$. An explicit formula for such function is provided in Aguirregabiria and Mira (2002, Equation (8)), which we succinctly express as follows:

$$\{V_\theta(x)\}_{x \in X} \;=\; \varphi_\theta(\{P_\theta(a|x)\}_{(a,x) \in \tilde{A}X}). \quad (2.6)$$

By composing the mappings in Eqs. 2.4 and 2.6, we arrive to the following fixed point representation of $\{P_\theta(a|x)\}_{(a,x) \in \tilde{A}X}$:

$$\{P_\theta(a|x)\}_{(a,x) \in \tilde{A}X} \;=\; \Psi_\theta(\{P_\theta(a|x)\}_{(a,x) \in \tilde{A}X}), \quad (2.7)$$

where $\Psi_\theta \equiv \Lambda_\theta \circ \varphi_\theta$ is the policy iteration operator. As explained by Aguirregabiria and Mira (2002), this operator can be evaluated at any vector of conditional choice probabilities, optimal or not. For any arbitrary $P \equiv \{P(a|x)\}_{(a,x) \in \tilde{A}X}$, $\Psi_\theta(P)$ provides the current optimal choice probabilities of an agent whose future behavior will be to randomize over alternatives according to P.

Under the current assumptions, the policy operator Ψ_θ in Eq. 2.7 has several properties that are central to the results of this paper.

Lemma 20. *Assume Assumption 9. Then, Ψ_θ satisfies the following properties:*

(a) Ψ_θ has a unique fixed point $P_\theta \equiv \{P_\theta(a|x)\}_{(a,x)\in\tilde{A}X}$,

(b) The sequence $P^K = \Psi_\theta(P^{K-1})$ for $K \geqslant 1$, converges to P_θ for any initial $P^0 \equiv \{P^0(a|x)\}_{(a,x)\in\tilde{A}X}$,

(c) $\Psi_\theta(P)$ is twice continuously differentiable in θ and P,

(d) The Jacobian matrix of Ψ_θ with respect to P is zero at P_θ,

(e) $\Psi_\theta(P)(a|x) > 0$ for any $(a,x) \in AX$ and any θ and P.[2]

The policy operator Ψ_θ in Eq. 2.7 is a complicated function of the components of the model. The researcher intends to use this object to estimate the parameter of interest $\theta = (\alpha, \theta_f)$. In this paper, we estimate θ using a two stage procedure where, in a first stage, we use f_{θ_f} to estimate θ_f and, in a second stage, we use $\Psi_{(\alpha,\theta_f)}(P)$ and our first stage results to estimate α. In order to achieve this task successfully, we impose the following assumption.

Assumption 10. *The parameter $\theta = (\alpha, \theta_f) \in \Theta$ is identified as follows:*

(a) The parameter θ_f is uniquely identified by f_{θ_f}, i.e., $f_{\theta_{f,a}} = f_{\theta_{f,b}}$ implies $\theta_{f,a} = \theta_{f,b}$.

(b) The parameter α is identified by the fixed point condition $\Psi_{(\alpha,\theta_f)}(P) = P$ for any $(\theta_f, P) \in \Theta_f \times \Theta_P$, i.e., $\Psi_{(\alpha_a,\theta_f)}(P) = P$ and $\Psi_{(\alpha_b,\theta_f)}(P) = P$ implies $\alpha_a = \alpha_b$.

[2] This expression is an abuse of the notation for $a = |A|$ in the sense that $\Psi_\theta(P)(a,x)$ is only defined for $(a,x) \in \tilde{A}X$, i.e., it is not defined when $a = |A|$. To complete the definition, we use $\Psi_\theta(P)(|A||x) \equiv 1 - \sum_{a\in\tilde{A}} \Psi_\theta(P)(a|x)$ for any $x \in X$.

Assumption 10 is an essential to the consistent estimation of θ using a two stage procedure where, in a first stage, we use f_{θ_f} to estimate θ_f and, in a second stage, we use $\Psi_{(\alpha,\theta_f)}(P)$ and our first stage results to estimate α.[3] Magnac and Thesmar (2002) provide conditions under which Assumption 10 holds when the discount factor β and the error distribution g_{θ_g} are known. Finally, notice that this assumption implies the high level condition used by Aguirregabiria and Mira (2002, conditions (e)-(f) in Proposition 4). Under these assumptions, we can deduce certain important properties for the model implied CCPs.

Lemma 21. *Assume Assumptions 9-10. Then,*

(a) $P_\theta(a|x) > 0$ *for any* $(a,x) \in \tilde{A}X$ *and* $\sum_{a\in\tilde{A}} P_\theta(a|x)$ *for any* $x \in X$*,*

(b) P_θ *is twice continuously differentiable in* θ*,*

(c) $D_\theta P_\theta = D_\theta \Psi_\theta(P_\theta)$*,*

(d) The parameter α *is identified by* $P_{(\alpha,\theta_f)}$ *for any* $\theta_f \in \Theta_f$*, i.e.,* $\forall \theta_f \in \Theta_f$*,* $P_{(\alpha_a,\theta_f)} = P_{(\alpha_b,\theta_f)}$ *implies* $\alpha_a = \alpha_b$*.*

Thus far, we have described how the model specifies two conditional distributions: the CCPs and the transition probabilities. To complete the description of the structural model, the only remaining piece is the marginal distribution of the observed state variable. In this respect, the researcher is unwilling to make assumptions regarding the marginal distributions of observed variables, or even impose its

[3] Notice that Assumption 10 is sufficient for the identification of θ, but not necessary. For example, Assumption 10(a) could fail and θ could be identified by Ψ_θ. However, if this were the case, then the two stage procedure would not be used, and it would be replaced by a single step estimation procedure based on Ψ_θ.

stationarity. As a result, the marginal distribution of observed variables is completely unrestricted according to the model.

2.2.2 Local misspecification

The previous subsection has focused on the econometric model. This section focuses on the data generating process (DGP) and its relationship with the model. In particular, later sections will impose that the researcher observes an i.i.d. $\{(a_i, x_i, x_i')\}_{i \leqslant n}$ that is distributed according to the a DGP denoted by $\Pi_n^*(a, x, x')$. Notice that the DGP, which is a population object, is indexed by the sample size n. This indexing is important for the locally misspecification framework and will be explained throughout this section.

By definition, the underlying DGP $\Pi_n^*(a, x, x')$ can be expressed as the product of the underlying transition probability, CCPs, and marginal distribution of the state variable, i.e.,

$$\Pi_n^*(a, x, x') \;=\; f_n^*(x'|a, x) \;\times\; P_n^*(a|x) \;\times\; m_n^*(x), \tag{2.8}$$

where the superscript with asterisk denotes true values and $m_n^*(\cdot)$ denotes the true marginal distributions of the state variable.

The econometric model in Section 2.2.1 specifies $P_\theta(a|x)$ as a model for $P_n^*(a|x)$, $f_{\theta_f}(x'|a, x)$ as a model for $f_n^*(x'|a, x)$, and avoids imposing any restrictions on $m_n^*(x)$ (i.e. it treats this non-parametrically). In this paper, the proposed econometric model is allowed to be incorrect or misspecified, i.e., the model is an incorrect representation of the DGP. Formally speaking, this implies that:

$$\inf_{(\alpha, \theta_f) \in \Theta_\alpha \times \Theta_f} \| \; (P_{(\alpha, \theta_f)} - P_n^*) \; , \; (f_{\theta_f} - f_n^*) \; \| \geqslant 0. \tag{2.9}$$

The above equation can hold strictly whenever the mapping between parameters and DGPs given by $(\alpha, \theta_f) \to (P_\theta, f_{\theta_f})$ is not surjective or onto. This should not be considered unusual given that these econometric models are heavily parametrized in practice.

Eq. 2.9 allows the econometric model to be incorrect. If we allow for arbitrary, i.e., global misspecification, it will be hard to obtain strong conclusions about the properties of the estimators. For this reason (among others), this paper will consider an asymptotic framework such that the degree of misspecification for the two stage population problem vanishes as the sample size increases.[4] This is the content of the following assumption.

Assumption 11. *(a)* $\{(J_n^*, f_n^*)\}_{n \geqslant 1}$ *converge to the limiting distributions* (J^*, f^*) *at* n^δ*-rate for some* $\delta > 0$*. In particular,*

$$n^\delta \begin{pmatrix} J_n^* - J^* \\ f_n^* - f^* \end{pmatrix} \to \begin{pmatrix} B_J \\ B_f \end{pmatrix}$$

with $||(B_J, B_f)|| < \infty$*.*

(b) The econometric model is correctly specified according to the limiting distributions (J^*, f^*)*. In the absence of a model for the marginal distribution of the state variable, this is equivalent to:*

$$\inf_{(\alpha,\theta_f) \in \Theta_\alpha \times \Theta_f} \| \ (P_{(\alpha,\theta_f)} - P^*) \ , \ (f_{\theta_f} - f^*) \ \| = 0,$$

[4] In principle, this assumption still allows for a single stage procedure to be non-locally misspecified (e.g. even correctly specified). There are several reasons why this possibility is not particularly problematic. First, the first step of the two stage procedure is typically non parametric and thus, the two and single stage procedures are necessarily misspecified in the same fashion. Second, as long as researchers only consider two stage estimation methods, the possibility of a correctly specified single stage method becomes irrelevant.

where $P^*(a|x) \equiv J^*(a,x)/\sum_{\tilde{a}\in A} J^*(\tilde{a},x)$ *for all* $(a,x) \in \tilde{A}X$.

Assumption 11 defines the local misspecification framework used in our paper. While Eq. 2.9 allows the econometric model to be incorrect (i.e. the DGP and the model can differ), Assumption 11 implies that the limiting distribution of the data can be correctly represented by the model. It is relevant to note that Assumption 11(b) allows for misspecification to vanish at an arbitrarily low rate.

In practice, researchers specify econometric models that may contain non vanishing specification errors, i.e., global misspecification. As we reveal in the paper, our local misspecification approach has two advantages relative to the global approach. First, regular misspecification analysis is typically intractable and general results are almost impossible to obtain. In contrast, under local misspecification, we can derive very general results (see Theorems 23, 24, and 25). Also, under global misspecification, the different estimation methods typically converge to different pseudo-true parameters, making results hard to interpret and compare. In contrast, under local misspecification, Theorem 22 shows there is a unique well-defined true limiting parameter value (α^*, θ_f^*).

Theorem 22. *Under Assumptions 10-11(b), there is a unique parameter* $(\alpha^*, \theta_f^*) \in \Theta$ *that solves* $P_{(\alpha^*,\theta_f^*)} = P^*$ *and* $f_{\theta_f^*} = f^*$.

2.3 Inference in the locally misspecified model

Following the literature on CCP estimation (e.g. Rust (1987), Hotz and Miller (1993), and Aguirregabiria and Mira (2002)), this paper focuses on two stage estimation procedures. In the first stage, the researcher estimates $\theta_f \in \Theta_f$ and, in a second

stage, the researcher estimates $\alpha \in \Theta_\alpha$ based on the sample CCPs, taking the first stage results as given. The main reason for their popularity of two stage estimators is their simplicity, relative to a single stage estimation method.[5]

Our paper focuses on two stage estimators based on the K-step sequential policy iteration (PI) algorithm developed by Aguirregabiria and Mira (2002). By definition, the K-step PI estimator of α is given by:

$$\hat{\alpha}_n^K \equiv \arg\max_{\alpha \in \Theta_\alpha} Q_n(\alpha, \hat{\theta}_{f_n}, \hat{P}_n^{K-1}), \tag{2.10}$$

where $Q_n : \Theta_\alpha \times \Theta_f \times \Theta_P \to \mathbb{R}$ is a sample objective function chosen by the researcher, $\hat{\theta}_{f_n}$ is the first step estimator, and $\hat{P}_n^K$ denotes the K-step estimator of the CCPs which is iteratively defined as follows. If we let $\hat{P}_n^0$ denote the zero-step or preliminary CCP estimator, the K-step CCP estimator $\hat{P}_n^K$ for some $K \in \mathbb{N}$ is iteratively defined as follows:

$$\hat{P}_n^K \equiv \Psi_{(\hat{\alpha}_n^K, \hat{\theta}_{f_n})}(\hat{P}_n^{K-1}).$$

In this paper, we consider two possible choices of sample objective functions Q_n, leading to different estimation procedures. Section 2.4.2 considers maximum likelihood type (ML) estimation while Section 2.4.3 considers minimum distance (MD) estimation. As mentioned earlier, the K-step PI estimator is a very general class of estimators that include popular estimators such as Rust (1987)'s nested fixed point estimator ($K = \infty$ and ML estimation), Hotz and Miller (1993)'s conditional choice probability estimator ($K = 1$ and MD estimation), Aguirregabiria and Mira

[5] Results for single step estimation procedures are relatively easy to deduce from our analysis. Basically, all one needs to do is to consider that the entire parameter vector is estimated on the second stage.

(2002)'s nested algorithm estimator ($K \in \mathbb{N}$ and ML estimation), and certain cases of Pesendorfer and Schmidt-Dengler (2008)'s least squares estimator ($K = 1$ and MD estimation).

In order to obtain results, we impose two high-level assumptions that involve the sample objective function Q_n in Eq. 2.10 and the first stage estimator in θ_f^*. While these conditions are admittedly high-level, they are standard in extremum estimator problems and they will be verified for ML and MD estimators in Sections 2.4.2 and 2.4.3, respectively.

Assumption 12. $\alpha^* \in Int(\Theta_\alpha)$.

Assumption 13. *Let $\mathcal{N}$ denote an arbitrary small neighborhood of (α, θ_f, P) around $(\alpha^*, \theta_f^*, P^*)$. Then,*

(a) $\sup_{\alpha \in \Theta_\alpha} |Q_n(\alpha, \hat{\theta}_{f_n}, \hat{P}_n^0) - Q_\infty(\alpha, \theta_f^*, P^*)| = o_{P_n}(1)$.

(b) $Q_\infty(\alpha, \theta_f^*, P^*)$ *is uniquely maximized at* α^*.

(c) $Q_n(\alpha, \theta_f, P)$ *is twice continuously differentiable in* α *and* (α, θ_f, P) *for all* $(\alpha, \theta_f, P) \in \mathcal{N}$ *w.p.a.1.*

(d)

$$\sup_{(\alpha, \theta_f, P) \in \mathcal{N}} \|\partial Q_n(\alpha, \theta_f, P)/\partial\alpha\partial\lambda - \partial Q_\infty(\alpha, \theta_f, P)/\partial\alpha\partial\lambda\| = o_{P_n}(1)$$

for $\lambda \in \{\alpha, \theta_f, P\}$.

(e) $\partial Q_\infty(\alpha, \theta_f, P)/\partial\alpha\partial\alpha'$ *is continuous and non-singular at* $(\alpha^*, \theta_f^*, P^*)$.

(f) $\partial Q_\infty(\alpha^*, \theta_f^*, P^*)/\partial\lambda\partial P = \mathbf{0}_{d_\lambda \times |AX|}$ *for* $\lambda \in \{\alpha, \theta_f\}$.

Assumption 14. *(a)*

$$n^{\min\{1/2,\delta\}}[\partial Q_n(\alpha^*, \theta_f^*, P^*)/\partial\alpha', (\hat{\theta}_{f,n} - \theta_f^*)] \xrightarrow{d} [\zeta_1, \zeta_2]$$

for some random variable ζ.

(b) $n^{\min\{1/2,\delta\}}(\hat{P}_n^0 - P^*) = O_{p_n}(1)$.

We now briefly comment on these conditions. Assumption 12 is a standard assumption in extremum estimators and is also required in Aguirregabiria and Mira (2002). Assumptions 13-14 are high level assumptions that will be shown as results for specific estimators in future sections. For the most part, Assumption 13 is relatively standard condition used to derive results for extremum estimators. The exception to this is Assumption 13(f), which will be shown to direct consequence of the zero Jacobian property derived in Lemma 20(d). Assumption 14 reflects the fact that the econometric model is locally misspecification at a rate of n^δ in that the estimators of the components of the model are converging to the limiting parameter at rate of $n^{\min\{1/2,\delta\}}$, i.e., the slowest rate between local misspecification and sampling variation. These high level assumptions will be shown as results for specific estimators in future sections.

Under these high-level assumptions, Theorem 23 establishes the asymptotic distributions of general two stage K-step policy iteration estimator.

Theorem 23. *Assume Assumptions 9-14. For any* $K \geqslant 1$,

$$n^{\min\{1/2,\delta\}}(\hat{\alpha}_n^K - \alpha^*) \xrightarrow{d} (-\frac{\partial Q_\infty(\alpha^*, \theta_f^*, P^*)}{\partial\alpha\partial\alpha'})^{-1}[\zeta_1 + \frac{\partial Q_\infty(\alpha^*, \theta_f^*, P^*)}{\partial\alpha'\partial\theta_f}\zeta_2].$$

Theorem 23 reveals two important properties of the asymptotic distributions of two stage K-step policy iteration estimator. First, as a consequence of the local misspecification at a rate of n^{δ}, the estimator converges to the limiting parameter value at a rate of $n^{\min\{1/2,\delta\}}$, i.e., the slowest rate among the local misspecification and the sampling variation. Second, the asymptotic distribution of the estimator is completely unaffected by the number of iteration steps K. In the particular case of ML estimators, this second feature effectively extends the main result of Aguirregabiria and Mira (2002) from correctly specified models to arbitrarily locally misspecified ones.

The invariance of the asymptotic distribution of the two stage K-step policy iteration estimator to K is perhaps one of the main results in this paper. The intuition for this result is as follows. The fact that the model is locally misspecified at a rate of n^{δ} affects this asymptotic distribution of $(\hat{\alpha}_n^K - \alpha^*)$ in three ways, as it follows from Eq. 2.10. First, it affects the convergence of the gradient of the sample criterion function Q_n. Second, it affects the convergence of the first stage estimator $\hat{\theta}_{f,n}$. Third, it affects the convergence of the $(K-1)$-step CCP estimator $\hat{P}_n^{K-1}$. By definition, the only one of these pieces that varies with the number of iteration steps K is the third one, i.e., $\hat{P}_n^{K-1}$. However, the extremum estimator problem under consideration satisfies the zero Jacobian property imposed in Assumption 13(f). According to our formal arguments, the zero Jacobian property is responsible for effectively erasing the influence of the CCP estimation at each iteration step. As a consequence of this, the asymptotic distribution of the two stage K-step policy iteration estimator is invariant to the number of iteration steps K. In fact, we show in Theorem 23 that all K-step policy iteration estimator are asymptotically equivalent. One of the

remarkable features of this result is that it holds regardless of the rate of the local misspecification (determined by δ).

2.4 Applications of the general result

In this section, we apply the main result in Theorem 23 to classes of estimators used in practice. The main sections are Sections 2.4.2 and 2.4.3, where we apply our general result in Theorem 23 to two stage K-ML or K-MD estimation, respectively. The remainder of the sections provide supporting derivations for these two subsections. Section 2.4.1 discusses a reasonable framework for the first stage and the preliminary CCP estimators.

Throughout this section, we presume that the researcher observes an i.i.d. sample $\{(a_i, x_i, x_i')\}_{i \leqslant n}$ that is distributed according to the true DGP.

Assumption 15. *$\{(a_i, x_i, x_i')\}_{i \leqslant n}$ is an i.i.d. sample distributed according to the joint distribution $\Pi_n^*(a, x, x')$.*

Under this context, it is natural to consider the sample analogue estimator of the true DGP Π_n^*, denoted by $\hat{\Pi}_n = \{\hat{\Pi}_n(a, x, x')\}_{(a,x,x') \in AX^2}$ and defined by:

$$\hat{\Pi}_n(a, x, x') \equiv \sum_{i=1}^{n} 1[x_i = x, a_i = a, x_i' = x']/n.$$

2.4.1 Preliminary estimators

This section discusses our formal framework of the first-stage estimator $\hat{\theta}_{f,n}$ and the zero*th*-step CCP estimator $\hat{P}_n^0$. Rather than assuming a specific estimator for these objects, we consider a general framework that covers several popular examples of estimators and that satisfy our high-level assumption 14.

Under Assumptions 10-11, Theorem 22 implies that the true limiting CCPs P^* and state transition probabilities f^* are identified. In turn, Assumption 10 implies that the latter identifies the true limiting parameter θ_f^*. Under these conditions and Assumption 15, it is reasonable to presume that the researcher proposes preliminary estimators of θ_f^* and P^* that are smooth functions of the sample analogue estimator of $\hat{\Pi}_n$.

Assumption 16. *The preliminary first stage and CCP estimators are defined according to the following equation:*

$$(\hat{\theta}_{f,n}, \hat{P}_n^0) \;=\; G_n(\hat{\Pi}_n),$$

where $\{G_n : \mathbb{R}^{|AX^2|} \to \mathbb{R}^{d_f} \times \mathbb{R}^{|AX^2|}\}_{n \geqslant 1}$ *is a sequence of functions that satisfies the following properties. For an arbitrary small neighborhood of* Π^* *denoted by* $\mathcal{N}_{\Pi^*}$*,*

(a) $\sup_{\Pi \in \mathcal{N}_{\Pi^*}} ||G_n(\Pi) - G(\Pi)|| = o_{p_n}(n^{-min\{1/2,\delta\}})$.

(b) $G(\Pi)$ *is continuously differentiable for any* $\Pi \in \mathcal{N}_{\Pi^*}$*,*

(c) $(\theta_f^*, P^*) = G(\Pi^*)$.

Assumption 16 is very mild. In fact, given that the state and action spaces are finite, this is automatically satisfied if we use a sample analogue estimator of the CCP, i.e.,

$$\hat{P}_n^0(a|x) \;\equiv\; \frac{\sum_{i=1}^n 1[a_i = a, x_i = x]}{\sum_{i=1}^n 1[x_i = x]} = \frac{\sum_{\tilde{x}' \in X} \hat{\Pi}_n(a, x, \tilde{x}')}{\sum_{(\check{a},\check{x}') \in AX} \hat{\Pi}_n(\check{a}, x, \check{x}')} \tag{2.11}$$

and a non-parametric model for the first stage, i.e., $\theta_f \equiv \{f(x'|a,x)\}_{(a,x,x')\in AX^2}$ that is also estimated by sample analogues, i.e., $\hat{\theta}_{f,n} \equiv \{\hat{f}_n(x'|a,x)\}_{(a,x,x')\in AX^2}$ with:

$$\hat{f}_n(x'|a,x) \equiv \frac{\sum_{i=1}^n 1[a_i = a, x_i = x, x_i' = x']}{\sum_{i=1}^n 1[a_i = a, x_i = x]} = \frac{\hat{\Pi}_n(a,x,x')}{\sum_{\tilde{x}'\in X} \hat{\Pi}_n(a,x,\tilde{x}')}. \tag{2.12}$$

For the sake of completeness, Lemma 30 in the appendix formally shows that the sample analogue estimators in Eqs. 2.11 and 2.12 satisfy Assumption 16.[6]

2.4.2 ML estimation

We now consider the case of partial ML estimation, as considered by Aguirregabiria and Mira (2002). This is a special case of a two stage K-step PI estimation in Eq. 2.10 in which the sample objective function Q_n is the pseudo-likelihood function. By definition, this is given by:

$$Q_n^{ML}(\alpha, \theta_f, P) \equiv n^{-1} \sum_{i=1}^n \ln \Psi_{(\alpha,\theta_f)}(P)(a_i, x_i).$$

Deriving the asymptotic distribution of the K-ML estimator requires the following regularity condition.

Assumption 17. *$D_{\theta_\alpha}\Psi_\theta(P)$ is a full rank matrix at $(\theta^*, P_{\theta*})$.*

Assumption 17 is the low-level condition connected to the non-singularity requirement in Assumption 13(e). This condition is critical for the consistency of the K-ML

[6] One practical problem with the sample analogue estimators proposed in Eqs. 2.11 and 2.12 is that they would not be properly defined if either $\sum_{(\check{a},\check{x}')\in AX} \hat{\Pi}_n(\check{a},x,\check{x}') = 0$ or $\sum_{\tilde{x}'\in X} \hat{\Pi}_n(a,x,\tilde{x}') = 0$ for any $(a,x) \in AX$. See Hotz et al. (1994) and Pesendorfer and Schmidt-Dengler (2008, Page 914). Of course, this is a small sample issue that will disappear asymptotically. In the light of this difficulty, the sample analogue estimators are often replaced by Kernel smoother estimators. These estimators can also be shown to satisfy Assumption 16 under appropriate restrictions on the bandwidth parameter. We omit this proof for reasons of brevity.

estimators and it is also assumed in Aguirregabiria and Mira (2002). Since the α has been assumed to be identified by $\Psi_\theta(P) = P$, (and, thus, locally identified by it), Assumption 17 is equivalent to the regularity conditions in Rothenberg (1971, Theorem 1).

Theorem 24 characterizes the asymptotic distribution of the K-ML estimator under local misspecification. As we have already discussed, it is a direct consequence of Theorem 23.

Theorem 24. *Assume Assumptions 9-12 and 16-17. Then, for any* $K, \tilde{K} \geqslant 1$,

$$n^{\min\{1/2,\delta\}}(\hat{\alpha}_n^{K-ML} - \alpha^*) = n^{\min\{1/2,\delta\}}(\hat{\alpha}_n^{\tilde{K}-ML} - \alpha^*) + o_{P_n}(1)$$

and its asymptotic distribution is

$$\Upsilon_{ML} \cdot \left[\begin{array}{cc} \Sigma & -\frac{\partial P_{\theta *}}{\partial \theta_f}' \end{array} \right] \cdot N\left(B \cdot 1[\delta \leqslant 1/2], \Omega \cdot 1[\delta \geqslant 1/2]\right),$$

where the vector B *and the matrix* Ω *are as defined in Lemma 27, the matrix* Σ *is as defined in Lemma 2.23, and the matrix* Υ_{ML} *is given by:*

$$\Upsilon_{ML} \equiv \left(\frac{\partial P_{\theta *}}{\partial \alpha} (\Sigma \Omega_{JJ} \Sigma')^{-1} \frac{\partial P_{\theta *}}{\partial \alpha}' \right)^{-1} \frac{\partial P_{\theta *}}{\partial \alpha} (\Sigma \Omega_{JJ} \Sigma')^{-1},$$

and Ω_{JJ} *is as defined in Lemma 27 as one of the components of the matrix* Ω*, i.e.,*

$$\Omega \equiv \left(\begin{array}{cc} \Omega_{JJ} & \Omega_{Jf} \\ \Omega_{Jf}' & \Omega_{ff} \end{array} \right). \tag{2.13}$$

The qualitative conclusions of Theorem 24 are as in Theorem 23, i.e., all K-ML estimators are asymptotically equivalent and thus share the asymptotic distribution,

regardless of the number of iterations K. As we also knew from Theorem 23, the rate of convergence of the K-ML estimator is $n^{\min\{1/2,\delta\}}$, which clearly depends on the rate of the local misspecification δ. In quantitative terms, Theorem 24 provides a precise formula for the asymptotic distribution, which is normal and with mean and variance given by:

$$\begin{aligned} AB_{ML} &= \Upsilon_{ML} \left[\begin{array}{cc} \Sigma & -\frac{\partial P_{\theta *}}{\partial \theta_f}' \end{array} \right] B \cdot 1[\delta \leqslant 1/2], \\ AV_{ML} &= \Upsilon_{ML} \left[\begin{array}{cc} \Sigma & -\frac{\partial P_{\theta *}}{\partial \theta_f}' \end{array} \right] \Omega \left[\begin{array}{c} \Sigma' \\ -\frac{\partial P_{\theta *}}{\partial \theta_f} \end{array} \right] \Upsilon'_{ML} \cdot 1[\delta \geqslant 1/2]. \end{aligned}$$

As these equations reveal, the presence of (asymptotic) bias and variance depends on the rate of local misspecification δ relative to the rate of sampling variation. In the case of $\delta < 1/2$, the asymptotic distribution has zero bias and the variance coincides exactly with the one obtained under correct specification. In this case, the local misspecification is irrelevant relative to sampling error and it can be safely ignored for practical purposes. The opposing situation occurs when $\delta > 1/2$, as the asymptotic distribution has zero variance and the estimator converges in probability to the bias term. In such a case, the local misspecification is overwhelming relative to sampling error and it entire dominates the asymptotic distribution. Finally, we have the interesting knife edge case in which $\delta = 1/2$ and non-negligible asymptotic variance and bias coexist. In such a case, an adequate characterization of the precision of the estimator is given by the asymptotic mean squared error, given by:

$$AMSE_{ML} = \Upsilon_{ML} \left[\begin{array}{cc} \Sigma & -\frac{\partial P_{\theta *}}{\partial \theta_f}' \end{array} \right] [\Omega + BB'] \left[\begin{array}{c} \Sigma' \\ -\frac{\partial P_{\theta *}}{\partial \theta_f} \end{array} \right] \Upsilon'_{ML}. \tag{2.14}$$

We conclude the subsection with a minor point regarding the asymptotic optimality of the K-ML estimator. The K-ML estimator considered in this section is a partial maximum likelihood estimator in the sense that the maximum likelihood is only applied to the second step, effectively ignoring the sampling error in the preliminary estimation step. Because of this feature, usual optimality results for maximum likelihood estimation need not apply. In fact, the next subsection will describe a K-MD estimator that is more efficient than the K-ML, both in asymptotic variance and in asymptotic mean squared error.

2.4.3 MD estimation

We now consider the case of MD estimation, defined as special case of a two stage K-step PI estimation in Eq. 2.10 in which the sample objective function Q_n is given by:

$$Q_n^{MD}(\alpha, \theta_f, P) \equiv -[\hat{P}_n - \Psi_{(\alpha,\theta_f)}(P)]'\hat{W}_n[\hat{P}_n - \Psi_{(\alpha,\theta_f)}(P)]. \tag{2.15}$$

where $\hat{W}_n$ is the weight function. In the special case in which $K = 1$ and the preliminary estimator $\hat{P}_n^0$ is equal to sample frequency estimator $\hat{P}_n$, our two step K-MD estimator coincides with the least-squares estimator considered in Pesendorfer and Schmidt-Dengler (2008, Eqs. (18)-(19)).[7] Deriving the asymptotic distribution of the K-MD estimator requires the following condition regarding the weight matrix.[8]

---

[7] In all fairness, their framework allows also for $\hat{P}_n$ used in Eq. 2.15 to differ from the sample frequency estimator. Nonetheless, this observation would also apply to our K-MD estimation framework.

[8] In principle, we could allow for the weight matrices to be functions of the parameters of the problem, i.e, we could allow for $\hat{W}_n(\alpha, \theta_f, P)$ and $W^*(\alpha, \theta_f, P)$. In such cases, one could obtain the same results by imposing additional conditions in Assumption 18 and by using slightly longer theoretical arguments.

Assumption 18. *$\hat{W}_n = W^* + o_{p_n}(1)$, where W^* is positive definite and symmetric.*

Theorem 25 characterizes the asymptotic distribution of the K-MD estimator under local misspecification. As in the case of Theorem 24, it is a direct application of the general result in Theorem 23.

Theorem 25. *Assume Assumptions 9-12 and 16-18. Then, for any $K, \tilde{K} \geqslant 1$,*

$$n^{\min\{1/2,\delta\}}(\hat{\alpha}_n^{K-MD} - \alpha^*) = n^{\min\{1/2,\delta\}}(\hat{\alpha}_n^{\tilde{K}-MD} - \alpha^*) + o_{P_n}(1)$$

and its asymptotic distribution is

$$\Upsilon_{MD(W^*)} \cdot \left[\begin{array}{cc} \Sigma & -\frac{\partial P_{\theta^*}}{\partial \theta_f}' \end{array} \right] \cdot N\left(B \cdot 1[\delta \leqslant 1/2], \Omega \cdot 1[\delta \geqslant 1/2]\right),$$

where the vector B and the matrix Ω are as defined in Lemma 27, the matrix Σ is as defined in Lemma 2.23, and the matrix $\Upsilon_{MD}(W^*)$ is given by:

$$\Upsilon_{MD}(W^*) \equiv \left(\frac{\partial P_{\theta^*}}{\partial \alpha} W^* \frac{\partial P_{\theta^*}}{\partial \alpha}' \right)^{-1} \frac{\partial P_{\theta^*}}{\partial \alpha} W^*.$$

Remark 26. *The asymptotic distribution of the K-ML estimator is a special case of that of the K-MD estimator with $W^* = W_{ML} \equiv (\Sigma \Omega_{JJ} \Sigma')^{-1}$ where Ω_{JJ} is as defined in Lemma 27 as one of the components of the matrix Ω.*

Once again, the qualitative conclusions of Theorem 25 are as in Theorem 23, i.e., all K-MD estimators are asymptotically equivalent and thus share the asymptotic distribution, regardless of the number of iterations K. The rate of convergence of the K-MD estimator is $n^{\min\{1/2,\delta\}}$, which depends crucially on the rate of the local

misspecification δ. In quantitative terms, Theorem 25 shows that the asymptotic distribution of the K-MD estimator is normal and with mean and variance given by:

$$\begin{aligned} AB_{MD}(W^*) &= \Upsilon_{MD}(W^*) \left[\begin{array}{cc} \Sigma & -\frac{\partial P_{\theta *}}{\partial \theta_f}' \end{array} \right] B \cdot 1[\delta \leqslant 1/2], \\ AV_{MD}(W^*) &= \Upsilon_{MD}(W^*) \left[\begin{array}{cc} \Sigma & -\frac{\partial P_{\theta *}}{\partial \theta_f}' \end{array} \right] \Omega \left[\begin{array}{c} \Sigma' \\ -\frac{\partial P_{\theta *}}{\partial \theta_f} \end{array} \right] \Upsilon_{MD}(W^*)' \cdot 1[\delta \geqslant 1/2]. \end{aligned}$$

The presence of (asymptotic) bias and variance depends heavily on the value of δ and in exactly the same way as for the K-ML estimator. In the interesting knife edge case in which $\delta = 1/2$ and non-negligible asymptotic variance and bias coexist with asymptotic mean-squared error given by:

$$AMSE_{MD}(W^*) = \Upsilon_{MD}(W^*) \left[\begin{array}{cc} \Sigma & -\frac{\partial P_{\theta *}}{\partial \theta_f}' \end{array} \right] [\Omega + BB'] \left[\begin{array}{c} \Sigma' \\ -\frac{\partial P_{\theta *}}{\partial \theta_f} \end{array} \right] \Upsilon_{MD}(W^*)'. \tag{2.16}$$

Once again, the qualitative conclusions of Theorem 25 are as in Theorem 23, i.e., all K-MD estimators are asymptotically equivalent and thus share the asymptotic distribution, regardless of the number of iterations K. The rate of convergence of the K-MD estimator is $n^{\min\{1/2,\delta\}}$, which depends crucially on the rate of the local misspecification δ. In quantitative terms, Theorem 25 shows that the asymptotic distribution of the K-MD estimator is normal and with mean and variance given by:

$$\begin{aligned} AB_{MD}(W^*) &= \Upsilon_{MD}(W^*) \left[\begin{array}{cc} \Sigma & -\frac{\partial P_{\theta *}}{\partial \theta_f}' \end{array} \right] B \cdot 1[\delta \leqslant 1/2], \\ AV_{MD}(W^*) &= \Upsilon_{MD}(W^*) \left[\begin{array}{cc} \Sigma & -\frac{\partial P_{\theta *}}{\partial \theta_f}' \end{array} \right] \Omega \left[\begin{array}{c} \Sigma' \\ -\frac{\partial P_{\theta *}}{\partial \theta_f} \end{array} \right] \Upsilon_{MD}(W^*)' \cdot 1[\delta \geqslant 1/2]. \end{aligned}$$

The presence of (asymptotic) bias and variance depends heavily on the value of δ and in exactly the same way as for the K-ML estimator. In the interesting knife edge case in which $\delta = 1/2$ and non-negligible asymptotic variance and bias coexist with asymptotic mean-squared error given by:

$$AMSE_{MD}\left(W^*\right) = \Upsilon_{MD}\left(W^*\right)\left[\begin{array}{cc} \Sigma & -\frac{\partial P_{\theta *}}{\partial \theta_f}' \end{array}\right]\left[\Omega + BB'\right]\left[\begin{array}{c} \Sigma' \\ -\frac{\partial P_{\theta *}}{\partial \theta_f} \end{array}\right]\Upsilon_{MD}\left(W^*\right)'. \tag{2.17}$$

We can now briefly discuss the optimality in the choice of W^* in K-MD estimation. Since these estimators can have bias, variance, and both, we deem the asymptotic means squared error to be a reasonable optimality criterion. We divide the analysis into the three cases. First, consider the case in which local misspecification is asymptotically irrevant, i.e., $\delta < 1/2$. In this case, the K-MD estimator presents no asymptotic bias and a variance (and mean squared error) equal to Eq. 2.17. By standard argument in GMM estimation, the minimum asymptotic variance (in matrix sense) among K-MD estimators is given by:

$$AV^* \equiv \left(\frac{\partial P_{\theta *}}{\partial \alpha}\left[\left[\begin{array}{cc} \Sigma & -\frac{\partial P_{\theta *}}{\partial \theta_f}' \end{array}\right]\Omega\left[\begin{array}{c} \Sigma' \\ -\frac{\partial P_{\theta *}}{\partial \theta_f} \end{array}\right]\right]^{-1}\frac{\partial P_{\theta *}}{\partial \alpha}'\right)^{-1}.$$

This can be achieved by the following feasible choice of W^* as follows:

$$W^*_{AV} \equiv \left(\left[\begin{array}{cc} \Sigma & -\frac{\partial P_{\theta *}}{\partial \theta_f}' \end{array}\right]\Omega\left[\begin{array}{c} \Sigma' \\ -\frac{\partial P_{\theta *}}{\partial \theta_f} \end{array}\right]\right)^{-1}.$$

As we point out in Remark 26, notice that the K-ML estimator the same asymptotic distribution as the K-MD estimator with $W^*_{ML} = \left(\Sigma\Omega_{JJ}\Sigma'\right)^{-1}$. Unless there are

special conditions on the econometric model (e.g. $\partial P_{\theta *}/\partial \theta_f = \mathbf{0}_{|\tilde{A}X| \times d_f}$), the K-ML is not necessarily optimal in the sense of achieving a minimum variance among K-MD estimator.

Next, consider the case in which local misspecification is asymptotically overwhelming, i.e., $\delta > 1/2$. In this case, the K-MD estimator presents no asymptotic variance and it converges at a pure bias term equal to Eq. 2.17. By definition, the asymptotic squared bias coincides with the asymptotic mean squared error. Once again, standard arguments imply that the minimum asymptotic mean squared error (in matrix sense) among all K-MD estimators is given by:

$$AMSE^* \equiv \left(\frac{\partial P_{\theta *}}{\partial \alpha} \left[\left[\begin{array}{cc} \Sigma & -\frac{\partial P_{\theta *}}{\partial \theta_f}' \end{array} \right] BB' \left[\begin{array}{c} \Sigma' \\ -\frac{\partial P_{\theta *}}{\partial \theta_f} \end{array} \right] \right]^{-1} \frac{\partial P_{\theta *}}{\partial \alpha}' \right)^{-1}.$$

This can be achieved by the following infeasible choice of W^* as follows:

$$W^*_{AMSE^*} \equiv \left(\left[\begin{array}{cc} \Sigma & -\frac{\partial P_{\theta *}}{\partial \theta_f}' \end{array} \right] BB' \left[\begin{array}{c} \Sigma' \\ -\frac{\partial P_{\theta *}}{\partial \theta_f} \end{array} \right] \right)^{-1}.$$

This choice of weight matrix is unfeasible because it depends on the bias of the econometric model which, by definition, an unknown feature to the researcher.

Finally, consider the knife-edge case in which local misspecification is of the same rate as sampling error, i.e., $\delta = 1/2$. In this case, the K -MD estimator presents both bias and variance given by Eqs. 2.17 and 2.17, respectively, and the asymptotic mean squared error is as in Eq. 2.17. One more time, standard arguments imply that the minimum asymptotic mean squared error (in matrix sense) among all K-

MD estimators is given by:

$$AMSE^{*} \equiv \left(\frac{\partial P_{\theta *}}{\partial \alpha} \left[\left[\begin{array}{cc} \Sigma & -\frac{\partial P_{\theta *}}{\partial \theta_f}' \end{array} \right] (\Omega + BB') \left[\begin{array}{c} \Sigma' \\ -\frac{\partial P_{\theta *}}{\partial \theta_f} \end{array} \right] \right]^{-1} \frac{\partial P_{\theta *}}{\partial \alpha}' \right)^{-1}.$$

This can be achieved by the following infeasible choice of W^{*} as follows:

$$W^{*}_{AMSE*} \equiv \left(\left[\begin{array}{cc} \Sigma & -\frac{\partial P_{\theta *}}{\partial \theta_f}' \end{array} \right] (\Omega + BB') \left[\begin{array}{c} \Sigma' \\ -\frac{\partial P_{\theta *}}{\partial \theta_f} \end{array} \right] \right)^{-1}.$$

Once again, this choice of weight matrix is unfeasible because it depends on the bias of the econometric model which, by definition, an unknown feature to the researcher. However, we point out that the presence of bias in the estimation might generate a situation in which a feasible choice of weight matrix that minimizes asymptotic variance could result in a large value of asymptotic squared bias and, consequently, asymptotic mean squared error.[9]

2.5 Monte Carlo simulations

This section investigates the finite sample properties of the two-stage estimators considered in previous sections by means of a Monte Carlo experiment. We simulate data using the classical bus engine replacement problem studied by Rust (1987).

2.5.1 A misspecified econometric model

In each period $t = 1, \ldots, T \equiv \infty$, the bus owner has to decide whether to replace the bus engine or not to minimize the discounted present value of his costs. In any

[9] This is the case in some of our Monte Carlo simulations. Using the feasible weight matrix W^{*}_{AV} that minimizes asymptotic variance can generate a large asymptotic mean squared error, even larger than the one that could be achieved by simply choosing an identity weight matrix $W^{*} = \mathbf{I}$.

representative period, his choice is denoted by $a \in A = \{1, 2\}$, where $a = 2$ represents replacing the engine and $a = 1$ represents not replacing the engine, and the current engine mileage is denoted by $x \in X \equiv \{1, \ldots, |X|\}$ with $|X| \equiv 25$.

The researcher assumes the following specification for the deterministic part of the utility (profit) function:

$$u_{\theta_u}(x, a) \;=\; -\theta_{u,1} \cdot 1[a = 2] \; - \theta_{u,2} \cdot 1[a = 1]x, \tag{2.18}$$

where $\theta_u \equiv (\theta_{u,1}, \theta_{u,2}) \in \Theta_u \equiv [-B, B]^2$ with $B = 10$. In addition, the researcher also assumes that the unobserved errors are distributed according to an extreme value type I distribution, independent of x, i.e.,

$$g_{\theta_g}(\epsilon = e|x) = \prod_{a \in A} \exp(e(a)) \exp(-\exp(e(a))), \tag{2.19}$$

which does not have any unknown parameters (i.e. θ_g is known). Finally, the observed state is assumed to evolve according to the following Markov chain:

$$\begin{aligned} f_{\theta_f}(x'|x, a) \;=\; & (1 - \theta_f) \cdot 1[a = 1, x' = \min\{x + 1, |X|\}] \\ & + \theta_f \cdot 1[a = 1, x' = x] \; + \; 1[a = 2, x' = 1], \end{aligned} \tag{2.20}$$

where $\theta_f \in \Theta_f \equiv [0, 1]$. The researcher correctly assumes that $\beta = 0.9999$. His goal is to estimate $\theta = (\alpha, \theta_f) \in \Theta = \Theta_\alpha \times \Theta_f$ with $\alpha \equiv \theta_u \in \Theta_\alpha = \Theta_u$ and $\theta_f \in \Theta_f$.

The researcher has correctly specified the error distribution and the state transition distribution, which satisfies Eq. 2.20 with $\theta_f = 0.25$. Unfortunately, he has incorrectly specified the utility function. Instead of the linear function in Eq. 2.18, the utility function is the following quadratic function:

$$u_{\theta_u,n}(x, a) \;=\; -\theta_{u,1} \cdot 1[a = 2] \; - \theta_{u,2} \cdot 1[a = 1]x \; - \theta_{u,3,n} \cdot 1[a = 1]x^2 \tag{2.21}$$

with $\theta_{u,1} = 1$, $\theta_{u,2} = 1/|X|$, and $\theta_{u,3,n} = (10/|X|^2) \cdot n^{-\delta}$ with $\delta \in \{1/4, 1/3, 1/2, 1\}$. The fact that $\theta_{u,3,n} \neq 0$ implies that the econometric model specified by the researcher is incorrectly specified. However, the fact that $\{\theta_{u,3,n} n^{\delta}\}_{n \geqslant 1}$ is a convergent sequence implies that the econometric model is locally misspecified in the sense of Assumption 11. Notice that our choices of δ include a case in which the local misspecification is asymptotically irrelevant relative to sampling error (i.e. $\delta = 1$), one case in which local misspecification is at the exact rate of sampling error (i.e. $\delta = 1/2$), and two cases in which the local misspecification is overwhelming relative to sampling error (i.e. $\delta \in 1/3$). For the sake of completeness of the analysis, we also consider a case in which the econometric model is correctly specified, i.e., $\theta_{u,3,n} = 0$.

By our arguments in Section 2.2, the correctly specified versions of error distribution (Eq. 2.19), state transition probability distribution (Eq. 2.20), and sequence of utility functions (Eq. 2.21), determines the sequence of true CCPs $\{P_n^*\}_{n \geqslant 1}$. We generate marginal observations of the state variables according to the following distribution:

$$m_n^*(x) \propto 1 + \log(x).^{10}$$

By combining the true state transition probability distribution f_n^* (given by f_{θ_f} with $\theta_f = 0.25$), the true CCPs P_n^*, and the true marginal distribution m_n^*, we have completely specified the true joint distribution Π_n^* in Eq. 2.8.

Our results will be the average of $S = 20,000$ independent datasets of observations $\{(a_i, x_i, x_i')\}_{i \leqslant n}$ that are i.i.d. distributed according to Π_n^*. We present simulation results for sample sizes of $n \in \{1,000, 5,000, 10,000\}$. For each $i = 1, \ldots, n$, the observation (a_i, x_i, x_i') is generated according to the following three step sampling

[10] Recall from Section 2.2 that this function is left unspecified in the econometric model.

procedure:

- First, we sample an independent observation of x_i distributed according to m_n^*.
- Second, given the realization of x_i, we sample an independent observation of a_i distributed according to $P_n^*(a|x = x_i)$.
- Third, given the realization of (x_i, a_i), we sample an independent observation of x_i' distributed according to $f_n^*(x'|x = x_i, a = a_i)$.

2.5.2 Estimation

Given any sample of observations $\{(a_i, x_i, x_i')\}_{i \leqslant n}$, the researcher estimates the parameters of interest $\theta = (\theta_{u,1}, \theta_{u,2}, \theta_f)$ using a two-stage K-step policy iteration estimator described in Sections 2.3-2.4.

- In a first stage, the researcher estimates θ_f by using the following sample analogue estimator:

$$\hat{\theta}_{f_n} \equiv \frac{\sum_{i=1}^n 1[a_i = 1, x_i' = x_i, x_i \neq |X|]}{\sum_{i=1}^n 1[a_i = 1, x_i \neq |X|]}. \tag{2.22}$$

- In a second stage, the researcher estimates $(\theta_{u,1}, \theta_{u,2})$ by using the K-step policy-iteration as described in Sections 2.3. The researcher computes the policy iteration mapping Ψ_θ as in Eq. 2.7,[11] and solves the estimation problem in Eq. 2.10 by making the following choices:

[11] Notice that this is the result of combining the incorrectly specified model for the utility function (Eq. 2.18) with the correctly specified error and state transition probability distributions (Eqs. 2.19 and 2.20, respectively).

– The zero-step CCP estimator $\hat{P}_n^0$ is set to be the non-parametric estimator in Eq. 2.11, i.e.,

$$\hat{P}_n^0(a|x) \equiv \frac{\sum_{\tilde{x}' \in X} \hat{\Pi}_n(a, x, \tilde{x}')}{\sum_{(\check{a}, \check{x}') \in AX} \hat{\Pi}_n(\check{a}, x, \check{x}')}.$$

– The number of steps used is $K \in \{1, 2, 3, 10\}$.[12]

– The criterion function Q_n is equal to: (a) pseudo-likelihood function Q_n^{ML} in Section 2.4.2 and (b) the weighted minimum distance function Q_n^{MD} in Section 2.4.3 with three choices of limiting weight matrix W^*: identity matrix (i.e. $W^* = \mathbf{I}_{|AX|}$), asymptotic variance minimizer (i.e. $W^* = W_{AV}^*$ and asymptotic mean square error minimizer (i.e. $W^* = W_{AMSE}^*$).[13]

For each of the 20,000 simulated datasets, the researcher computes sixteen estimators. These are the natural result of having four choices of $K \in \{1, 2, 3, 10\}$ and four choices of criterion functions $Q_n \in \{Q_n^{ML}, Q_n^{MD}(\mathbf{I}), Q_n^{MD}(W_{AV}^*), Q_n^{MD}(W_{AMSE}^*)\}$.

2.5.3 Results

We used our simulations to investigate the finite sample behavior of the estimators for $\theta_u = (\theta_{u,1}, \theta_{u,2})$. For reasons of brevity, we focus on the main text on the coefficient $\theta_{u,2}$ as we expect it to be more affected by the misspecification of the utility function with respect to $x \in X$.

[12] In accordance to our asymptotic theory, the simulation results with $K \in \{4, \ldots, 9\}$ are almost identical to those with $K \in \{3, 10\}$. These were eliminated from the paper for reasons of brevity but they are available from the authors upon request.

[13] These were approximated using Monte Carlo integration and numerical derivates with sample size that are significantly larger than those used in the actual Monte Carlo simulations.

Table 2.1 describes results under correct specification. As expected, when scaled by $\sqrt{n}$, all estimators appear to converge to a distribution with zero mean and finite and positive variance. As predicted by Aguirregabiria and Mira (2002), the number of iterations K does not appear to affect the bias or variance of any of the estimators under consideration. Also, since the estimators are asymptotically unbiased, the optimal MD estimator in terms of variance or mean squared error are numerically identical. In addition, these seem to be very similar to the ML estimator and slightly more efficient than MD estimator with identity weight matrix.

Table 2.2 provides results under local misspecification that is asymptotically irrelevant relative to sampling error, i.e., $\delta = 1$. According to our theoretical results, the asymptotic behavior of all estimators should identical to the correctly specified model. These predictions are confirmed in our simulations, as the results in Tables 2.1 and 2.2 are virtually identical.

Table 2.3 describes results under local misspecification that is of the exact rate of the sampling error, i.e., $\delta = 1/2$. According to our theoretical results, the presence of this local misspecification should result in asymptotically biased estimator. Nevertheless, our results indicate that the number of iterations K should not affect the asymptotic bias or variance of any of the estimators under consideration. Given the presence of (asymptotic) bias and variance, we now consider the mean squared error as the criterion to compare estimators. As expected, the optimal MD estimator in terms of mean squared error appears to be slightly more efficient than any other estimators. Interestingly, in this simulation design, the MD estimator with identity weight matrix is now also slightly more efficient than the ML and optimal MD estimator in terms of variance. In particular, while the MD estimator with iden-

tity matrix has more variance than the ML estimator it appears to have less bias, resulting in less mean squared error.

Tables 2.4 and 2.5 provide results under local misspecification that is asymptotically overwhelming relative to sampling error, i.e., $\delta = 1/3$. According to our theoretical results, the presence of this local misspecification dramatically changes the asymptotic distribution of all estimators under consideration. In particular, these no longer converge at the regular $n^{1/2}$-rate. In fact, at the said rate, the asymptotic bias is no longer asymptotically bounded as Table 2.4 clearly shows. Once we scale the estimators appropriately (at the $n^{1/3}$-rate) all estimators under consideration should converge to an asymptotic distribution that is entirely composed of bias. Furthermore, our theoretical results indicate that the number of iterations K does not affect this asymptotic distribution. These two facts are clearly depicted in Table 2.5. In this case, the optimal MD estimator in terms of mean squared error appears to be significantly more efficient than any other estimators. In line with previous results, the MD estimator with identity weight matrix is slightly more efficient than the ML and optimal MD estimator in terms of variance.

2.6 Conclusion

This paper considers the problem of inference in dynamic discrete choice problems when the structural model is locally misspecified in an arbitrary fashion.

We consider the class of two stage estimators based on the K-step sequential policy iteration algorithm developed by Aguirregabiria and Mira (2002), where K denotes the number of iterations employed in the estimation. By appropriate choice of the criterion function, this class captures the K-step maximum likelihood estimators

(K-ML) and the K-step minimum distance estimators (K-MD). Special cases of our framework are Rust (1987)'s nested fixed point estimator, Hotz and Miller (1993)'s conditional choice probability estimator, Aguirregabiria and Mira (2002)'s nested algorithm estimator, and Pesendorfer and Schmidt-Dengler (2008)'s least squares estimator.

We derive and compare the asymptotic distributions of K-ML and K-MD estimators when the model is arbitrarily locally misspecified and we obtain three main results. In the absence of misspecification, Aguirregabiria and Mira (2002) show that all K-ML estimators are asymptotically equivalent regardless of the choice of K. Our first result shows that this finding extends to a locally misspecified model, regardless of the degree of local misspecification. As a second result, we show that an analogous result holds for all K-MD estimators, i.e., all K-MD estimator are asymptotically equivalent regardless of the choice of K. Our third and final result is to compare K-MD and K-ML estimators in terms of asymptotic mean squared error (AMSE). Under local misspecification, the optimally weighted K-MD estimator depends on the unknown asymptotic bias and is no longer feasible. In turn, feasible K-MD estimators could have an AMSE that is higher or lower than that of the K-ML estimators. To demonstrate the relevance of our asymptotic analysis, we illustrate our findings using in a simulation exercise based on a misspecified version of Rust (1987) bus engine problem.

2.7 Proofs

This appendix collects all proofs of results in the paper and several intermediate results.

2.7.1 Notation

Throughout this appendix, "LLN" refers to the strong law of large numbers, "CLT" refers to the central limit theorem, and "CMT" refers to the continuous mapping theorem. Also, "s.t." abbreviates "such that", and "RHS" and " LHS" abbreviate "right hand side" and "left hand side", respectively.

Throughout this appendix, we employ the following notation. Given a DGP Π_n^*, we define the probability of actions and states J_n^*, CCPs P_n^* and transition probabilities f_n^* in the usual fashion, i.e.,

$$\begin{aligned} J_n^*(a,x) &\equiv \sum_{\tilde{x}' \in X} \Pi_n^*(a,x,\tilde{x}') \quad \forall (a,x) \in AX, \\ P_n^*(a|x) &\equiv \frac{\sum_{\tilde{x}' \in X} \Pi_n^*(a,x,\tilde{x}')}{\sum_{(\check{a},\check{x}') \in AX} \Pi_n^*(\check{a},x,\check{x}')} \quad \forall (a,x) \in AX, \\ f_n^*(x'|a,x) &\equiv \frac{\Pi_n^*(a,x,x')}{\sum_{\tilde{x}' \in X} \Pi_n^*(a,x,\tilde{x}')} \quad \forall (a,x,x') \in AX^2. \end{aligned}$$

Their limiting values are defined analogously, as shown in Assumption 11. The only one that is not defined there is the limiting probability of actions and states J^*, which we define by replacing Π_n^* by Π^* in the previous display, i.e.,

$$J^*(a,x) \equiv \sum_{\tilde{x}' \in X} \Pi^*(a,x,\tilde{x}') \quad \forall (a,x) \in AX.$$

Finally, it convenient to define sample analogues of these objects. For a sample DGP $\hat{\Pi}_n$, we define the sample analogues of actions and states $\hat{J}_n$, sample analogues of CCPs $\hat{P}_n$ and sample analogues of transition probabilities f_n^* by replacing Π_n^* by $\hat{\Pi}_n$,

i.e.

$$\begin{aligned}
\hat{J}_n(a,x) &\equiv \sum_{\tilde{x}' \in X} \hat{\Pi}_n(a,x,\tilde{x}') \quad \forall (a,x) \in AX, \\
\hat{P}_n(a|x) &\equiv \frac{\sum_{\tilde{x}' \in X} \hat{\Pi}_n(a,x,\tilde{x}')}{\sum_{(\check{a},\check{x}') \in AX} \hat{\Pi}_n(\check{a},x,\check{x}')} \quad \forall (a,x) \in AX, \\
\hat{f}_n(x'|a,x) &\equiv \frac{\hat{\Pi}_n(a,x,x')}{\sum_{\tilde{x}' \in X} \hat{\Pi}_n(a,x,\tilde{x}')} \quad \forall (a,x,x') \in AX^2.
\end{aligned}$$

In the current setup, these sample analogue estimators have regular asymptotic properties as we show in Lemmas 27 and 28.

Finally, we now define a matrix $\Sigma \in \mathbb{R}^{|AX| \times |\tilde{A}X|}$ that appears repeatedly in our formal arguments. This is a block diagonal matrix:

$$\Sigma = \begin{bmatrix} \Sigma_1 & \mathbf{0}_{|A\tilde{A}|} & \cdots & \mathbf{0}_{|A\tilde{A}|} \\ \mathbf{0}_{|A\tilde{A}|} & \Sigma_2 & \mathbf{0}_{|A\tilde{A}|} & \mathbf{0}_{|A\tilde{A}|} \\ \vdots & \mathbf{0}_{|A\tilde{A}|} & \ddots & \mathbf{0}_{|A\tilde{A}|} \\ \mathbf{0}_{|A\tilde{A}|} & \mathbf{0}_{|A\tilde{A}|} & \mathbf{0}_{|A\tilde{A}|} & \Sigma_{|X|} \end{bmatrix}, \tag{2.23}$$

where $\Sigma_x \equiv \frac{1}{J^*(x)}[\mathbf{I}_{|\tilde{A}| \times A} - \{P^*(\tilde{a},x)\}_{\tilde{a} \in \tilde{A}} \mathbf{1}_{1 \times A}]$ for every $x \in X$.

2.7.2 Results on the econometric model

Proof of Lemma 20. The econometric model imposes all assumptions in Aguirregabiria and Mira (2002, Sections 2-3). Thus, Parts (a)-(b) follow from Aguirregabiria and Mira (2002, Proposition 1), Pard (d) follows from Aguirregabiria and Mira (2002, Proposition 2), and Parts (c) and (e) are a corollary of the discussion in Aguirregabiria and Mira (2002, Page 1532). □

Proof of Lemma 21. Part (a) follows from $P_\theta = \Psi_\theta(P_\theta)$ and that $\Psi_\theta(P)(a|x) > 0$ for any $(a, x) \in AX$ and any θ and P. Part (b) follows from Rust (1988, Pages 1015-6). Part (c) follows from $P_\theta = \Psi_\theta(P_\theta)$ and $\partial\Psi_\theta(P)/\partial P = \mathbf{0}$ at $P = P_\theta$. Part (d) follows from the following argument. Suppose that $\exists\theta_f \in \Theta_f$ and $\exists\alpha_a, \alpha_b \in \Theta_\alpha$ s.t. $P_{(\alpha_a,\theta_f)} = P_{(\alpha_b,\theta_f)}$. Then, $P_\theta = \Psi_\theta(P_\theta)$ implies that $\Psi_{(\alpha_a,\theta_f)}(P) = P$ and $\Psi_{(\alpha_b,\theta_f)}(P)) = P$ for $P = P_{(\alpha_a,\theta_f)} = P_{(\alpha_b,\theta_f)}$. In turn, since this condition identifies α, we conclude that $\alpha_a = \alpha_b$. □

2.7.3 Results on local misspecification

Proof of Theorem 22. Since $\Theta = \Theta_\alpha \times \Theta_f$ is compact and $||(P_{(\alpha,\theta_f)} - P_n^*), (f_{\theta_f} - f_n^*)||$ is a continuous function of (α, θ_f), the arguments in (Royden, 1988, pages 193-195) implies that $\exists(\alpha^*, \theta_f^*) \in \Theta$ that minimizes $||(P_{(\alpha,\theta_f)} - P_n^*), (f_{\theta_f} - f_n^*)||$. By Assumption 11(b), this minimum value is zero, i.e., $\exists(\alpha^*, \theta_f^*) \in \Theta$ s.t. $||(P_{(\alpha^*,\theta_f^*)} - P^*), (f_{\theta_f^*} - f^*)|| = 0$ or, equivalently, $P_{(\alpha^*,\theta_f^*)} = P^*$ and $f_{\theta_f^*} = f^*$.

Now suppose that this also occurs for $(\tilde{\theta}_f, \tilde{\alpha}) \in \Theta$. We now show that $(\theta_f^*, \alpha^*) = (\tilde{\theta}_f, \tilde{\alpha})$. By triangle inequality $||f_{\theta_f^*} - f_{\tilde{\theta}_f}|| \leqslant ||f_{\theta_f^*} - f^*|| + ||f_{\tilde{\theta}_f} - f^*\|$ and since θ_f^* and $\tilde{\theta}_f$ both satisfy $||f_{\theta_f} - f^*|| = 0$, we conclude that $||f_{\theta_f^*} - f_{\tilde{\theta}_f}|| = 0$ and so $f_{\theta_f^*} = f_{\tilde{\theta}_f}$. By Assumption 10, this implies that $\theta_f^* = \tilde{\theta}_f$. Using this and by repeating the previous argument with $P_{(\alpha,\theta_f^*)}$ instead of f_{θ_f}, we conclude that $\alpha^* = \tilde{\alpha}$. □

2.7.4 Results on inference

Proof of Theorem 23. Throughout this proof, let $\mathcal{N}_\alpha$ denote an arbitrarily neighborhood of α^* that results from projecting $\mathcal{N}$ onto its α-coordinate.

Part 1. Fix $K \geqslant 1$ arbitrarily. We prove the result by assuming that:

$$n^{\min\{\delta,1/2\}}(\hat{P}_n^{K-1} - P^*) = O_{P_n}(1). \tag{2.24}$$

By definition, $\hat{\alpha}_n^K = \arg\max_{\alpha\in\Theta_\alpha} Q_n(\alpha)$ with $Q_n(\alpha) \equiv Q_n(\alpha, \hat{\theta}_{f,n}, \hat{P}_n^{K-1})$ and $\hat{P}_n^{K-1} = \Psi_{(\hat{\alpha}_n^{K-1}, \hat{\theta}_{f,n})}(\hat{P}_n^{K-2})$ for $K > 1$ and $\hat{P}_n^{K-1} = \hat{P}_n^0$ for $K = 1$. The result is a direct consequence of Theorem 34, provided that verify its conditions.

The compactness of Θ and Assumption 13 (items 1-3) imply that $\hat{\alpha}_n^K = \alpha^* + o_{p_n}(1)$ follows from Theorem 33. In turn, $\hat{\alpha}_n^K = \alpha^* + o_{p_n}(1)$ is condition (i) in Theorem 34. Assumption 13 (4) and (5) are conditions (ii) and (iii) in Theorem 34, respectively.

Eq. 2.24 implies that $(\alpha^*, \hat{\theta}_{f,n}, \hat{P}_n) \in \mathcal{N}$ w.p.a.1. This and Assumption 13 (3,5) imply that the following derivation holds w.p.a.1.

$$\begin{aligned}
& n^{\min\{\delta,1/2\}} \frac{\partial Q_n(\alpha^*)}{\partial \alpha} \\
&= n^{\min\{\delta,1/2\}} \frac{\partial Q_n(\alpha^*, \hat{\theta}_{f,n}, \hat{P}_n^{K-1})}{\partial \alpha} \\
&= n^{\min\{\delta,1/2\}} \frac{\partial Q_n(\alpha^*, \theta_f^*, P^*)}{\partial \alpha} \\
&\quad + \frac{\partial Q_n(\alpha^*, \tilde{\theta}_{f,n}, \tilde{P}_n)}{\partial \alpha \partial \theta_f'} n^{\min\{\delta,1/2\}}(\hat{\theta}_{f,n} - \theta_f^*) \\
&\quad + \frac{\partial Q_n(\alpha^*, \tilde{\theta}_{f,n}, \tilde{P}_n)}{\partial \alpha \partial P'} n^{\min\{\delta,1/2\}}(\hat{P}_n^{K-1} - P^*) \\
&= n^{\min\{\delta,1/2\}} \Big[\frac{\partial Q_n(\alpha^*, \theta_f^*, P^*)}{\partial \alpha} \\
&\quad + \frac{\partial Q_\infty(\alpha^*, \theta_f^*, P^*)}{\partial \alpha \partial \theta_f'} (\hat{\theta}_{f,n} - \theta_f^*) \Big] + o_{P_n}(1),
\end{aligned}$$

where $(\tilde{\theta}_{f,n}, \tilde{P}_n)$ is some sequence between $(\hat{\theta}_{f,n}, \hat{P}_n^{K-1})$ and (θ_f^*, P^*). From this and Eq. 2.24,

$$n^{\min\{\delta,1/2\}} \frac{\partial Q_n(\alpha^*)}{\partial \alpha} \xrightarrow{d} \zeta_1 + \frac{\partial Q_\infty(\alpha^*, \theta_f^*, P^*)}{\partial \alpha \partial \theta_f'} \zeta_2,$$

which verifies condition (iv) in Theorem 34.

By previous arguments, $(\alpha, \hat{\theta}_{f,n}, \hat{P}_n) \in \mathcal{N}$ w.p.a.1 for all $\alpha \in \mathcal{N}_\alpha$. This and Assumption 13 (3,5,6) imply that the following derivation holds w.p.a.1.

$$\begin{aligned}
\frac{\partial Q_n(\alpha)}{\partial \alpha \partial \alpha'} &= \frac{\partial Q_n(\alpha, \hat{\theta}_{f,n}, \hat{P}_n^{K-1})}{\partial \alpha \partial \alpha'} \\
&= \frac{\partial Q_\infty(\alpha, \hat{\theta}_{f_n}, \hat{P}_n^{K-1})}{\partial \alpha \partial \alpha'} + o_{P_n}(1) \\
&= \frac{\partial Q_\infty(\alpha, \theta_f^*, P^*)}{\partial \alpha \partial \alpha'} + o_{P_n}(1),
\end{aligned}$$

where convergence is uniformly in $\alpha \in \mathcal{N}_\alpha$. This verifies condition (v) in Theorem 34. In turn, this and Assumption 13 (6) imply condition (vi) in Theorem 34. Since we have verified all condition in Theorem 34, we conclude that:

$$n^{\min\{\delta,1/2\}}(\hat{\alpha}_n^K - \alpha^*) = A_1 n^{\min\{\delta,1/2\}} \frac{\partial Q_n(\alpha^*, \theta_f^*, P^*)}{\partial \alpha'} + A_2 n^{\min\{\delta,1/2\}}(\hat{\theta}_{f,n} - \theta_f^*) + o_{P_n}(1), \tag{2.25}$$

with

$$A_1 \equiv \left(\frac{\partial Q_\infty(\alpha, \theta_f^*, P^*)}{\partial \alpha \partial \alpha'}\right)^{-1}$$

$$A_2 \equiv \left(\frac{\partial Q_\infty(\alpha, \theta_f^*, P^*)}{\partial \alpha \partial \alpha'}\right)^{-1} \frac{\partial Q_\infty(\alpha^*, \theta_f^*, P^*)}{\partial \alpha \partial \theta_f'}.$$

Part 2. The objective of this part is to show Eq. 2.24 holds for all $K \geqslant 1$. We prove the result by induction.

We begin with the initial step. For $K = 1$, the result holds by Assumption 14. In addition, part 1 implies that Eq. 2.25 with $K = 1$, i.e.,

$$n^{\min\{\delta,1/2\}}(\hat{\alpha}_n^1 - \alpha^*) = A_1 n^{\min\{\delta,1/2\}} \frac{\partial Q_n(\alpha^*, \theta_f^*, P^*)}{\partial \alpha'} + A_2 n^{\min\{\delta,1/2\}}(\hat{\theta}_{f,n} - \theta_f^*) + o_{P_n}(1),$$

This concludes the initial step.

We now verify the inductive step. Suppose that for some $K \geqslant 1$, Eqs. 2.24-2.25 hold. Based on this, we show that Eqs. 2.24-2.25 hold with K replaced by $K + 1$. Consider the following argument. By inductive assumption, $(\hat{\alpha}_n^K, \hat{\theta}_{f,n}, \hat{P}_n^{K-1}) = (\alpha^*, \theta_f^*, P^*) + o_{P_n}(1)$ and so $(\hat{\alpha}_n^K, \hat{\theta}_{f,n}, \hat{P}_n^{K-1}) \in \mathcal{N}$ w.p.a.1. Then, Assumption 13 implies that

$$\begin{aligned}
&n^{\min\{\delta,1/2\}}(\hat{P}_n^K - P^*) \\
&\quad = n^{\min\{\delta,1/2\}}(\Psi_{(\hat{\alpha}_n^K, \hat{\theta}_{f,n})}(\hat{P}_n^{K-1}) - \Psi_{(\alpha^*, \theta_f^*)}(P^*)) \\
&\quad = n^{\min\{\delta,1/2\}}\Big[\frac{\partial \Psi_{(\hat{\alpha}_n^K, \hat{\theta}_{f,n})}(\hat{P}_n^{K-1})}{\partial \alpha'}(\hat{\alpha}_n^K - \alpha) \\
&\qquad + \frac{\partial \Psi_{(\hat{\alpha}_n^K, \hat{\theta}_{f,n})}(\hat{P}_n^{K-1})}{\partial \theta_f'}(\hat{\theta}_{f,n} - \theta^*) \\
&\qquad + \frac{\partial \Psi_{(\tilde{\alpha}_n^K, \tilde{\theta}_{f,n})}(\hat{P}_n^{K-1})}{\partial P'}(\hat{P}_n^{K-1} - P^*)\Big] + o_{P_n}(1) \\
&\quad = B_1 n^{\min\{\delta,1/2\}}(\hat{\alpha}_n^K - \alpha) + B_2 n^{\min\{\delta,1/2\}}(\hat{\theta}_{f,n} - \theta^*) + o_{P_n}(1),
\end{aligned}$$

where the first line uses that $P^* = \Psi_{(\alpha^*, \theta_f^*)}(P^*)$, the second equality holds for $(\hat{\alpha}_n^K, \hat{\theta}_{f_n}, \hat{P}_n^{K-1}) \in \mathcal{N}$, which holds w.p.a.1, and the final line holds by plugging in

Eq. 2.25 and by defining:

$$B_1 \equiv \frac{\partial \Psi_{(\alpha^*,\theta_f^*)}(P^*)}{\partial \alpha'}$$

$$B_2 \equiv \frac{\partial \Psi_{(\alpha^*,\theta_f^*)}(P^*)}{\partial \theta_f'}.$$

From this, we conclude that $n^{\min\{\delta,1/2\}}(\hat{P}_n^K - P^*) = O_{P_n}(1)$, i.e., Eq. 2.24 holds with K replaced by $K+1$. In turn, this and part 1 then implies that Eq. 2.25 holds with K replaced by $K+1$. This concludes the inductive step and the proof. □

2.7.5 Proofs of theorems

Proof of Theorem 24. This result is a corollary of Theorem 23 and Lemma 32. To apply Theorem 23, we need to verify Assumptions 13-14. We anticipate that

$$Q_\infty^{ML}(\theta, P) = \sum_{(a,x)\in AX} J^*(a,x) \ln \Psi_\theta(P)(a,x).$$

We first verify the conditions in Assumption 13.

Condition (a). First, notice that $\hat{J}_n - J^* = o_{P_n}(1)$ and $\Psi_\theta(P)(a,x) > 0$ for all $(\theta, P) \in \Theta \times \Theta_P$ implies that $Q_n^{ML}(\theta, P) - Q_\infty^{ML}(\theta, P) = o_{P_n}(1)$. Furthermore, notice that

$$\begin{aligned}
&\sup_{(\theta,P)\in\Theta\times\Theta_P} |Q_n^{ML}(\theta,P) - Q_\infty^{ML}(\theta,P)| \\
&= \sup_{(\theta,P)\in\Theta\times\Theta_P} |\sum_{(a,x)\in AX} (\hat{J}_n(a,x) - J^*(a,x)) \ln \Psi_\theta(P)(a,x)| \\
&\leqslant |\sum_{(a,x)\in AX} (\hat{J}_n(a,x) - J^*(a,x))| \ln[\min_{(a,x)\in AX} \inf_{(\theta,P)\in\Theta\times\Theta_P} \Psi_\theta(P)(a,x)]
\end{aligned}$$

Since $\Psi_\theta(P)(a,x) > 0$ for all $(\theta,P) \in \Theta \times \Theta_P$ and all (a,x), $\Psi_\theta(P)(a,x) : \Theta \times \Theta_P \to \mathbb{R}$ is continuous in (θ,P) for all (a,x), and $\Theta \times \Theta_P$ is compact, this implies that $\min_{(a,x)\in AX} \inf_{(\theta,P)\in\Theta\times\Theta_P} \Psi_\theta(P)(a,x) > 0$. From this and $\hat{J}_n - J^* = o_{P_n}(1)$, we conclude that $\sup_{(\theta,P)\in\Theta\times\Theta_P} |Q_n^{ML}(\theta,P) - Q_\infty^{ML}(\theta,P)| = o_{P_n}(1)$. Second, since $\Psi_\theta(P)(a,x) > 0$ for all $(\theta,P) \in \Theta \times \Theta_P$ and since $\Psi_\theta(P) : \Theta \times \Theta_P \to \mathbb{R}$ is continuous in (θ,P), it follows that $Q_\infty^{ML}(\theta,P) : \Theta \times \Theta_P \to \mathbb{R}$ is continuous in (θ,P). In turn, since $\Theta \times \Theta_P$ is compact it follows that $Q_\infty^{ML}(\theta,P) : \Theta \times \Theta_P \to \mathbb{R}$ is uniformly continuous in (θ,P). Third, $(\hat{\theta}_{f,n}, \hat{P}_n) - (\theta_f^*, P^*) = o_{P_n}(1)$. By combining these with Gourieroux and Monfort (1995, Lemma 24.1), the result follows.

Condition (b). This result is a consequence of the information inequality (see, e.g., (White, 1996, Theorem 2.3)), which we now review for completeness. Notice that:

$$\begin{aligned} &Q_\infty^{ML}(\alpha,\theta_f^*,P^*) - Q_\infty^{ML}(\alpha^*,\theta_f^*,P^*) \\ &\quad = E_{J^*}[\ln \frac{\Psi_{(\alpha,\theta_f^*)}(P^*)(a,x)}{\Psi_{(\alpha^*,\theta_f^*)}(P^*)(a,x)}] = E_{J^*}[\ln \frac{\Psi_{(\alpha,\theta_f^*)}(P^*)(a,x)}{P^*(a|x)}], \end{aligned}$$

where the last equality uses that $\Psi_{(\alpha^*,\theta_f^*)}(P^*) = P^*$. Clearly, $Q_\infty^{ML}(\alpha,\theta_f^*,P^*) - Q_\infty^{ML}(\alpha^*,\theta_f^*,P^*) = 0$ for $\alpha = \alpha^*$. Consider any $\tilde{\alpha} \in \Theta_\alpha \backslash \alpha^*$. By the identification assumption, $\Psi_{(\tilde{\alpha},\theta_f^*)}(P^*) \neq \Psi_{(\alpha^*,\theta_f^*)}(P^*) = P^*$ and so

$$\Psi_{(\alpha,\theta_f^*)}(P^*)(a,x)/\Psi_{(\alpha^*,\theta_f^*)}(P^*)(a,x) \neq 1$$

for some $(a,x) \in A \times X$, implying that $\{\ln(\Psi_{(\tilde{\alpha},\theta_f^*)}(P^*)(a,x)/P^*(x,a))\}_{(x,a)\in AX} \neq \mathbf{0}_{AX}$.

By Jensen's inequality, it follows that

$$\begin{aligned}
&Q_\infty^{ML}(\tilde{\alpha}, \theta_f^*, P^*) - Q_\infty^{ML}(\alpha^*, \theta_f^*, P^*) \\
&\leqslant \ln E_{J^*}[\frac{\Psi_{(\alpha,\theta_f^*)}(P^*)(a,x)}{P^*(a|x)}] \\
&= \ln \sum_{(x,a)\in X\times A} m^*(x)\Psi_{(\alpha,\theta_f^*)}(P^*)(a,x) \\
&= \ln 1 \\
&= 0,
\end{aligned}$$

where the first equality uses that $J^*(a,x) = P^*(a|x)m^*(x)$ and the second equality uses that $\sum_{a\in A}\Psi_{(\tilde{\alpha},\theta_f^*)}(P^*)(a,x) = 1$ for all $x \in X$ and $\sum_{x\in X} m^*(x) = 1$.

Condition (c). Since $\Psi_\theta(P)(a,x) > 0$ for all $(\theta, P) \in \Theta \times \Theta_P$ and all (a,x), and $\Psi_\theta(P)(a,x) : \Theta \times \Theta_P \to \mathbb{R}$ is twice continuously differentiable in (θ, P) for all (a,x), we conclude that $\ln \Psi_\theta(P)(a,x) : \Theta \times \Theta_P \to \mathbb{R}$ is twice continuously differentiable in (θ, P) for all (a,x). From here, the result follows.

Condition (d). In the verification of condition (c), we have shown that the function $Q_n^{ML}(\theta, P) : \Theta \times \Theta_P \to \mathbb{R}$ is twice continuously differentiable in $(\theta, P) \in \Theta \times \Theta_P$. By the same argument, $Q_\infty^{ML}(\theta, P) : \Theta \times \Theta_P \to \mathbb{R}$ is also twice continuously differentiable in $(\theta, P) \in \Theta \times \Theta_P$. Then, by direct computation:

$$\begin{aligned}
&\sup_{(\theta,P)\in\mathcal{N}} |\frac{\partial Q_n^{ML}(\alpha,\theta_f,P)}{\partial\lambda'\partial\alpha} - \frac{\partial Q_\infty^{ML}(\alpha,\theta_f,P)}{\partial\lambda'\partial\alpha}| \\
&= \sup_{(\theta,P)\in\mathcal{N}} |\sum_{(a,x)\in AX}(J_n^*(a,x) - J^*(a,x))M_{\theta,P}(a,x)| \\
&\leqslant |\sum_{(a,x)\in AX}(J_n^*(a,x) - J^*(a,x))| \max_{(a,x)\in AX}\sup_{(\theta,P)\in\Theta\times\Theta_P}|M_{\theta,P}(a,x)|,
\end{aligned}$$

with $M_{\theta,P}(a,x)$ defined by:

$$\frac{-1}{(\Psi_\theta(P)(a,x))^2}\frac{\partial\Psi_\theta(P)(a,x)}{\partial\lambda'}\frac{\partial\Psi_\theta(P)(a,x)}{\partial\alpha}+\frac{1}{\Psi_\theta(P)(a,x)}\frac{\partial\Psi_\theta(P)(a,x)}{\partial\lambda'\partial\alpha}.$$

Since $\Psi_\theta(P)(a,x) > 0$ for all $(\theta,P) \in \Theta \times \Theta_P$ and all (a,x), $\Psi_\theta(P)(a,x) : \Theta \times \Theta_P \to \mathbb{R}$ is continuous in (θ,P) for all (a,x), and $\Theta \times \Theta_P$ is compact, this implies that $\inf_{(\theta,P)\in\Theta\times\Theta_P} \Psi_\theta(P)(a,x) > 0$ for all (a,x). From this and that $\Psi_\theta(P)$ is twice continuously differentiable in (θ,P) for all (a,x), we conclude that $M_{\theta,P}(a,x)$ is continuous in (θ,P) for all (a,x). Since $\Theta \times \Theta_P$ is compact,

$$\max_{(a,x)\in AX}\sup_{(\theta,P)\in\Theta\times\Theta_P}|M_{\theta,P}(a,x)| < \infty.$$

From this and $\hat{J}_n - J^* = o_{P_n}(1)$, the result follows.

Condition (e). By direct computation, for any $\lambda \in (\alpha,\theta_f,P)$,

$$\begin{aligned}\frac{\partial Q_\infty^{ML}(\alpha,\theta_f,P)}{\partial\lambda\partial\alpha'} &= \sum_{(a,x)\in AX} J^*(a,x)\{\frac{-1}{(\Psi_\theta(P)(a,x))^2}\frac{\partial\Psi_\theta(P)(a,x)}{\partial\lambda}\frac{\partial\Psi_\theta(P)(a,x)}{\partial\alpha'}\\ &\quad+\frac{1}{\Psi_\theta(P)(a,x)}\frac{\partial\Psi_\theta(P)(a,x)}{\partial\lambda\partial\alpha'}\}. \qquad (2.26)\end{aligned}$$

This function is continuous and, when evaluated at $(\alpha,\theta_f,P) = (\alpha^*,\theta_f^*,P^*)$, we

obtain:

$$
\begin{aligned}
&\frac{\partial Q_{\infty}^{ML}(\alpha^*, \theta_f^*, P^*)}{\partial\lambda\partial\alpha'} \\
&= \sum_{(a,x)\in AX} J^*(a,x)\{\frac{-1}{(\Psi_{\theta^*}(P^*)(a,x))^2}\frac{\partial\Psi_{\theta^*}(P^*)(a,x)}{\partial\lambda}\frac{\partial\Psi_{\theta^*}(P^*)(a,x)}{\partial\alpha'} \\
&\quad + \frac{1}{\Psi_{\theta^*}(P^*)(a,x)}\frac{\partial\Psi_{\theta}(P^*)(a,x)}{\partial\lambda\partial\alpha'}\} \\
&= \sum_{(a,x)\in AX} \pi^*(x)\{\frac{-1}{P_{\theta^*}(a|x)}\frac{\partial P_{\theta^*}(a|x)}{\partial\lambda}\frac{\partial P_{\theta^*}(a|x)}{\partial\alpha'} \\
&\quad + \frac{\partial P_{\theta^*}(a|x)}{\partial\lambda\partial\alpha'}\} \\
&= -E_{J^*}[\frac{\partial\{\ln P_{\theta^*}(a|x)\}_{(a,x)\in AX}}{\partial\lambda}\frac{\partial\{\ln P_{\theta^*}(a|x)\}_{(a,x)\in AX}}{\partial\alpha}'] \\
&= -\frac{\partial P_{\theta^*}}{\partial\lambda}(\Sigma\Omega_{JJ}\Sigma')^{-1}\frac{\partial P_{\theta^*}}{\partial\alpha}'
\end{aligned}
$$

where the second line uses $J^*(a,x) = P^*(a|x)m^*(x)$, $\partial\Psi_{\theta^*}(P^*)/\partial\alpha = \partial P_{\theta^*}/\partial\alpha$, and $P_{\theta^*} = P^*$, and the second line interchanges summation and differentiation and uses that $\sum_{(a,x)\in AX}\pi^*(x)P_{\theta^*}(a|x) = 1$, and the final equality in the third line uses Lemma 32. To verify the result, it suffices to consider the last expression with $\lambda = \alpha$. Since $(\Sigma\Omega_{JJ}\Sigma')^{-1}$ is a non-singular matrix and $\partial P_{\theta^*}/\partial\alpha$ is a full rank matrix, we conclude that the expression is square, symmetric, and negative definite, and, consequently, it must be non-singular.

Condition (f). If we focus Eq. 2.26 on $\lambda = P$ and evaluate at $(\alpha, \theta_f, P) =$

$(\alpha^*, \theta_f^*, P^*)$,

$$\begin{aligned}
&\frac{\partial Q_\infty^{ML}(\alpha^*, \theta_f^*, P^*)}{\partial P \partial \alpha'} \\
&= \sum_{(a,x)\in AX} J^*(a,x)\{\frac{-1}{(\Psi_{\theta*}(P^*)(a,x))^2}\frac{\partial \Psi_{\theta*}(P^*)(a,x)}{\partial P}\frac{\partial \Psi_{\theta*}(P^*)(a,x)}{\partial \alpha'} \\
&\quad + \frac{1}{\Psi_\theta(P)(a,x)}\frac{\partial \Psi_{\theta*}(P^*)(a,x)}{\partial P \partial \alpha'}\} \\
&= \sum_{(a,x)\in AX} J^*(a,x)\{\frac{-1}{(\Psi_{\theta*}(P^*)(a,x))^2}\frac{\partial \Psi_{\theta*}(P_{\theta*})(a,x)}{\partial P}\frac{\partial \Psi_{\theta*}(P^*)(a,x)}{\partial \alpha'} \\
&\quad + \frac{1}{\Psi_\theta(P)(a,x)}\frac{\partial}{\partial \alpha'}(\frac{\partial \Psi_{\theta*}(P_{\theta*})(a,x)}{\partial P})\}.
\end{aligned}$$

where the second line uses $P_{\theta*} = P^*$ and Young's theorem. Since the Jacobian matrix of $\Psi_{\theta*}$ with respect to P is zero at $P_{\theta*}$, the result follows.

We now verify the Assumption 14. Assumption 14 (b) holds by Lemma 29. To verify Assumption 14(a), consider the following argument. By the verification of Assumption 13(c)-(d), Q_n^{ML} and Q_∞^{ML} are twice continuously differentiable in $(\theta, P) \in \mathcal{N}$. By direct computation,

$$\begin{aligned}
\frac{\partial Q_n^{ML}(\alpha^*, \theta_f^*, P^*)}{\partial \alpha} &= \sum_{(a,x)\in AX} \hat{J}_n(a,x)\frac{1}{\Psi_{\theta*}(P^*)(a,x)}\frac{\partial \Psi_{\theta*}(P^*)(a,x)}{\partial \alpha} \\
&= \frac{\partial\{\ln P_{\theta*}(a|x)\}_{(a,x)\in AX}}{\partial \alpha}\hat{J}_n = \frac{\partial P_{\theta*}}{\partial \lambda}(\Sigma \Omega_{JJ} \Sigma')^{-1}\Sigma \hat{J}_n,
\end{aligned}$$

where the last equality uses that $\partial \Psi_{\theta*}(P^*)/\partial \alpha = \partial P_{\theta*}/\partial \alpha$. Also, by using an analogous argument but applied to the population,

$$\frac{\partial Q_\infty^{ML}(\alpha^*, \theta_f^*, P^*)}{\partial \alpha} = \frac{\partial P_{\theta*}}{\partial \lambda}(\Sigma \Omega_{JJ} \Sigma')^{-1}\Sigma J^* = \mathbf{0},$$

where the equality to zero holds by Assumptions 12 and 13(b). By combining both equations and using Lemma 32, we conclude that:

$$n^{\min\{\delta,1/2\}} \left[\begin{array}{c} \partial Q_n^{ML}(\alpha^*, \theta_f^*, P^*)/\partial\alpha \\ (\hat{\theta}_{f,n} - \theta_f^*) \end{array} \right]$$

$$= \left[\begin{array}{cc} \frac{\partial P_{\theta^*}}{\partial\lambda}(\Sigma\Omega_{JJ}\Sigma')^{-1}\Sigma & \mathbf{0}_{|AX|\times d_f} \\ \mathbf{0}_{d_f\times|AX|} & \mathbf{I}_{d_f} \end{array} \right] n^{\min\{\delta,1/2\}} \left(\begin{array}{c} \hat{J}_n - J^* \\ \hat{\theta}_{f,n} - \theta_f^* \end{array} \right).$$

From this and Lemma 27, we conclude that the desired result holds with (ζ_1, ζ_2) being distributed according the normal distribution with mean

$$\left(\begin{array}{c} \frac{\partial P_{\theta^*}}{\partial\lambda}(\Sigma\Omega_{JJ}\Sigma')^{-1}\Sigma B_J \\ B_f \end{array} \right)$$

and variance

$$\left(\begin{array}{cc} \frac{\partial P_{\theta^*}}{\partial\lambda}(\Sigma\Omega_{JJ}\Sigma')^{-1}\Sigma\Omega_{JJ}\Sigma'(\Sigma\Omega_{JJ}\Sigma')^{-1}\frac{\partial P_{\theta^*}}{\partial\lambda}' & \frac{\partial P_{\theta^*}}{\partial\lambda}(\Sigma\Omega_{JJ}\Sigma')^{-1}\Sigma\Omega_{Jf} \\ \Omega_{Jf}'\Sigma'(\Sigma\Omega_{JJ}\Sigma')^{-1}\frac{\partial P_{\theta^*}}{\partial\lambda}' & \Omega_{ff} \end{array} \right). \quad (2.27)$$

This completes the verification of Assumptions 13-14 and so Theorem 23 applies. The specific formula for the asymptotic distribution relies on the expressions in Eqs. 2.27-2.27 and Lemma 32. □

Proof of Theorem 25. This result is a corollary of Theorem 23. To complete the proof, we need to verify Assumptions 13-14. We anticipate that $Q_\infty^{MD}(\theta, P) = -[P^* - \Psi_\theta(P)]'W^*[P^* - \Psi_\theta(P)]$. We first verify the conditions in Assumption 13.

Condition (a). First, we show that $\sup_{(\theta,P)\in\Theta\times\Theta_P} |Q_n^{MD}(\theta, P) - Q_\infty^{MD}(\theta, P)| =$

$o_{p_n}(1)$. Consider the following argument:

$$\sup_{(\theta,P)\in\Theta\times\Theta_P} |Q_n^{MD}(\theta,P) - Q_\infty^{MD}(\theta,P)|$$

$$= \sup_{(\theta,P)\in\Theta\times\Theta_P} \left| \begin{array}{c} -(\hat{P}_n - P^*)'\hat{W}_n[\hat{P}_n - \Psi_\theta(P)] \\ -(P^* - \Psi_\theta(P))'[\hat{W}_n - W^*][\hat{P}_n - \Psi_\theta(P)] \\ -(P^* - \Psi_\theta(P))'W^*(\hat{P}_n - P^*) \end{array} \right|$$

$$\leqslant \|\hat{P}_n - P^*\|[\|\hat{W}_n - W^*\| + 2\|W^*\|] + \|\hat{W}_n - W^*\|$$

Second, since $\Psi_\theta(P)(a,x) : \Theta \times \Theta_P \to \mathbb{R}$ is continuous in (θ,P) for all (a,x), it follows that $Q_\infty^{MD}(\theta,P) : \Theta \times \Theta_P \to \mathbb{R}$ is continuous in (θ,P). In turn, since $\Theta \times \Theta_P$ is compact it follows that $Q_\infty^{MD}(\theta,P) : \Theta \times \Theta_P \to \mathbb{R}$ is uniformly continuous in (θ,P). Third, $(\hat{\theta}_{f,n}, \hat{P}_n) - (\theta_f^*, P^*) = o_{P_n}(1)$. By combining these with Gourieroux and Monfort (1995, Lemma 24.1), the result follows.

Condition (b). $Q_\infty^{MD}(\alpha, \theta_f^*, P^*) = -[P^* - \Psi_{(\alpha,\theta_f^*)}(P^*)]'W^*[P^* - \Psi_{(\alpha,\theta_f^*)}(P^*)]$ is uniquely maximized at α^*. First, notice that $\Psi_{(\alpha^*,\theta_f^*)}(P^*) = P^*$ and so

$$Q_\infty^{MD}(\alpha^*, \theta_f^*, P^*) = 0.$$

Second, consider any $\tilde{\alpha} \in \Theta_\alpha \backslash \alpha^*$. By the identification assumption, $\Psi_{(\tilde{\alpha},\theta_f^*)}(P^*) \neq \Psi_{(\alpha^*,\theta_f^*)}(P^*) = P^*$. Since W^* is positive definite, $Q_\infty^{MD}(\tilde{\alpha}, \theta_f^*, P^*) > 0$.

Condition (c). This result follows from the fact that $\Psi_\theta(P)(a,x) : \Theta \times \Theta_P \to \mathbb{R}$ is twice continuously differentiable in (θ,P) for all (a,x).

Condition (d). By the same argument as in the verification of condition (c), it follows that $Q_\infty^{MD}(\theta,P) : \Theta \times \Theta_P \to \mathbb{R}$ is twice continuously differentiable in (θ,P). Since $\Psi_\theta(P)(a,x)$ is twice continuously differentiable in (θ,P) for all (a,x), we conclude that $\partial\Psi_\theta(P)(a,x)/\partial\lambda$ and $\partial\Psi_\theta(P)(a,x)/(\partial\lambda'\partial\alpha)$ are continuous in (θ,P)

for $\lambda \in \{\theta, P\}$ for all (a, x). From this and the fact that $\Theta \times \Theta_P$ is compact, we conclude that $\max_{(a,x)\in AX} \sup_{(\theta,P)\in\Theta\times\Theta_P} \|\partial\Psi_\theta(P)(a,x)/\partial\lambda\| < \infty$ and

$$\max_{(a,x)\in AX} \sup_{(\theta,P)\in\Theta\times\Theta_P} \|\partial\Psi_\theta(P)(a,x)/\partial\lambda'\partial\alpha\| < \infty.$$

From this, $\hat{P}_n - P^* = o_{P_n}(1)$, and $\hat{W}_n - W^* = o_{P_n}(1)$, the result follows.

Condition (e). By direct computation, for any $\lambda \in (\alpha, \theta_f, P)$,

$$\frac{\partial Q_\infty^{MD}(\alpha, \theta_f, P)}{\partial\lambda\partial\alpha'} = 2[\frac{\partial}{\partial\lambda}\frac{\Psi_\theta(P)}{\partial\alpha}W^*(P^* - \Psi_\theta(P)) - \frac{\partial\Psi_\theta(P)}{\partial\alpha}W^*\frac{\partial\Psi_\theta(P)}{\partial\lambda}'] \quad (2.28)$$

This function is continuous and, when evaluated at $(\alpha, \theta_f, P) = (\alpha^*, \theta_f^*, P^*)$, we obtain:

$$\frac{\partial Q_\infty^{MD}(\alpha^*, \theta_f^*, P^*)}{\partial\lambda\partial\alpha'} = -2\frac{\partial\Psi_{\theta*}(P^*)}{\partial\alpha}W^*\frac{\partial\Psi_{\theta*}(P^*)}{\partial\lambda}' = -2\frac{\partial P_{\theta*}}{\partial\alpha}W^*\frac{\partial P_{\theta*}}{\partial\lambda}'$$

where the first line uses that $P^* = \Psi_{\theta*}(P^*)$ and $\partial\Psi_{\theta*}(P^*)/\partial\alpha = \partial P_{\theta*}/\partial\alpha$. To verify the result, it suffices to consider the last expression with $\lambda = \alpha$. By assumption, this expression is square, symmetric, and negative definite, and, consequently, it must be non-singular.

Condition (f). If we focus Eq. 2.28 on $\lambda = P$ and evaluate at $(\alpha, \theta_f, P) = (\alpha^*, \theta_f^*, P^*)$,

$$\frac{\partial Q_\infty^{MD}(\alpha^*, \theta_f^*, P^*)}{\partial P\partial\alpha'} = -2\frac{\Psi_{\theta*}(P^*)}{\partial\alpha}W^*\frac{\Psi_{\theta*}(P^*)}{\partial P}' = 0.$$

where the we have used that the Jacobian matrix of $\Psi_{\theta*}$ with respect to P is zero at $P_{\theta*} = P^*$, the result follows.

We now verify the Assumption 14. Assumption 14 (b) holds by Lemma 29. To verify Assumption 14(a), consider the following argument. By the verification

of conditions (c) and (d), Q_n^{MD} and Q_∞^{MD} are twice continuously differentiable in $(\theta, P) \in \mathcal{N}$. By direct computation,

$$\frac{\partial Q_n^{MD}(\alpha^*, \theta_f^*, P^*)}{\partial \alpha} = 2\frac{\Psi_{\theta*}(P^*)}{\partial \alpha}\hat{W}_n[\hat{P}_n - \Psi_{\theta*}(P^*)] = 2\frac{\partial P_\theta^*}{\partial \alpha}W^*[\hat{P}_n - P^*] + o_{p_n}(1),$$

where the last equality uses that $\Psi_{\theta*}(P^*) = P^*$, $\partial\Psi_{\theta*}(P^*)/\partial\alpha = \partial P_{\theta*}/\partial\alpha$, $\hat{P}_n - P^* = o_{P_n}(1)$, and $\hat{W}_n - W^* = o_{P_n}(1)$. We conclude that:

$$\begin{aligned} & n^{\min\{\delta,1/2\}} \left[\begin{array}{c} \partial Q_n^{MD}(\alpha^*, \theta_f^*, P^*)/\partial\alpha \\ (\hat{\theta}_{f,n} - \theta_f^*) \end{array} \right] \\ & = \left[\begin{array}{cc} 2\frac{\partial P_\theta^*}{\partial \alpha}W^* & \mathbf{0}_{|AX|\times d_f} \\ \mathbf{0}_{d_f\times|AX|} & \mathbf{I}_{d_f} \end{array} \right] n^{\min\{\delta,1/2\}} \left[\begin{array}{c} (\hat{P}_n - P^*) \\ (\hat{\theta}_{f,n} - \theta_f^*) \end{array} \right] + o_{p_n}(1). \end{aligned}$$

From this and Lemma 28, we conclude that (ζ_1, ζ_2) is distributed according to

$$N\left(\left(\begin{array}{c} 2\frac{\partial P_\theta^*}{\partial \alpha}W^*\Sigma B_J \\ B_f \end{array} \right), \left(\begin{array}{cc} 4\frac{\partial P_\theta^*}{\partial \alpha}W^*\Sigma\Omega_{JJ}\Sigma' W^{*\prime}\frac{\partial P_\theta^{*\prime}}{\partial \alpha} & 2\frac{\partial P_\theta^*}{\partial \alpha}W^*\Sigma\Omega_{Jf} \\ \Omega_{Jf}'\Sigma' W^{*\prime}\frac{\partial P_\theta^{*\prime}}{\partial \alpha} & \Omega_{ff} \end{array} \right)\right). \quad (2.29)$$

This completes the verification of Assumptions 13-14 and so Theorem 23 applies. The specific formula for the asymptotic distribution relies on the expressions in Eqs. 2.28-2.29. □

2.7.6 Proofs of other results

Lemma 27. *Assume Assumptions 11-15. Then,*

$$n^{\min\{\delta,1/2\}} \left(\begin{array}{c} \hat{J}_n - J^* \\ \hat{\theta}_{f,n} - \theta_f^* \end{array} \right) \xrightarrow{d} N(B \times 1[\delta \leqslant 1/2], \Omega \times 1[\delta \geqslant 1/2]), \quad (2.30)$$

where

$$B \equiv \begin{pmatrix} B_J \\ B_f \end{pmatrix} \equiv \Delta B_{\Pi^*}$$

$$\Omega \equiv \begin{pmatrix} \Omega_{JJ} & \Omega_{Jf} \\ \Omega'_{Jf} & \Omega_{ff} \end{pmatrix} \equiv \Delta(diag(\Pi^*) - \Pi^*\Pi^{*\prime})\Delta',$$

with

$$\Delta \equiv \begin{pmatrix} \{\{1[(a,x) = (\tilde{a},\tilde{x})]\}'_{(a,x,x')\in AX^2}\}_{(\tilde{a},\tilde{x})\in AX}, \\ DG_{\theta_f}(\Pi^*) \end{pmatrix},$$

and $DG_{\theta_f}(\Pi^*)$ denotes the gradient of θ_f-component of G in Assumption 16.

Proof. Under Assumptions 11 and 15, the triangular array CLT implies that:

$$\sqrt{n}(\hat{\Pi}_n - \Pi^*_n) \xrightarrow{d} N(0, diag(\Pi^*) - \Pi^*\Pi^{*\prime}).$$

If this is combined with Assumptions 11, we conclude that:

$$n^{\min\{\delta,1/2\}}(\hat{\Pi}_n - \Pi^*) \xrightarrow{d} N(B_{\Pi^*} 1\{\delta \leqslant 1/2\}, (diag(\Pi^*) - \Pi^*\Pi^{*\prime}) 1\{\delta \geqslant 1/2\}). \quad (2.31)$$

Consider the following argument.

$$\begin{aligned} n^{\min\{\delta,1/2\}} \begin{pmatrix} \hat{J}_n - J^* \\ \hat{\theta}_{f,n} - \theta^*_f \end{pmatrix} &= n^{\min\{\delta,1/2\}} \begin{pmatrix} \hat{J}_n - J^* \\ G_{n,\theta_f}(\hat{\Pi}_n) - G_{\theta_f}(\Pi^*) \end{pmatrix} \\ &= n^{\min\{\delta,1/2\}} \begin{pmatrix} \hat{J}_n - J^* \\ G_{\theta_f}(\hat{\Pi}_n) - G_{\theta_f}(\Pi^*) \end{pmatrix} + o_{p_n}(1) \\ &= n^{\min\{\delta,1/2\}} \left(F\left(\hat{\Pi}_n\right) - F\left(\Pi^*\right) \right) + o_{p_n}(1) \end{aligned}$$

where the first two equalities follow from Assumption 16, where G_{n,θ_f} and G_{θ_f} denote the θ_f-component of G_n and G, respectively, and the last equality follows from defining the function $F : \mathbb{R}^{|AX^2|} \to \mathbb{R}^{|AX|+d_{\theta_f}}$ as follows. For coordinates $j = 1, \ldots, |AX|$

where j represents coordinate $(a,x) \in AX$, $F_j(z) \equiv \sum_{\tilde{x}' \in X} z_{(a,x,\tilde{x}')}$, and for coordinates $j = |AX|+1, \ldots, |AX|+d_{\theta_f}$, $F_j(z) \equiv G_{\theta_f,j}(z)$. By definition, $F(\hat{\Pi}_n) \equiv (\hat{J}_n, G_{\theta_f}(\hat{\Pi}_n))$ and $F(\Pi^*) \equiv (J^*, G_{\theta_f}(\Pi^*))$. By Assumptions 11 and Eq. 2.31, $\hat{\Pi}_n$,belongs to any arbitrarily small neighborhood of Π^* w.p.a.1. Eq. 2.30 holds by the delta method provided that F is continuously differentiable for any Π in a neighborhood of Π^* and its gradient at Π^* is equal to Δ, as we now verify.

For a coordinate $j > |AX|$, G_{θ_f} is differentiable at Π^* by Assumption 16 and for $j \leqslant |AX|$ that represents a certain (a,x), the gradient is:

$$\partial F_j(z)/\partial z_{(\tilde{a},\tilde{x},\tilde{x}')} = 1[(a,x) = (\tilde{a},\tilde{x})]$$

which is also differentiable. □

Lemma 28. *Assume Assumptions 11-15. Then,*

$$n^{\min\{\delta,1/2\}} \begin{pmatrix} \hat{P}_n - P^* \\ \hat{\theta}_{f,n} - \theta_f^* \end{pmatrix} \xrightarrow{d} N(\tilde{B} \times 1[\delta \leqslant 1/2], \tilde{\Omega} \times 1[\delta \geqslant 1/2]), \tag{2.32}$$

where

$$\tilde{B} \equiv \begin{pmatrix} \Sigma B_J \\ B_f \end{pmatrix} \equiv \begin{bmatrix} \Sigma & \mathbf{0}_{|X\tilde{A}| \times d_{\theta_f}} \\ \mathbf{0}_{d_{\theta_f} \times |\tilde{A}X|} & \mathbf{I}_{d_{\theta_f} \times d_{\theta_f}} \end{bmatrix} \Delta B_{\Pi^*}$$

$$\begin{aligned} \tilde{\Omega} &\equiv \begin{pmatrix} \Sigma\Omega_{JJ}\Sigma' & \Sigma\Omega_{Jf} \\ \Omega'_{Jf}\Sigma' & \Omega_{ff} \end{pmatrix} \\ &\equiv \begin{bmatrix} \Sigma & \mathbf{0}_{|X\tilde{A}| \times d_{\theta_f}} \\ \mathbf{0}_{d_{\theta_f} \times |\tilde{A}X|} & \mathbf{I}_{d_{\theta_f} \times d_{\theta_f}} \end{bmatrix} \Delta(diag(\Pi^*) - \Pi^*\Pi^{*\prime})\Delta' \begin{bmatrix} \Sigma & \mathbf{0}_{|X\tilde{A}| \times d_{\theta_f}} \\ \mathbf{0}_{d_{\theta_f} \times |\tilde{A}X|} & \mathbf{I}_{d_{\theta_f} \times d_{\theta_f}} \end{bmatrix}', \end{aligned}$$

with

$$\Delta \equiv \begin{pmatrix} \{\{1[(a,x) = (\tilde{a},\tilde{x})]\}'_{(a,x,x') \in AX^2}\}_{(\tilde{a},\tilde{x}) \in AX}, \\ DG_{\theta_f}(\Pi^*) \end{pmatrix},$$

and $DG_{\theta_f}(\Pi^*)$ denotes the gradient of θ_f-component of G in Assumption 16.

Proof. This result is a direct consequence of Lemma 29 and the delta method. $F : \mathbb{R}^{|AX|+d_{\theta_f}} \to \mathbb{R}^{|\tilde{A}X|+d_{\theta_f}}$ as follows. For coordinates $j = 1, \ldots, |\tilde{A}X|$ where j represents coordinate $(a, x) \in \tilde{A}X$, $F_j(z) \equiv z_{(a,x)} / \sum_{a \in A} z_{(\tilde{a},x)}$, and for $j = |\tilde{A}X| + 1, \ldots, |\tilde{A}X| + d_{\theta_f}$, $F_j(z) = z_j$. By definition, $F((\hat{J}_n, \hat{\theta}_{f,n})) \equiv (\hat{P}_n, \hat{\theta}_{f,n})$ and $F((J^*, \theta_f^*)) \equiv (P^*, \theta_f^*)$. Eq. 2.32 holds by the delta method provided that F is continuously differentiable, as we now verify.

For a coordinate $j = 1, \ldots, |\tilde{A}X|$ and $\tilde{j} = 1, \ldots, |AX|$ representing coordinates $(a, x) \in \tilde{A}X$ and $(\tilde{a}, \tilde{x}) \in AX$, respectively,

$$\begin{aligned}
\frac{\partial F_{(a,x)}(z)}{\partial z_{(\tilde{a},\tilde{x})}} &= 1[x = \tilde{x}] \left[\left(\frac{\sum_{\check{a} \in A} z_{(\check{a},\tilde{x})} - z_{(a,\tilde{x})}}{(\sum_{\check{a} \in A} z_{(\check{a},\tilde{x})})^2} \right) 1[a = \tilde{a}] + \left(\frac{-z_{(a,\tilde{x})}}{(\sum_{\check{a} \in A} z_{(\check{a},\tilde{x})})^2} \right) 1[a \neq \tilde{a}] \right] \\
&= \frac{1[x = \tilde{x}]}{(\sum_{\check{a} \in A} z_{(\check{a},\tilde{x})})} [1[a = \tilde{a}] - \frac{z_{(\tilde{a},\tilde{x})}}{(\sum_{\check{a} \in A} z_{(\check{a},\tilde{x})})}],
\end{aligned}$$

provided that $\sum_{\check{a} \in A} z_{(\check{a},\tilde{x})} > 0$. Since $\sum_{\tilde{a} \in A} J^*(\tilde{a}, x) = J^*(x) > 0$ for all $x \in X$, we then conclude that F is continuously differentiable at (J^*, θ_f^*) and $DF(J^*, \theta_f^*)$ is a block diagonal matrix given by

$$DF(J^*, \theta_f^*) = \begin{bmatrix} \Sigma & \mathbf{0}_{|X\tilde{A}| \times d_{\theta_f}} \\ \mathbf{0}_{d_{\theta_f} \times |\tilde{A}X|} & \mathbf{I}_{d_{\theta_f} \times d_{\theta_f}} \end{bmatrix}.$$

By application of the delta method, the desired result follows. □

Lemma 29. *Assume Assumptions 11-15. Then,*

$$n^{\min\{\delta, 1/2\}} (\hat{P}_n^0 - P^*) \xrightarrow{d} N(B_P 1\{\delta \leqslant 1/2\}, \Omega_P 1\{\delta \geqslant 1/2\}),$$

where $B_P \equiv DG_P(\Pi^*)B_{\Pi^*}$, $\Omega_P \equiv DG_P(\Pi^*)(diag(\Pi^*) - \Pi^*\Pi^{*\prime})DG_P(\Pi^*)'$, *and* $DG_P(\Pi^*)$ *denotes the gradient of* P*-component of* G *in Assumption 16.*

Proof. This proof is analogous to that of Lemma 27 and is therefore omitted. □

Lemma 30. *Assume a non-parametric model for the first stage, i.e.,*

$$\theta_f \equiv \{f(x'|a,x)\}_{(a,x,x')\in AX^2}.$$

Then, the preliminary estimators $(\hat{\theta}_{f,n}, \hat{P}_n^0)$ *defined in Eqs. 2.12-2.11 satisfy Assumption 16.*

Proof. By definition, $(\hat{\theta}_{f,n}, \hat{P}_n^0)$ satisfies $(\hat{\theta}_{f,n}, \hat{P}_n^0) = G(\hat{\Pi}_n)$ with $G : \mathbb{R}^{|AX^2|} \to \mathbb{R}^{|AX^2|} \times \mathbb{R}^{|AX|}$ defined as follows: $G_j(\Pi) \equiv \sum_{\tilde{x}'\in X} \Pi(a,x,\tilde{x}') / \sum_{(\check{a},\check{x}')\in AX} \Pi(\check{a}x,\check{x}')$ for $j = 1,\ldots,|AX^2|$ and $G_j(\Pi) \equiv \Pi(a,x,x') / \sum_{\tilde{x}'\in X} \Pi(a,x,\tilde{x}')$ for $j = |AX^2| + 1,\ldots,|AX^2| + |AX|$. Also, let us define $\mathcal{N}_{\Pi^*} \equiv \{\Pi \in \mathbb{R}^{|AX^2|} : \Pi^*(a,x,x') \geqslant \eta/2\}$ for $\eta \equiv \inf_{(a,x,x')\in AX^2} \Pi^*(a,x,x') > 0$. Notice that this implies that $\Pi^* \in \mathcal{N}_{\Pi^*}$.

Assumption 16(a) is automatically satisfied because G is a constant function. Assumption 16(b) follows from the fact that $\Pi(a,x,x') \geqslant \eta/2 > 0$ for all $\Pi \in \mathcal{N}_{\Pi^*}$. Finally, Assumption 16(c) follows from the definition of G as it implies that $G(\Pi^*) = (f^*, P^*) = (\theta_f^*, P^*)$. □

Lemma 31. *The following results hold:*

1. $\Sigma J^* = \mathbf{0}_{|\tilde{A}X|}$,

2. $\Sigma B_J = \{B_J(a,x)/\pi^*(x)\}_{(a,x)\in\tilde{A}X}$,

3. $\Sigma diag\{J^*\}\Sigma' = diag\{\Omega_{PP}(x)/m^*(x)\}_{x\in X}$*, where for every* $x \in X$*,*

$$\Omega_{PP}(x) = [diag\{P^*(a|x)\}_{a\in\tilde{A}} - \{P^*(a|x)\}_{a\in\tilde{A}}\{P^*(a|x)\}'_{a\in\tilde{A}}]$$

and so

$$\Omega^{-1}_{PP}(x) = [diag\{\{1/P^*(a|x)\}_{a\in\tilde{A}}\} + (1/P^*(|A||x))\mathbf{1}_{|\tilde{A}-1|\times|\tilde{A}-1|}].$$

4. $(\Sigma\Omega_{JJ}\Sigma')^{-1} = diag\{m^*(x)[diag\{\{1/P^*(a|x)\}_{a\in\tilde{A}}\} + (1/P^*(|A||x))\mathbf{1}_{|\tilde{A}|\times|\tilde{A}|}]\}_{x\in X}$,

5. $(\Sigma\Omega_{JJ}\Sigma')^{-1}\Sigma = diag\{[diag\{\{1/P^*(a|x)\}_{a\in\tilde{A}}\}, -P^*(|A||x)^{-1}\mathbf{1}_{|\tilde{A}|\times 1}]\}_{x\in X}$.

Proof. Part 1. Notice that:

$$\begin{aligned}
\Sigma J^* &= diag\{\Sigma_x\}_{x\in X}J^* = diag\{\Sigma_x J^*(\cdot,x)\}_{x\in X} \\
&= diag\{\frac{1}{m^*(x)}[\mathbf{I}_{|\tilde{A}|\times|A|}J^*(\cdot,x) - \{P^*(a|x)\}_{a\in\tilde{A}}\mathbf{1}_{1\times|A|}J^*(\cdot,x)]\}_{x\in X} \\
&= diag\{\frac{1}{m^*(x)}[\{J^*(a,x)\}_{a\in\tilde{A}} - \{P^*(\tilde{a}|x)\}_{\tilde{a}\in\tilde{A}}m^*(x)]\}_{x\in X} \\
&= diag\{\frac{1}{m^*(x)}\mathbf{0}_{|\tilde{A}|}\}_{x\in X} = \mathbf{0}_{|\tilde{A}|\times|X|}
\end{aligned}$$

where we have used that $\mathbf{I}_{|\tilde{A}|\times|A|}J^*(\cdot,x) = \{J^*(a,x)\}_{a\in\tilde{A}}$, $\mathbf{1}_{1\times|A|}J^*(\cdot,x) = m^*(x)$, and $P^*(a|x)m^*(x) = J^*(a,x)$.

Part 2. Notice that:

$$\begin{aligned}
\Sigma B_J &= diag\{\Sigma_x\}_{x\in X}B_J = diag\{\Sigma_x B_J(\cdot,x)\}_{x\in X} \\
&= diag\{\frac{1}{\pi^*(x)}[\mathbf{I}_{|\tilde{A}|\times|A|}B_J(\cdot,x) - \{P^*(\tilde{a},x)\}_{\tilde{a}\in\tilde{A}}\mathbf{1}_{1\times|A|}B_J(\cdot,x)]\}_{x\in X} \\
&= \{B_J(a,x)/\pi^*(x)\}_{(a,x)\in\tilde{A}X},
\end{aligned}$$

Part 3. The first display is the result of algebraic derivations. The second display follows from Seber (2008, result 15.5).

Part 4. Consider the following argument. By definition, $\Omega_{JJ} \equiv diag\{J^*\} - J^*J^{*\prime}$. This and previous parts imply that

$$\Sigma\Omega_{JJ}\Sigma' = \Sigma diag\{J^*\}\Sigma' = diag\{\Omega_{PP}(x)/m^*(x)\}_{x\in X}.$$

Notice then that $diag\{\Omega_{PP}(x)/m^*(x)\}_{x\in X}$ is a block-diagonal matrix and each block is invertible. Then, by elementary properties of block-diagonal matrices, the result follows.

Part 5. Notice that:

$$(\Sigma\Omega_{JJ}\Sigma')^{-1}\Sigma = diag\{m^*(x)\tilde{\Omega}_{PP}^{-1}(x)\}_{x\in X} diag\{\Sigma_x\}_{x\in X} = diag\{m^*(x)\Omega_{PP}^{-1}(x)\Sigma_x\}_{x\in X},$$

where:

$$\begin{aligned} & m^*(x)\Omega_{PP}^{-1}(x)\Sigma_x \\ = & \left\{ \begin{array}{c} [diag\{\{1/P^*(a|x)\}_{a\in\tilde{A}}\} + (1/P^*(|A||x))\mathbf{1}_{|\tilde{A}-1|\times|\tilde{A}-1|}, \mathbf{0}_{\tilde{A}\times 1}] \\ -\mathbf{1}_{|\tilde{A}|\times|A|} + (1-(1/P^*(|A||x)))\mathbf{1}_{|\tilde{A}|\times|A|} \end{array} \right\} \\ = & [diag\{\{1/P^*(a|x)\}_{a\in\tilde{A}}\} + (1/P^*(|A||x))\mathbf{1}_{|\tilde{A}-1|\times|\tilde{A}-1|}]. \end{aligned}$$

This completes the step and the proof. □

Lemma 32. *Assume Assumptions 9-10. For any* $\lambda, \tilde{\lambda} \in \{\theta_f, \alpha\}$,

$$\begin{aligned} \frac{\partial\{\ln P_{\theta*}(a|x)\}_{(a,x)\in AX}}{\partial\lambda} &= \frac{\partial P_{\theta*}}{\partial\lambda}(\Sigma\Omega_{JJ}\Sigma')^{-1}\Sigma, \\ E_{J*}[\frac{\partial\{\ln P_{\theta*}(a|x)\}_{(a,x)\in AX}}{\partial\lambda}\frac{\partial\{\ln P_{\theta*}(a|x)\}_{(a,x)\in AX}}{\partial\lambda}'] &= \frac{\partial P_{\theta*}}{\partial\lambda}(\Sigma\Omega_{JJ}\Sigma')^{-1}\frac{\partial P_{\theta*}}{\partial\tilde{\lambda}}'. \end{aligned}$$

Proof. Before deriving the results, consider some preliminary observations. Notice that $\sum_{a\in A} P_{\theta*}(a|x) = 1$ and so $\partial P_{\theta*}(|A||x)/\partial\lambda' = -\sum_{a\in\tilde{A}} \partial P_{\theta*}(|A||x)/\partial\lambda'$. Also, notice that $P^*(a|x) = P_{\theta*}(a|x)$ and so $(\partial P_{\theta*}(a|x)/\partial\lambda)(1/P^*(a|x)) = \partial \ln P_{\theta*}(a|x)/\partial\lambda$.

For the first result, consider the following derivation:

$$\begin{aligned}
&\frac{\partial P_{\theta*}}{\partial\lambda}(\Sigma\Omega_{JJ}\Sigma')^{-1}\Sigma \\
= \;& [\frac{\partial\{P_{\theta*}(a|x)\}_{(a,x)\in\tilde{A}X}}{\partial\lambda}]diag\{[diag\{\{1/P^*(a|x)\}_{a\in\tilde{A}}\}, -P^*(|A||x)^{-1}\mathbf{1}_{|\tilde{A}|\times 1}]\}_{x\in X} \\
= \;& \{[\frac{\partial\{\ln P_{\theta*}(a|x)\}_{a\in\tilde{A}}}{\partial\lambda}, -\sum_{a\in\tilde{A}} \frac{\partial P_{\theta*}(a|x)}{\partial\lambda}\frac{1}{P^*(|A||x)}]\}_{x\in X} \\
= \;& \{\frac{\partial\{\ln P_{\theta*}(a|x)\}_{a\in A}}{\partial\lambda}\}_{x\in X} = \frac{\partial\{\ln P_{\theta*}(a|x)\}_{(a,x)\in AX}}{\partial\lambda}
\end{aligned}$$

where the first equality uses Lemma 31 and rest of the equalities use the preliminary observations.

For the second result, consider the following derivation:

$$\begin{aligned}
&\frac{\partial P_{\theta*}}{\partial\lambda}(\Sigma\Omega_{JJ}\Sigma')^{-1}\frac{\partial P_{\theta*}}{\partial\tilde{\lambda}}' \\
&= \sum_{x\in X} m^*(x) \sum_{a\in A} \frac{\partial P_{\theta*}(a|x)}{\partial\lambda}\frac{\partial P_{\theta*}(a|x)}{\partial\tilde{\lambda}'}\frac{1}{P^*(a|x)} \\
= \;& \sum_{(a,x)\in AX} J^*(a,x)\frac{\partial \ln P_{\theta*}(a|x)}{\partial\lambda}\frac{\partial \ln P_{\theta*}(a|x)}{\partial\tilde{\lambda}'} \\
&= E_{J*}[\frac{\partial\{\ln P_{\theta*}(a|x)\}_{(a,x)\in AX}}{\partial\lambda}\frac{\partial\{\ln P_{\theta*}(a|x)\}_{(a,x)\in AX}}{\partial\tilde{\lambda}}'],
\end{aligned}$$

where the first equality uses Lemma 31 and rest of the equalities use the preliminary observations. □

2.8 Review of results on extremum estimators

The purpose of this section is to prove the consistency and asymptotic normality of extremum estimators under certain regularity conditions. Relative to the standard results in the literature (e.g. McFadden and Newey (1994)), our arguments allow for (a) a rate of convergence that is not necessarily $\sqrt{n}$ and (b) sequences of data generating processes that change as a function of the sample size. Both of these modifications are important for our theoretical results.

Theorem 33. *Assume the following:*

(a) $Q_n(\theta)$ *converges uniformly in probability to* $Q(\theta)$ *along* $\{P_n\}_{n\geqslant 1}$*, i.e., for any* $\varepsilon > 0$*,*

$$P_n(\sup_{\theta\in\Theta} |Q_n(\theta) - Q(\theta)| > \varepsilon) = o(1).$$

(b) $Q(\theta)$ *is upper semi-continuous, i.e., for any* $\{\theta_n\}_{n\geqslant 1}$ *with* $\theta_n \to \tilde{\theta}$*,*

$$\limsup Q(\theta_n) \leqslant Q(\tilde{\theta}).$$

(c) $Q(\theta)$ *is uniquely maximized at* $\theta = \theta^*$*.*

Consider an estimator $\hat{\theta}_n \in \Theta$ *that satisfies* $Q_n(\hat{\theta}_n) = \max_{\theta\in\Theta} Q_n(\theta)$*. Then,* $\hat{\theta}_n = \theta^* + o_{P_n}(1)$*.*

Proof. Part 1. Fix $\mu > 0$ arbitrarily. First, we show that $\Theta \cap \{\theta : \|\theta - \theta^*\| \geqslant \mu\}$ is compact. Since Θ is compact, it is Θ is closed. Since Θ and $\{\theta : \|\theta - \theta^*\| \geqslant \mu\}$ are closed, $\Theta \cap \{\theta : \|\theta - \theta^*\| \geqslant \mu\}$ is closed. Hence, $\Theta \cap \{\theta : \|\theta - \theta^*\| \geqslant \mu\}$ is a closed subset of Θ and, hence, $\Theta \cap \{\theta : \|\theta - \theta^*\| \geqslant \mu\}$ is compact.

We show that $\exists \varepsilon = \varepsilon(\mu) > 0$ s.t.

$$\{\theta \in \Theta : \|\theta - \theta^*\| > \mu\} \subseteq \{\Theta \cap \{\theta : \|\theta - \theta^*\| \geqslant \mu\}\} \subseteq \{Q(\theta) - Q(\theta^*) < \varepsilon\}.$$

The first inclusion is trivial. To prove the second inclusion, suppose not, i.e., suppose that $\forall \{\varepsilon_n\}_{n \geqslant 1} \downarrow 0$, $\exists \{\theta_n\}_{n \geqslant 1}$ with

$$\theta_n \in \{\Theta \cap \{\theta : \|\theta - \theta^*\| \geqslant \mu\}\} \cap \{Q(\theta_n) - Q(\theta^*) \geqslant \varepsilon_n\}.$$

By this and our assumptions, $Q(\theta_n) \geqslant \varepsilon_n + Q(\theta^*) \geqslant \varepsilon_n + Q(\theta_n)$ and so $\lim Q(\theta_n) = Q(\theta^*)$. Since Θ is compact, there is a subsequence $\{a_n\}_{n \geqslant 1}$ of $\{n\}_{n \geqslant 1}$ s.t. $\theta_{a_n} \to \tilde{\theta} \in \Theta \cap \{\theta : \|\theta - \theta^*\| \geqslant \mu\}$. By upper semi-continuity and subsequence converging to the same limit,

$$Q(\tilde{\theta}) \geqslant \limsup Q(\theta_{a_n}) = \lim Q(\theta_{a_n}) = Q(\theta^*).$$

By our assumptions, $\tilde{\theta} = \theta^*$. But this implies that $\theta^* \in \{\theta : \|\theta - \theta^*\| > \mu\}$, which is a contradiction.

Part 2. Fix $\mu > 0$ arbitrarily. By part 1, $\exists \varepsilon > 0$ s.t.

$$P_n(||\hat{\theta}_n - \theta^*|| > \mu) \leqslant P_n(Q(\hat{\theta}_n) - Q(\theta^*) < \varepsilon).$$

The proof is completed by showing that the RHS is $o(1)$.

Consider the following derivation:

$$\begin{aligned} Q(\hat{\theta}_n) - Q(\theta^*) &= Q(\hat{\theta}_n) - Q_n(\hat{\theta}_n) + Q_n(\hat{\theta}_n) - Q_n(\theta^*) + Q_n(\theta^*) - Q(\theta^*) \\ &\geqslant 2 \sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta)|, \end{aligned}$$

where we have used that $Q_n(\hat{\theta}_n) - Q_n(\theta^*) \geqslant 0$ by definition of $\hat{\theta}_n$. Therefore:

$$P_n(Q(\hat{\theta}_n) - Q(\theta^*) > \varepsilon) \leqslant P_n(\sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta)| > \frac{\varepsilon}{2}) = o(1),$$

where the convergence occurs by our assumptions. This completes the proof. □

Theorem 34. *Consider an estimator $\hat{\theta}_n$ of a parameter θ^* that satisfies $\hat{\theta}_n \in \Theta$ and $Q_n(\hat{\theta}_n) = \max_{\theta \in \Theta} Q_n(\theta)$. Furthermore,*

(a) $\hat{\theta}_n = \theta^* + o_{p_n}(1)$,

(b) $\theta^* \in Int(\Theta)$,

(c) Q_n *is twice continuously differentiable in a neighborhood* $\mathcal{N}$ *of* θ^* *w.p.a.1,*

(d) For some $\delta > 0$, $n^\delta \partial Q_n(\theta^*)/\partial\theta \xrightarrow{d} Z$ *for some random variable* Z *along* $\{P_n\}_{n \geqslant 1}$,

(e) $\sup_{\theta \in \mathcal{N}} ||\partial^2 Q_n(\theta)/\partial\theta\partial\theta' - H(\theta)|| = o_{p_n}(1)$ *for some function* $H : \mathcal{N} \to \mathbb{R}^{k \times k}$ *that is continuous at* θ^*,

(f) $H(\theta^*)$ *is non-singular.*

Then,

1. $n^\delta(\hat{\theta}_n - \theta^*) = -H(\theta^*)^{-1} n^\delta \partial Q_n(\theta^*)/\partial\theta + o_{p_n}(1)$.

2. $n^\delta(\hat{\theta}_n - \theta^*) \xrightarrow{d} -H(\theta^*)^{-1} Z$ *along* $\{P_n\}_{n \geqslant 1}$.

Proof. Given our assumptions, part 1 implies part 2, so we only show part 1. Without loss of generality, we assume that $\mathcal{N}$ is an open convex set contained in Θ. Let $\mathcal{E}_n$ denote the event that Q_n is twice continuously differentiable in $\mathcal{N}$ and $\partial^2 Q_n(\theta)/\partial\theta\partial\theta'$ is non-singular in $\mathcal{N}$. Throughout this proof, let us define $G_{n,1}(\theta) \equiv \partial Q_n(\theta)/\partial\theta \times 1[\theta \in \mathcal{N}]$, $G_{n,2}(\theta) \equiv \partial^2 Q_n(\theta)/\partial\theta\partial\theta' \times 1[\theta \in \mathcal{N}]$, and $G_{n,3}(\theta) \equiv (\partial^2 Q_n(\theta)/\partial\theta\partial\theta')^{-1} \times 1[\theta \in \mathcal{N}]$ if $\mathcal{E}_n$ occurs and $G_{n,1}(\theta) \equiv G_{n,2}(\theta) \equiv G_{n,3}(\theta) \equiv 0$ if $\mathcal{E}_n$ does not occur.

For the sake of clarity, we divide the argument into several steps.

Step 1. Throughout this step, we assume that Q_n is twice continuously differentiable in $\mathcal{N}$. Fix $n \in \mathbb{N}$ and $\theta \in \mathcal{N}$ arbitrarily. Let $f : [0,1] \to \mathbb{R}$ be defined by $f(\lambda) \equiv \partial Q_n(\theta_\lambda)/\partial\theta$, where $\theta_\lambda \equiv \lambda\theta + (1-\lambda)\theta^*$. Since Q_n is twice continuously differentiable with respect to θ_λ, $f(\lambda)$ is continuously differentiable in $\lambda \in [0,1]$. A one-term Taylor expansion evaluated at $\lambda = 1$ implies that for some $\mu \in (0,1)$,

$$\partial Q_n(\theta)/\partial\theta = \partial Q_n(\theta^*)/\partial\theta + \partial Q_n(\theta_\mu)/\partial\theta\partial\theta' \times (\theta - \theta^*).$$

Step 2. Throughout this step, we assume that Q_n is twice continuously differentiable in $\mathcal{N}$, $\partial^2 Q_n(\theta)/\partial\theta\partial\theta'$ is non-singular in $\mathcal{N}$, and that $\hat{\theta}_n \in \mathcal{N}$. First, $\hat{\theta}_n$ is the minimum of a twice continuous differentiable function and, thus, $\partial Q_n(\hat{\theta}_n)/\partial\theta = \mathbf{0}_k$. Second, step 1 implies that there exists a value $\hat{\lambda}_n \in (0,1)$ (possibly dependent on n and $\hat{\theta}_n$) s.t.

$$\mathbf{0}_k = \partial Q_n(\theta^*)/\partial\theta + \partial Q_n(\tilde{\theta}_n)/\partial\theta\partial\theta' \times (\hat{\theta}_n - \theta^*),$$

where we have used the definition

$$\tilde{\theta}_n \equiv (\hat{\lambda}_n\hat{\theta}_n + (1-\hat{\lambda}_n)\theta^*)1(\hat{\theta}_n \in \mathcal{N}) + \theta^* 1(\hat{\theta}_n \notin \mathcal{N}). \tag{2.33}$$

Notice that $\tilde{\theta}_n \in \mathcal{N}$ as a result of $\hat{\theta}_n, \theta^* \in \mathcal{N}$ and $\mathcal{N}$ being a convex set. From this, we conclude that:

$$-(\partial Q_n(\tilde{\theta}_n)/\partial\theta\partial\theta')^{-1} n^\delta \partial Q_n(\theta^*)/\partial\theta = n^\delta(\hat{\theta}_n - \theta^*).$$

Step 3. The goal of this step is to show that:

$$(G_{n,3}(\tilde{\theta}_n) - H(\theta^*)^{-1}) n^\delta \partial Q_n(\theta^*)/\partial\theta = o_{p_n}(1),$$

where $\tilde{\theta}_n$ is as in Eq. 2.33. We divide the argument into steps.

First, we show that:

$$G_{n,2}(\tilde{\theta}_n) - H(\tilde{\theta}_n) = o_{p_n}(1). \tag{2.34}$$

Fix $\varepsilon > 0$ arbitrarily and consider the following derivation:

$$\begin{aligned}
&P_n(||G_{n,2}(\tilde{\theta}_n) - H(\tilde{\theta}_n)|| > \eta) \\
&\quad = P_n(\{||G_{n,2}(\tilde{\theta}_n) - H(\tilde{\theta}_n)|| > \eta\} \cap \{\hat{\theta}_n \in \mathcal{N}\} \cap \mathcal{E}_n) \\
&\qquad + P_n(\{||G_{n,2}(\tilde{\theta}_n) - H(\tilde{\theta}_n)|| > \eta\} \cap \{\{\hat{\theta}_n \notin \mathcal{N}\} \cup \mathcal{E}_n^c\}) \\
&\quad \leqslant P_n(\sup_{\theta \in \mathcal{N}} ||\partial Q_n(\theta)/\partial\theta\partial\theta' - H(\theta)|| > \eta) + P_n(\hat{\theta}_n \notin \mathcal{N}) + P_n(\mathcal{E}_n^c).
\end{aligned}$$

The RHS is a sum of three terms and all of them converge to zero under our assumptions. In turn, from this and our assumptions, we conclude that $G_{n,2}(\tilde{\theta}_n) = H(\theta^*) + o_{p_n}(1)$ and, so, by the CMT, we conclude that:

$$G_{n,2}(\tilde{\theta}_n)^{-1} = H(\theta^*)^{-1} + o_{p_n}(1). \tag{2.35}$$

As a next step, we show that:

$$G_{n,3}(\tilde{\theta}_n) = H(\theta^*)^{-1} + o_{p_n}(1). \tag{2.36}$$

Fix $\varepsilon > 0$ arbitrarily and consider the following derivation:

$$\begin{aligned}
&P_n(||G_{n,3}(\tilde{\theta}_n) - H(\theta^*)^{-1}|| > \varepsilon) \\
&\quad = P_n(\{||G_{n,3}(\tilde{\theta}_n) - H(\theta^*)^{-1}|| > \varepsilon\} \cap \{\hat{\theta}_n \in \mathcal{N}\} \cap \mathcal{E}_n) \\
&\qquad + P_n(\{||G_{n,3}(\tilde{\theta}_n) - H(\theta^*)^{-1}|| > \varepsilon\} \cap \{\{\hat{\theta}_n \notin \mathcal{N}\} \cup \mathcal{E}_n^c\}) \\
&\quad \leqslant P_n(\{||G_{n,2}(\tilde{\theta}_n)^{-1} - H(\theta^*)^{-1}|| > \varepsilon\} \cap \{\hat{\theta}_n \in \mathcal{N}\} \cap \mathcal{E}_n) \\
&\qquad + P_n(\hat{\theta}_n \notin \mathcal{N}) + P_n(\mathcal{E}_n^c),
\end{aligned}$$

The RHS is a sum of three terms. The first term converges to zero by Eq. 2.35 and the other two converge to zero by our assumptions. Finally, Eq. 2.36 and our assumptions imply Eq. 2.34.

Step 4. Conclude the proof. Fix $\varepsilon > 0$ arbitrarily and consider the following derivation:

$$
\begin{aligned}
&P_n(||n^{\delta}(\hat{\theta}_n - \theta^*) - H(\theta^*)^{-1} n^{\delta} \partial Q_n(\theta^*)/\partial\theta|| > \varepsilon) \\
&\quad = \left\{ \begin{array}{c} P_n(||n^{\delta}(\hat{\theta}_n - \theta^*) - H(\theta^*)^{-1} n^{\delta} \partial Q_n(\theta^*)/\partial\theta|| > \varepsilon \cap \{\hat{\theta}_n \in \mathcal{N}\} \cap \mathcal{E}_n) + \\ P_n(||n^{\delta}(\hat{\theta}_n - \theta^*) - H(\theta^*)^{-1} n^{\delta} \partial Q_n(\theta^*)/\partial\theta|| > \varepsilon \cap \{\{\hat{\theta}_n \in \mathcal{N}\}^c \cup \mathcal{E}_n^c\}) \end{array} \right\} \\
&\quad \leqslant P_n(||(G_{n,3}(\tilde{\theta}_n) - H(\theta^*)^{-1}) n^{s} \partial Q_n(\theta^*)/\partial\theta|| > \varepsilon) \\
&\qquad + P_n(\hat{\theta}_n \notin \mathcal{N}) + P_n(\mathcal{E}_n^c),
\end{aligned}
$$

where the last inequality relies on step 2. The RHS is a sum of three terms. The first term converges to zero by step 3 and the last two converge zero by our assumptions.

□

2.9 Tables

Table 2.1: Monte Carlo results for $\theta_{u,2}$ under correct specification.

Steps	Statistic	ML			MD(W^*_{AV})			MD($\mathbf{I}$)			MD(W^*_{AMSE})		
		$1K$	$5K$	$10K$	$1K$	$5K$	$10K$	$1K$	$5K$	$10K$	$1K$	$5K$	$10K$
$K=1$	$\sqrt{n}$·Bias	0.01	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00
	$\sqrt{n}$·SD	0.22	0.22	0.22	0.22	0.22	0.22	0.24	0.23	0.24	0.22	0.22	0.22
	n·MSE	0.05	0.05	0.05	0.05	0.05	0.05	0.06	0.05	0.06	0.05	0.05	0.05
$K=2$	$\sqrt{n}$·Bias	0.01	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00
	$\sqrt{n}$·SD	0.22	0.22	0.22	0.22	0.22	0.22	0.24	0.23	0.24	0.22	0.22	0.22
	n·MSE	0.05	0.05	0.05	0.05	0.05	0.05	0.06	0.05	0.06	0.05	0.05	0.05
$K=3$	$\sqrt{n}$·Bias	0.01	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00
	$\sqrt{n}$·SD	0.22	0.22	0.22	0.22	0.22	0.22	0.24	0.23	0.24	0.22	0.22	0.22
	n·MSE	0.05	0.05	0.05	0.05	0.05	0.05	0.06	0.05	0.06	0.05	0.05	0.05
$K=10$	$\sqrt{n}$·Bias	0.01	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00
	$\sqrt{n}$·SD	0.22	0.22	0.22	0.22	0.22	0.22	0.24	0.23	0.24	0.22	0.22	0.22
	n·MSE	0.05	0.05	0.05	0.05	0.05	0.05	0.06	0.05	0.06	0.05	0.05	0.05

Table 2.2: Monte Carlo results for $\theta_{u,2}$ under local specification with $\theta_{u,3} \propto n^{-1}$.

Steps	Statistic	ML			MD(W^*_{AV})			MD($\mathbf{I}$)			MD(W^*_{AMSE})		
		$1K$	$5K$	$10K$	$1K$	$5K$	$10K$	$1K$	$5K$	$10K$	$1K$	$5K$	$10K$
$K=1$	$\sqrt{n}$·Bias	0.03	0.01	0.01	0.03	0.01	0.01	0.03	0.01	0.01	0.03	0.01	0.01
	$\sqrt{n}$·SD	0.22	0.22	0.22	0.22	0.22	0.22	0.24	0.23	0.23	0.22	0.22	0.22
	n·MSE	0.05	0.05	0.05	0.05	0.05	0.05	0.06	0.05	0.05	0.05	0.05	0.05
$K=2$	$\sqrt{n}$·Bias	0.02	0.01	0.01	0.02	0.01	0.01	0.02	0.01	0.01	0.02	0.01	0.01
	$\sqrt{n}$·SD	0.22	0.22	0.22	0.22	0.22	0.22	0.24	0.23	0.23	0.22	0.22	0.22
	n·MSE	0.05	0.05	0.05	0.05	0.05	0.05	0.06	0.05	0.05	0.05	0.05	0.05
$K=3$	$\sqrt{n}$·Bias	0.02	0.01	0.01	0.02	0.01	0.01	0.02	0.01	0.01	0.02	0.01	0.01
	$\sqrt{n}$·SD	0.22	0.22	0.22	0.22	0.22	0.22	0.24	0.23	0.23	0.22	0.22	0.22
	n·MSE	0.05	0.05	0.05	0.05	0.05	0.05	0.06	0.05	0.05	0.05	0.05	0.05
$K=10$	$\sqrt{n}$·Bias	0.02	0.01	0.01	0.02	0.01	0.01	0.02	0.01	0.01	0.02	0.01	0.01
	$\sqrt{n}$·SD	0.22	0.22	0.22	0.22	0.22	0.22	0.24	0.23	0.23	0.22	0.22	0.22
	n·MSE	0.05	0.05	0.05	0.05	0.05	0.05	0.06	0.05	0.05	0.05	0.05	0.05

Table 2.3: Monte Carlo results for $\theta_{u,2}$ under local specification with $\theta_{u,3} \propto n^{-1/2}$.

Steps	Statistic	ML			MD(W^*_{AV})			MD($\mathbf{I}$)			MD(W^*_{AMSE})		
		$1K$	$5K$	$10K$	$1K$	$5K$	$10K$	$1K$	$5K$	$10K$	$1K$	$5K$	$10K$
$K=1$	$\sqrt{n}$·Bias	0.54	0.54	0.55	0.54	0.54	0.55	0.50	0.50	0.51	0.40	0.40	0.41
	$\sqrt{n}$·SD	0.24	0.23	0.23	0.25	0.23	0.23	0.27	0.25	0.24	0.36	0.33	0.33
	n·MSE	0.35	0.35	0.35	0.35	0.35	0.35	0.33	0.31	0.32	0.29	0.27	0.28
$K=2$	$\sqrt{n}$·Bias	0.54	0.54	0.54	0.53	0.54	0.54	0.50	0.50	0.50	0.45	0.43	0.43
	$\sqrt{n}$·SD	0.24	0.23	0.23	0.25	0.23	0.23	0.27	0.25	0.24	0.34	0.32	0.32
	n·MSE	0.35	0.34	0.35	0.34	0.34	0.35	0.32	0.31	0.31	0.32	0.29	0.29
$K=3$	$\sqrt{n}$·Bias	0.54	0.54	0.54	0.53	0.54	0.54	0.50	0.50	0.50	0.45	0.43	0.43
	$\sqrt{n}$·SD	0.24	0.23	0.23	0.25	0.23	0.23	0.27	0.25	0.24	0.34	0.32	0.32
	n·MSE	0.35	0.34	0.35	0.34	0.34	0.35	0.32	0.31	0.31	0.32	0.29	0.29
$K=10$	$\sqrt{n}$·Bias	0.54	0.54	0.54	0.53	0.54	0.54	0.50	0.50	0.50	0.45	0.43	0.43
	$\sqrt{n}$·SD	0.24	0.23	0.23	0.25	0.23	0.23	0.27	0.25	0.24	0.34	0.32	0.32
	n·MSE	0.35	0.34	0.35	0.34	0.34	0.35	0.32	0.31	0.31	0.32	0.29	0.29

Table 2.4: Monte Carlo results for $\theta_{u,2}$ under local specification with $\theta_{u,3} \propto n^{-1/3}$ using the regular scaling.

Steps	Statistic	ML			MD(W^*_{AV})			MD($\mathbf{I}$)			MD(W^*_{AMSE})		
		$1K$	$5K$	$10K$	$1K$	$5K$	$10K$	$1K$	$5K$	$10K$	$1K$	$5K$	$10K$
$K=1$	$\sqrt{n}$·Bias	1.61	2.16	2.44	1.56	2.13	2.41	1.44	1.96	2.23	0.64	0.52	0.48
	$\sqrt{n}$·SD	0.29	0.26	0.25	0.29	0.26	0.25	0.33	0.28	0.27	0.75	0.87	0.91
	n·MSE	2.69	4.74	6.02	2.53	4.60	5.89	2.18	3.94	5.05	0.98	1.03	1.06
$K=2$	$\sqrt{n}$·Bias	1.61	2.16	2.44	1.56	2.13	2.41	1.44	1.97	2.23	0.80	0.64	0.56
	$\sqrt{n}$·SD	0.28	0.26	0.25	0.29	0.26	0.25	0.32	0.28	0.27	0.65	0.79	0.86
	n·MSE	2.66	4.73	6.02	2.52	4.60	5.89	2.17	3.95	5.07	1.06	1.03	1.05
$K=3$	$\sqrt{n}$·Bias	1.61	2.16	2.44	1.56	2.13	2.41	1.44	1.97	2.23	0.80	0.64	0.56
	$\sqrt{n}$·SD	0.28	0.26	0.25	0.29	0.26	0.25	0.32	0.28	0.27	0.65	0.78	0.85
	n·MSE	2.66	4.73	6.02	2.52	4.60	5.89	2.17	3.95	5.07	1.05	1.02	1.05
$K=10$	$\sqrt{n}$·Bias	1.61	2.16	2.44	1.56	2.13	2.41	1.44	1.97	2.23	0.80	0.64	0.56
	$\sqrt{n}$·SD	0.28	0.26	0.25	0.29	0.26	0.25	0.32	0.28	0.27	0.65	0.78	0.85
	n·MSE	2.66	4.73	6.02	2.52	4.60	5.89	2.17	3.95	5.07	1.05	1.02	1.05

Table 2.5: Monte Carlo results for $\theta_{u,2}$ under local specification with $\theta_{u,3} \propto n^{-1/3}$ using the correct scaling.

Steps	Statistic	ML			MD(W^*_{AV})			MD($\mathbf{I}$)			MD(W^*_{AMSE})		
		$1K$	$5K$	$10K$	$1K$	$5K$	$10K$	$1K$	$5K$	$10K$	$1K$	$5K$	$10K$
$K=1$	$n^{1/3}$·Bias	0.51	0.52	0.53	0.49	0.52	0.52	0.46	0.48	0.48	0.20	0.13	0.10
	$n^{1/3}$·SD	0.09	0.06	0.05	0.09	0.06	0.05	0.10	0.07	0.06	0.24	0.21	0.20
	$n^{2/3}$·MSE	0.27	0.28	0.28	0.25	0.27	0.27	0.22	0.23	0.23	0.10	0.06	0.05
$K=2$	$n^{1/3}$·Bias	0.51	0.52	0.53	0.49	0.51	0.52	0.45	0.48	0.48	0.25	0.15	0.12
	$n^{1/3}$·SD	0.09	0.06	0.05	0.09	0.06	0.05	0.10	0.07	0.06	0.20	0.19	0.19
	$n^{2/3}$·MSE	0.27	0.28	0.28	0.25	0.27	0.27	0.22	0.23	0.24	0.11	0.06	0.05
$K=3$	$n^{1/3}$·Bias	0.51	0.52	0.53	0.49	0.51	0.52	0.45	0.48	0.48	0.25	0.16	0.12
	$n^{1/3}$·SD	0.09	0.06	0.05	0.09	0.06	0.05	0.10	0.07	0.06	0.20	0.19	0.18
	$n^{2/3}$·MSE	0.27	0.28	0.28	0.25	0.27	0.27	0.22	0.23	0.24	0.11	0.06	0.05
$K=10$	$n^{1/3}$·Bias	0.51	0.52	0.53	0.49	0.51	0.52	0.45	0.48	0.48	0.25	0.16	0.12
	$n^{1/3}$·SD	0.09	0.06	0.05	0.09	0.06	0.05	0.10	0.07	0.06	0.20	0.19	0.18
	$n^{2/3}$·MSE	0.27	0.28	0.28	0.25	0.27	0.27	0.22	0.23	0.24	0.11	0.06	0.05

3

Unemployment misclassification and the earned income tax credit effect on unemployment

3.1 Introduction

The Earned Income Tax Credit (EITC) is one of the most influential anti-poverty programs in the United States for a few reasons. First, the EITC is one of the biggest means-tested transfer programs among the federal and state expenditures and the biggest among the cash transfer programs. According to the Internal Revenue Service (as found in http://www.eitc.irs.gov/ptoolkit/basicmaterials/ff/), the federal and state spending on the EITC was $66.7 billion in 2014 and, jointly with the Child Tax Credit, the EITC brought around 9.4 million people out of poverty in that year. Next, the EITC encourages individuals to work and, unlike other programs, it gives zero benefit to individuals with zero income. This feature makes the EITC popular to its political competitor, the negative income tax, in the sense that the negative

income tax discourages individuals from working by maximizing its benefit at zero hour worked. Hotz and Scholz (2003) discusses more details of the EITC features.

In addition to alleviating the current poverty, therefore, the EITC has an objective of alleviating the future poverty by encouraging self sufficiency among the low-income population and reducing their long-term welfare dependency. To evaluate this objective of the EITC, economists have investigated the consequence of the EITC implementation on the labor supply theoretically and empirically. The static labor supply model predicts that the EITC encourages nonworker to enter the labor force but that it has ambiguous effect on hours worked. The ambiguity comes from the non-convexity of the budget constraint created by the EITC system. In empirical research, a seminal paper by Eissa and Liebman (1996) concludes that this EITC expansion has positive effect on employment and negligible effect on hours worked.

The existing empirical analyses on the EITC and the labor supply have neither distinguished employment and labor force participation explicitly nor investigated them simultaneously (for example, Eissa and Liebman (1996) and Meyer and Rosenbaum (2001) define labor force participation as having positive annual hours worked). This treatment of employment and labor force participation is problematic in the sense that the previous literature cannot offer any insight into the effect of the EITC on unemployment, since unemployment is defined as the difference between labor force participation and employment. The previous empirical analyses cannot be directly extended to unemployment and labor force participation, because the definition of unemployment is controversial. The official definition of unemployment, given by Bureau of Labor Statistics (BLS), are based on employment status and the length of job search. According to their definition (as found in

http://www.bls.gov/bls/glossary.htm), unemployed persons are defined as "Persons aged 16 years and older who had no employment during the reference week, were available for work, except for temporary illness, and had made specific efforts to find employment sometime during the 4-week period ending with the reference week. Persons who were waiting to be recalled to a job from which they had been laid off need not have been looking for work to be classified as unemployed." However, the BLS definition of unemployment can be subject to misclassification. There are several criticisms on those definitions, e.g., Jones and Riddell (1999), Brandolini et al. (2006), Kingdon and Knight (2006), and Feng and Hu (2013), and the BLS itself provides six alternative definitions of unemployment (U1 through U6). First, the search intensity (job search activity) varies across individuals and the stated length is subject to reporting errors. Second, there are groups of individuals who are subtle to categorize in the labor force status. One example is marginally attached workers, who want a job but they are not actively searching for a job. Another example is future job starters, who have a job starting within the next four weeks.

This paper investigates the effect of the EITC on unemployment and labor force participation without adopting any particular definition of unemployment. Unemployment and labor force participation itself is a key outcome in the policy evaluation of work incentive and income redistribution programs. There exists no prediction from economic theory about the EITC effect on unemployment. The EITC effect on unemployment is empirically important to evaluate the EITC, because, even if economic theory predicts the EITC increases labor force participation, the increase in labor force participation can cause either an increase in employment or an increase in unemployment. Therefore, this paper investigates the EITC effect on unemploy-

ment with explicit distinction between employment and labor force participation. In other words, it evaluates the effects of the EITC on labor force participation and unemployment. Furthermore, it is meaningful to empirically re-evaluate the theoretical implication that the EITC encourages nonemployed individuals to work.

Instead of searching for an adequate definition of unemployment, this paper constructs lower and upper bounds for unemployment based on the three concepts in the Current Population Survey dataset: employment status, job search activity and willingness to work. In this framework, unemployment itself is not observable but an interval including unemployment is observable. From the econometric viewpoint, this is the same setting as Manski and Tamer (2002) and Molinari (2008), which theoretically investigate the identification and estimation of models with interval variable or misclassified variable. Their identification strategies assume that a variable of interest belongs to the observable interval. Since this paper wants to allow for any possible definition of unemployment, this approach is a suitable point to start for this paper. It does not guarantee the point identification but yeilds the identified set for the effect of the EITC on unemployment.

Given the above motivation and framework, this paper focuses on the effect of the Omnibus Budget Reconciliation Act of 1993, which caused the biggest change to the EITC in its history, on unemployment for single mothers. Single mothers are the biggest population in the welfare program in the United States and one of the most important groups in poverty analysis. Furthermore, the EITC has different schedules according to family structure and it favors the single parents the most, which makes single mothers the most relevant population to study in the EITC. As is often the case with the reduced form analysis, the results of this paper are applicable only to

the Omnibus Budget Reconciliation Act of 1993 and it cannot be generalized directly to the other change in EITC.

First, I use BLS definitions of unemployment and labor force participation. The estimated treatment effects on unemployment and labor force participation are both positive. Then I characterizes the bounds for the treatment effect of the EITC on unemployment and labor force participation, when the definition of unemployment is given by the interval. The inference results show that the confidence intervals imply positive treatment effect on labor force participation and negligible treatment effect on unemployment. This implies that the positive effect on unemployment is not robust to the ambiguity of unemployment definition; a small amount of the misclassification change the empirical results such that even the sign of the effect is not statistically significant.

The remainder of the paper is organized as follows. Section 3.2 discuss the EITC system and its effect on labor supply. Section 3.3 describes the data; Section 3.4 formulates the model and discusses the identification strategy; Section 3.5 estimates the model in Section 3.4; and Section 3.6 concludes.

3.2 Earned Income Tax Credit and its effect on employment

This section discuss the EITC system and the previous literature on its relationship with labor supply. The EITC is a near-cash transfer program to low-income working people, especially targeting on single parents. It provides the transfer to low-income households via refundable tax credit based on household earned income. The amount of the transfer is proportional to the income for the low income region (phrase-in region) in order to "make work pay".

Since its outset at 1975, the EITC has experienced several expansions in the tax acts of 1986, 1990, 1993, and 2001. Table 3.1 shows the changes in the maximum credits in EITC for 1990 through 1998. First, the credits are increasing over time, especially at the Omnibus Budget Reconciliation Act of 1993. Second, the EITC has different schedules for households with and without child. This paper ignores the requirements (age, relationship, and residential tests) for a child to be quantified for the EITC, because the CPS does not give information on those requirements. (The detailed explanation for the child quantification is found in http://www.irs.gov/Individuals/Qualifying-Child-Rules.)

In order to eligible for the EITC of 1998, an individual have only to have positive earned income below a specified amount (for example, $26,473 for households with only one child in 1998). However, the EITC neither offers substantial transfers to households without child nor to households with earned income near the upper bound of the EITC eligible income line. Therefore, a taxpayer should meet virtually two requirements: sufficiently low income and child. This supports the later usage of single women as a control group in comparison with the treatment group of single mothers.

A number of research papers have investigated the EITC effect on labor supply. The rest of this paragraph explains two papers which are relevant to this paper, and the extensive survey on this topic is found in Hotz and Scholz (2003). Eissa and Liebman (1996) uses the difference-in-difference approach to investigate the 1986 EITC expansion on single mothers' labor supply and concludes that this EITC expansion has positive effect on employment and negligible effect on hours worked. Meyer and Rosenbaum (2001) further proceed the analysis in Eissa and Liebman (1996) and an-

alyze the increase in employment for single mothers for 1984-1996. They construct the policy parameter variables from various welfare programs in the United States and evaluate the effect of these parameter variables on employment. They conclude that the EITC explains over 60% of the increase in employment during their sample period.

3.3 Data description

This paper uses the March Current Population Survey (March CPS) dataset for 1991 through 1997. The following analysis excludes the dataset for 1994 because it was the first year after the 1993 expansion of the EITC and individuals were likely to adjust their labor supply during this year, which would leads to bias in the empirical results. The March CPS is an annual statistical survey conducted by the U.S. BLS and the Census Bureau. They interviewed 60,000 civilian non-institutional individuals aged at least 16 years around the United States every year. The datasets includes the labor market outcomes of the previous year and various demographic variables for each sample. The following usage of the March CPS dataset closely follows Eissa and Liebman (1996) except that the years in the analysis are different and only the rotation groups 4 and 8 are extracted, in order to avoid missing values in the items that this paper uses.

To focus on single mothers, it uses the civilian single (widowed, divorced, and never married) female household heads who are 19-44 years old, who are not in school and who are not disable to work. The sample is further selected to the rotation groups 4 and 8 in the CPS sampling scheme, because a question on willingness to work ($Wantjob$), which is used to construct bounds for labor force participation in

the later section, is collected only for those groups. The resulting sample size is 15,961 in total. Table 3.7 tabulates the sample sizes by year with/without child. The summary statistics for demographic variables are in Table 3.3.

3.4 Model and specification

A number of economists have pointed out the difficulty to define the labor force participation and unemployment (for example, Jones and Riddell (1999), Brandolini et al. (2006), Kingdon and Knight (2006), and more recently Feng and Hu (2013)). This paper does not search for the "true" definitions of labor force participation and unemployment, but it takes more conservative approach of constructing intervals which covers multiple definitions of labor force participation and unemployment.

The rest of this subsection describes the construction of bounds for labor force participation and those for unemployment. To construct bounds for labor force participation, this paper combines two items, *Empstat* and $Wantjob$, from the CPS dataset. *Empstat* is the BLS definition of employment status, which takes values in {"employed", "unemployed", "not in labor force"}. $Wantjob$ asks individual with *Empstat* = "not in labor force" whether she wants regular job now, where the set of answers to this questions are {"yes", "no","maybe, it depends","do not know"} but this paper treats "maybe, it depends" and "do not know" as "no" for the simplicity. This treatment does not affect the results, because the portion of "maybe, it depends" and "do not know" in the sample is small (around 4 percent in the sample with *Empstat* = "not in labor force").

The bounds for labor force participation are based on the following two assumptions: (i) individuals with *Empstat* = "employed" or "unemployed" are in the labor

force and (ii) individuals with $Empstat$ = "not in labor force" and $Wantjob$ = "no" are out of the labor force. Denote by Y_{LF} the indicator of unobserved labor force participation where $Y_{LF} = 1$ is a participation and 0 is non-participation. This paper does not specify what is true labor force participation. The following arguments are valid for any definition of labor force participation satisfying the assumptions (i) and (ii). One bound for labor force participation is derived based on

$$Empstat = \text{"employed" or "unemployed"} \implies Y_{LF} = 1 \tag{3.1}$$

The other bound for labor force participation is derived by

$$Empstat = \text{"not in labor force" and } Wantjob = \text{"no"} \implies Y_{LF} = 0 \tag{3.2}$$

This usage of $Wantjob$ is similar to Jones and Riddell (1999) in that they use $Wantjob$ to define the marginally attached workers.

Table 3.4 summarizes the sample sizes of the categories according to $Empstat$ and $Wantjob$. Around five percentage points of the single women are categorized as out of the labor force according to the BLS but at the same time they shows that they want a regular job. The ratio goes up to more than seven percentage points for single mothers. $Wantjob$ indicates the subtlety of the definition of labor force participation, although this variable itself cannot be interpreted as labor force participation. Hurd et al. (1998) points out the measurement error of stated preferences such as "question ambiguity, subject concerns about confidentiality of sensitive information, incentives for strategic misrepresentation, imperfect knowledge of the facts, and psychometric context effects."

For the later use, the above concepts are notationally summarized as follows.

Denote

$$\bar{Y}_{LF} = \begin{cases} 1 & \text{if } Empstat = \text{“employed”} \\ 1 & \text{if } Empstat = \text{“unemployed”} \\ 1 & \text{if } Empstat = \text{“not in labor force” and } Wantjob = \text{“yes”} \\ 0 & \text{if } Empstat = \text{“not in labor force” and } Wantjob = \text{“no”} \end{cases} \tag{3.3}$$

$$Y_{LF}^{CPS} = \begin{cases} 1 & \text{if } Empstat = \text{“employed”} \\ 1 & \text{if } Empstat = \text{“unemployed”} \\ 0 & \text{if } Empstat = \text{“not in labor force” and } Wantjob = \text{“yes”} \\ 0 & \text{if } Empstat = \text{“not in labor force” and } Wantjob = \text{“no”}. \end{cases} \tag{3.4}$$

Y_{LF}^{CPS} is the indicator for the BLS definition of labor force participation, and $\bar{Y}_{LF} = 1$ implies either that an individual is in the labor force according to the BLS definition or that she is considered as a marginally attached worker as in Jones and Riddell (1999). Note that $\bar{Y}_{LF}$ is the indicator of a weaker definition than that of BLS. The equation (3.1) and (3.2) yield bounds for labor force participation as

$$Y_{LF}^{CPS} \leqslant Y_{LF} \leqslant \bar{Y}_{LF}. \tag{3.5}$$

In this setting, the variable Y_{LF} is not observed directly but it belongs to the observable interval $[Y_{LF}^{CPS}, \bar{Y}_{LF}]$. There are four possibilities for the values of the triplet $(Y_{LF}^{CPS}, Y_{LF}, \bar{Y}_{LF})$:

$$(Y_{LF}^{CPS}, Y_{LF}, \bar{Y}_{LF}) = (0,0,0),\ (0,0,1),\ (0,1,1),\ (1,1,1). \tag{3.6}$$

This paper allows for any definition of labor force participation within the interval (3.5), especially in the case with $(Y_{LF}^{CPS}, Y_{LF}, \bar{Y}_{LF}) = (0,1)$. The case with $(Y_{LF}^{CPS}, Y_{LF}, \bar{Y}_{LF}) = (0,1)$ corresponds to $Empstat =$ “*not in labor force*” but

$Wantjob$ = "*yes*", which is the case that the BLS definition of labor force participation is controversial.

The interval for unemployment is constructed in a similar fashion. Denote by Y_{UE} the indicator for unemployment where $Y_{UE} = 1$ is unemployment and $Y_{UE} = 0$ is not. Define

$$\bar{Y}_{UE} = \begin{cases} 0 & \text{if } Empstat = \text{“employed”} \\ 1 & \text{if } Empstat = \text{“unemployed”} \\ 1 & \text{if } Empstat = \text{“not in labor force” and } Wantjob = \text{“yes”} \\ 0 & \text{if } Empstat = \text{“not in labor force” and } Wantjob = \text{“no”} \end{cases} \tag{3.7}$$

$$Y_{UE}^{CPS} = \begin{cases} 0 & \text{if } Empstat = \text{“employed”} \\ 1 & \text{if } Empstat = \text{“unemployed”} \\ 0 & \text{if } Empstat = \text{“not in labor force” and } Wantjob = \text{“yes”} \\ 0 & \text{if } Empstat = \text{“not in labor force” and } Wantjob = \text{“no”}. \end{cases} \tag{3.8}$$

Then the variable Y_{UE} is not observed directly but it belongs to the observable interval

$$Y_{UE}^{CPS} \leqslant Y_{UE} \leqslant \bar{Y}_{UE}. \tag{3.9}$$

3.4.1 Difference-in-difference approach

This subsection describes the model and the parameter of interest. In order to identify and estimate the treatment effect, this papers adopts the difference-in-difference strategy for the repeated cross section. For $k = LF, UE$, the regression equation is defined as

$$Y_k = \beta_{k0} + \beta_{k1}G + \beta_{k2}T + \beta_{k3}(G \times T) + \epsilon_k \tag{3.10}$$

where $\theta_k = (\beta_{k0}, \beta_{k1}, \beta_{k2}, \beta_{k3})'$ is a parameter. G is the indicator of having at least one child, T is the indicator of whether the survey year is in 1995-97 and ϵ_k is an error

term. T indicates whether the survey is taken before or after this intervention and G indicates whether she is in the treatment group or in the control group. As in usual difference-in-difference framework, ϵ_k is assumed to be exogenous ($E[\epsilon_k \mid G, T] = 0$) and β_{k3} is the treatment effect.

The eligibility for the EITC is not directly observable in the CPS dataset. Depending on the adjusted gross income and the eligibility of children, some single mothers are not eligible to the EITC. To make the argument simple, however, this paper treats the group of single mothers as the treatment group of the 1993 EITC expansion as in Eissa and Liebman (1996).

One caveat of this approach is that there were other policy changes during the sample periods. Therefore, the coefficient of $G \times T$ includes those other changes. One support for this approach is that Meyer and Rosenbaum (2001) investigate the effect of various welfare programs on the labor supply during 1984-1996 and they conclude that the EITC is the most relevant policy change to the labor supply during that period.

3.4.2 Partial Identification

Since Y_k is not observed directly but it is in the observable interval, the model is written as

$$\begin{cases} Y_k = \beta_{k0} + \beta_{k1} G + \beta_{k2} T + \beta_{k3}(G \times T) + \epsilon_k \\ E[\epsilon_k \mid G, T] = 0 \\ Y_k^{CPS} \leqslant Y_k \leqslant \bar{Y}_k \end{cases} \tag{3.11}$$

for $k = LF, UE$. In the above model, the parameter would be point-identified in the ordinary least square if Y_k was observed. However, since Y_k is not observable,

the point identification does not hold for β_k in general but the identified set can be constructed.

This probability $Pr(Y_k \neq Y_k^{CPS} \mid G, T)$ is unknown and unidentifiable from the data, because Y_k is unobservable. I consider the upper bound on the probability that Y_k^{CPS} is misclassified:

$$Pr(Y_k \neq Y_k^{CPS} \mid G, T) \leqslant \lambda$$

where $\lambda \in [0, 1]$ is a known constant. $\lambda = 1$ means that there is no additional information from the above inequalities, and $\lambda = 0$ means that the variable Y_k^{CPS} represents the true outcome variable of interest. Molinari (2008) also uses the upper bound restrictions on the misclassification probability. Imposing additional assumption on the probability excludes some definitions of unemployment and labor force participation which satisfy the conditions (3.1) and (3.2), so it is possible to make the conclusion more likely to be decisive. Particularly I emphasize the case with $\lambda = 0.5$, in which at least 50% of the individuals with $Empstat =$ "not in labor force" and $Wantjob =$ "yes" in the dataset, are out of the labor force. Therefore, $\lambda = 0.5$ justifies the BLS definition of unemployment in that the BLS cosider the individuals with $Empstat =$ "not in labor force" and $Wantjob =$ "yes" as being out of the labor force.

The identified set for θ is characterized by the moment inequalities

$$\begin{aligned} E[1_{\{(G,T)=(g,t)\}}(\beta_{k0} + \beta_{k1}G + \beta_{k2}T + \beta_{k3}(G \times T) - Y_k^{CPS})] &\geqslant 0 \\ E[1_{\{(G,T)=(g,t)\}}((1-\lambda)Y_k^{CPS} + \lambda\bar{Y}_k - (\beta_{k0} + \beta_{k1}G + \beta_{k2}T + \beta_{k3}(G \times T)))] &\geqslant 0 \end{aligned}$$

for $g, t = 0, 1$. The size of the identified set for the parameter θ comes from the misclassification of unemployment, so the degree of under-identification depends on

the misclassification probability. I assume that the misclassification only occurs for those individuals with $Empstat =$ "not in labor force" and $Wantjob =$ "yes" in the dataset. Therefore, the degree of under-identification depends on

$$Pr(Empstat = \text{"not in labor force"}, Wantjob = \text{"yes"} \mid G, T). \tag{3.12}$$

The identified set is a singleton if the above probability is zero, and the identified set expands as the frequency increases.

3.5 Empirical results

The first part of the empirical exercise is based on BLS definitions of unemployment and labor force participation. In other words, this subsection assumes $\lambda = 0$ and regresses Y_k^{CPS} on $(G, T, G \times T)$. Tables 3.5 and 3.6 summarize the estimation results for each outcome variable. The estimated coefficient for $G \times T$ for unemployment is insignificant ,and the estimated coefficient for $G \times T$ for labor force participation is around 5%. This results can be interpreted that the 1993 EITC expansion has no effect on unemployment and its increase in labor force participation translates directly to the increase in employment. This analysis assumes $\lambda = 0$, that is, it assumes that BLS definitions of unemployment and labor force participation are correct, is one extreme approach to the ambiguity in definitions of unemployment and labor force participation. As explained in the introduction, however, these definitions are controversial in the labor literature.

The second analysis explores weaker assumption on the misclassification of unemployment, which yields the empirical results robust to the ambiguity in definitions of unemployment and labor force participation. To construct confidence intervals, I

use the generalized moment selection method in Andrews and Soares (2010). (Other methods for moment inequality models are surveyed by Canay and Shaikh (2016).) Tables 3.7 and 3.8 displays the 95% confidence interval for the difference-in-difference estimates on labor force participation and unemployment. The estimates are computed for three populations: all the samples, the high school dropouts and the individuals less than 30 years old.

Table 3.7 and Figure 3.1 show that the EITC effect on labor force participation is significantly positive even for $\lambda = 0.5$. This implies that the positive effect in Table 3.5 is robust to misclassification of unemployment. In contrast, the EITC effect on unemployment is not significantly different from zero once λ is at least 5%. This implies that the positive effect in Table 3.6 is not robust to misclassification of unemployment; a small amount of the misclassification change the empirical results such that even the sign of the effect is not statistically significant. The difference-in-difference result based on the BLS definition of unemployment can be misleading due to misclassification of unemployment.

3.6 Conclusion

There are several remaining issues in the analysis of this paper. First, this paper relies on the difference-in-difference assumption. However, it is possible that the treatment is endogenous and the difference-in-difference cannot recover the treatment effect. By applying the methodology in Manski (1995) and Pepper (2000), the endogeneity of the treatment can enlarge the identified set for the treatment effect. Second, the treatment group is single mothers in this paper. However, the EITC has different schedules even among single mothers according to income and

number of children. For example, single mothers with more than one children are the interesting treatment group, because the EITC favor this group after its 1993 expansion. Hotz et al. (2010) investigates the EITC effect on this group. Lasst, this paper uses the indicator variable for the reform in order to describe the EITC policy change. Alternatively it is possible to use the policy parameters from various welfare programs as in Meyer and Rosenbaum (2001). This change improves the analysis in two dimensions. First, it is possible to obtain treatment effects of one unit change of policy parameter instead of treatment effects of the whole change in welfare programs. Second, simultaneous analysis of multiple programs can separate the EITC effect out of the effects from other programs. One caution is that the above improvements assume that (i) the construction of the policy parameter variables is correct and (ii) that the taxpayers are aware of those parameters.

3.7 Tables

Table 3.1: EITC maximum credits (in nominal dollars)

Year	Max credits		
	No child	One child	More than one child
1990	0	953	953
1991	0	1192	1235
1992	0	1324	1384
1993	0	1434	1511
1994	306	2038	2538
1995	314	2094	3110
1996	323	2154	3556
1997	332	2210	3656
1998	341	2271	3756

Table 3.2: Sample size for single women during 1991-93 and 1995-97

Year	Single mothers Freq (Percent)	Single women without child Freq (Percent)
1991	1,688***	838***
	(18.66)	(16.34)
1992	1,689***	874***
	(18.68)	(17.04)
1993	1,622***	931***
	(17.93)	(18.15)
1995	1,494***	883***
	(16.52)	(17.21)
1996	1,286***	782***
	(14.22)	(15.24)
1997	1,265***	822***
	(13.99)	(16.02)
Total	9044	5130

Table 3.3: Summary statistics for single women during 1991-93 and 1995-97

	mean	sd	min	max
Employed	0.77	0.42	0	1
Unemployed	0.066	0.25	0	1
In the labor force	0.84	0.37	0	1
Family size	2.52	1.69	1	14
Number of preschool children	0.66	1.06	0	9
Number of preschool children	0.17	0.48	0	4
Age	30.3	7.20	19	44
High school dropout	0.20	0.40	0	1
High school graduate	0.57	0.49	0	1
College graduate	0.22	0.42	0	1
Nonwhite	0.23	0.42	0	1

Table 3.4: Employment status for single women during 1991-93 and 1995-97

Empstat	*Wantjob*	All	Single mothers
"employed"		76.6%	66.3%
"unemployed"		6.3%	7.6%
"not in labor force"	"yes"	4.2%	7.2%
"not in labor force"	"no"	12.9%	18.9%

Table 3.5: Difference-in-difference estimates for the the EITC Effect on labor force participation

	All	HS Dropouts	Under 30 yrs
G	-0.18***	-0.21***	-0.34***
	(0.0086)	(0.021)	(0.014)
T	-0.00013	-0.090***	-0.0026
	(0.0076)	(0.029)	(0.010)
$G \times T$	0.057***	0.073*	0.12***
	(0.013)	(0.038)	(0.020)
Constant	0.90***	0.76***	0.89***
	(0.0051)	(0.015)	(0.0067)
Observations	14,174	2,890	7,111
R-squared	0.045	0.042	0.102

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 3.6: Difference-in-difference estimates for the the EITC Effect on unemployment

	All	HS Dropouts	Under 30 yrs
G	0.018***	-0.0086	0.030***
	(0.0060)	(0.013)	(0.0098)
T	-0.016***	0.0013	-0.016**
	(0.0053)	(0.018)	(0.0071)
$G \times T$	0.0081	0.018	0.0023
	(0.0087)	(0.024)	(0.014)
Constant	0.065***	0.10***	0.072***
	(0.0035)	(0.0093)	(0.0047)
Observations	14,174	2,890	7,111
R-squared	0.002	0.001	0.003

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 3.7: 95% confidential intervals for the the EITC effect on labor force participation

	All		HS Dropouts		Under 30 yrs	
$\lambda = 0$	[0.035	0.073]	[0.057	0.169]	[0.058	0.128]
$\lambda = .05$	[0.034	0.076]	[0.056	0.174]	[0.058	0.131]
$\lambda = .1$	[0.032	0.079]	[0.052	0.18]	[0.054	0.136]
$\lambda = .25$	[0.027	0.088]	[0.038	0.202]	[0.047	0.15]
$\lambda = .5$	[0.016	0.106]	[0.013	0.237]	[0.035	0.175]
$\lambda = 1$	[-0.003	0.141]	[-0.039	0.308]	[0.006	0.226]

Table 3.8: 95% confidential intervals for the the EITC effect on unemployment

	All		HS Dropouts		Under 30 yrs	
$\lambda = 0$	[0.001	0.024]	[-0.013	0.054]	[-0.001	0.045]
$\lambda = .05$	[-0.002	0.026]	[-0.018	0.063]	[-0.003	0.049]
$\lambda = .1$	[-0.004	0.03]	[-0.023	0.067]	[-0.006	0.054]
$\lambda = .25$	[-0.009	0.04]	[-0.037	0.092]	[-0.013	0.068]
$\lambda = .5$	[-0.02	0.059]	[-0.063	0.126]	[-0.028	0.094]
$\lambda = 1$	[-0.04	0.096]	[-0.118	0.206]	[-0.055	0.147]

3.8 Figures

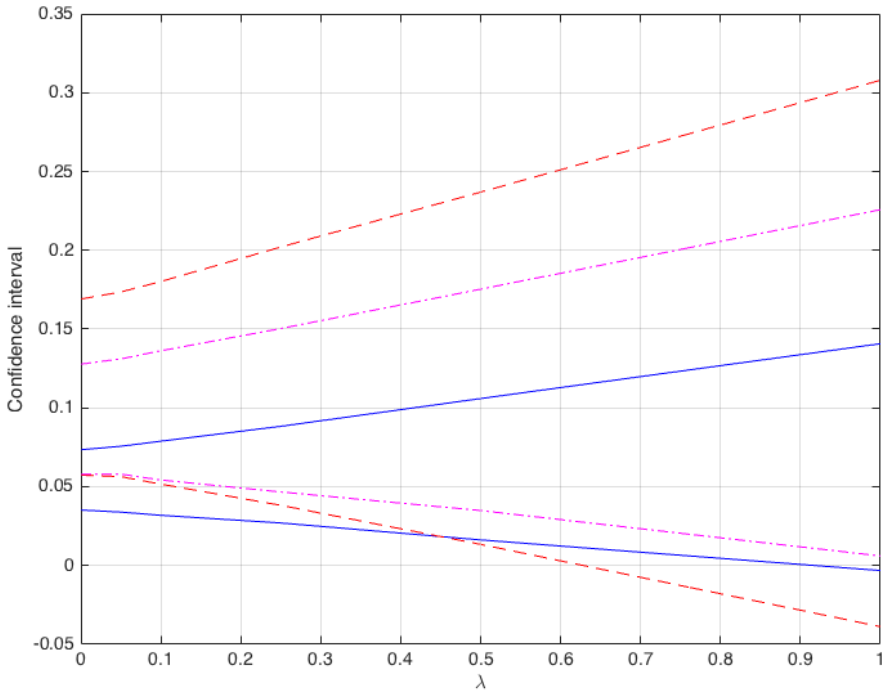

FIGURE 3.1: 95% confidence interval for the the EITC effect on labor force participation by λ. The blue solid line, —, represents the confidence interval for all the sample. The red dashed line, $--$, represents the confidence interval for the high school dropouts. The magenta dash-dotted line, $- \cdot -$, represents the confidence interval for the individuals younger than 30 years old.

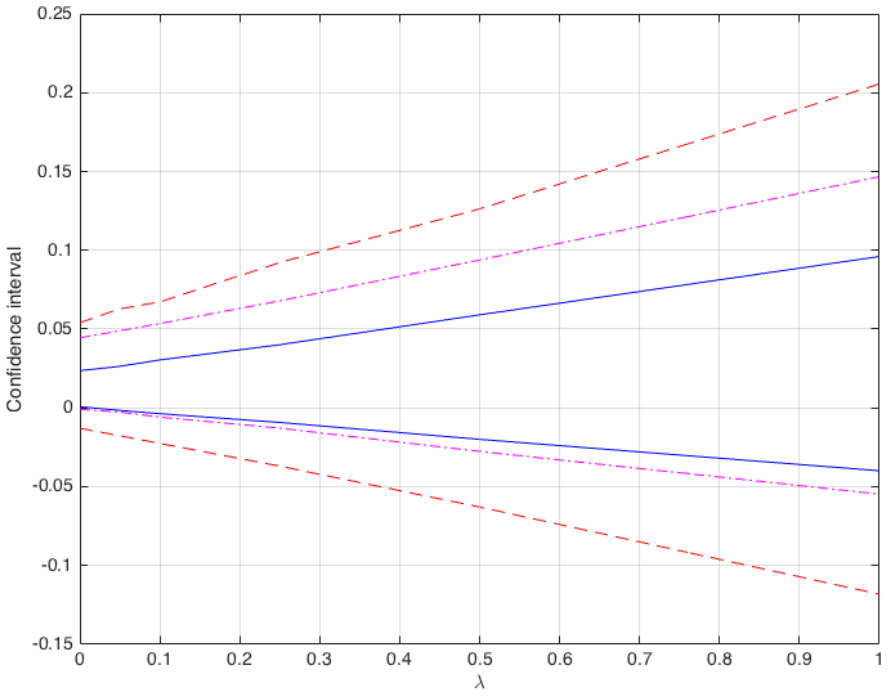

FIGURE 3.2: 95% confidence interval for the the EITC effect on unemployment by λ. The blue solid line, —, represents the confidence interval for all the sample. The red dashed line, $--$, represents the confidence interval for the high school dropouts. The magenta dash-dotted line, $-\cdot-$, represents the confidence interval for the individuals younger than 30 years old.

Bibliography

Aguirregabiria, V. and Mira, P. (2002), "Swapping the Nested Fixed Point Algorithm: A Class of Estimators for Discrete Markov Decision Models," *Econometrica*, 70, 1519–1543.

Aigner, D. J. (1973), "Regression with a Binary Independent Variable Subject to Errors of Observation," *Journal of Econometrics*, 1, 49–59.

Amemiya, Y. (1985), "Instrumental Variable Estimator for the Nonlinear Errors-in-Variables Model," *Journal of Econometrics*, 28, 273–289.

Andrews, D. W. K. and Guggenberger, P. (2009), "Hybrid and Size-Corrected Subsampling Methods," *Econometrica*, 77, 721–762.

Andrews, D. W. K. and Jia Barwick, P. (2012), "Inference for Parameters Defined by Moment Inequalities: A Recommended Moment Selection Procedure," *Econometrica*, 80, 2805–2826.

Andrews, D. W. K. and Shi, X. (2013), "Inference Based on Conditional Moment Inequalities," *Econometrica*, 81, 609–666.

Andrews, D. W. K. and Soares, G. (2010), "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection," *Econometrica*, 78, 119–157.

Angrist, J. D. and Krueger, A. B. (1999), "Empirical Strategies in Labor Economics," in *Handbook of Labor Economics*, eds. O. Ashenfelter and D. Card, vol. 3, pp. 1277–1366, North Holland, Amsterdam.

Angrist, J. D. and Krueger, A. B. (2001), "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments," *Journal of Economic Perspectives*, 15, 69–85.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996), "Identification of Causal Effects using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444–455.

Armstrong, T. B. (2014), "Weighted KS statistics for inference on conditional moment inequalities," *Journal of Econometrics*, 181, 92 – 116.

Armstrong, T. B. (2015), "Asymptotically exact inference in conditional moment inequality models," *Journal of Econometrics*, 186, 51 – 65.

Armstrong, T. B. and Chan, H. P. (2014), "Multiscale Adaptive Inference on Conditional Moment Inequalities," Working paper.

Balke, A. and Pearl, J. (1997), "Bounds on Treatment Effects from Studies with Imperfect Compliance," *Journal of the American Statistical Association*, 92, 1171–1176.

Besley, T. and Coate, S. (1992), "Understanding welfare stigma: Taxpayer resentment and statistical discrimination," *Journal of Public Economics*, 48, 165 – 183.

Black, D., Sanders, S., and Taylor, L. (2003), "Measurement of Higher Education in the Census and Current Population Survey," *Journal of the American Statistical Association*, 98, 545–554.

Black, D. A., Berger, M. C., and Scott, F. A. (2000), "Bounding Parameter Estimates with Nonclassical Measurement Error," *Journal of the American Statistical Association*, 95, 739–748.

Blackwell, D. (1965), "Discounted Dynamic Programming," *The Annals of Mathematical Statistics*, 36, 226–235.

Blundell, R., Gosling, A., Ichimura, H., and Meghir, C. (2007), "Changes in the Distribution of Male and Female Wages Accounting for Employment Composition Using Bounds," *Econometrica*, 75, 323–363.

Bollinger, C. R. (1996), "Bounding Mean Regressions When a Binary Regressor is Mismeasured," *Journal of Econometrics*, 73, 387–399.

Bound, J., Brown, C., and Mathiowetz, N. (2001), "Measurement Error in Survey Data," in *Handbook of Econometrics*, eds. J. Heckman and E. Leamer, vol. 5, chap. 59, pp. 3705–3843, Elsevier.

Brandolini, A., Cipollone, P., and Viviano, E. (2006), "Does the ILO Definition Capture All Unemployment?" *Journal of the European Economic Association*, 4, 153–179.

Bugni, F. A. (2010), "Bootstrap Inference in Partially Identified Models Defined by Moment Inequalities: Coverage of the Identified Set," *Econometrica*, 78, 735–753.

Bugni, F. A., Canay, I. A., and Guggenberger, P. (2012), "Distortions of Asymptotic Confidence Size in Locally Misspecified Moment Inequality Models," *Econometrica*, 80, 1741–1768.

Canay, I. A. (2010), "EL inference for partially identified models: Large deviations optimality and bootstrap validity," *Journal of Econometrics*, 156, 408 – 425.

Canay, I. A. and Shaikh, A. M. (2016), "Practical and Theoretical Advances in Inference for Partially Identified Models," Working paper.

Card, D. (1995), "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," in *Aspects of Labor Market Behaviour: Essays in Honour of John Vanderkamp*, eds. L. Christofides, E. Grant, and R. Swidinsky, pp. 201–222, University of Toronto Press, Toronto.

Card, D. (1999), "The Causal Effect of Education on Earnings," in *Handbook of Labor Economics*, eds. O. Ashenfelter and D. Card, vol. 3, chap. 30, pp. 1801–1863, Elsevier, Amsterdam.

Card, D. (2001), "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems," *Econometrica*, 69, 1127–1160.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2012), *Measurement Error in Nonlinear Models: A Modern Perspective*, Chapman & Hall/CRC, Boca Raton, 2nd edn.

Chalak, K. (2013), "Instrumental Variables Methods with Heterogeneity and Mismeasured Instruments," Working paper.

Chernozhukov, V., Hong, H., and Tamer, E. (2007), "Estimation and Confidence Regions for Parameter Sets in Econometric Models1," *Econometrica*, 75, 1243–1284.

Chernozhukov, V., Lee, S., and Rosen, A. M. (2013), "Intersection Bounds: Estimation and Inference," *Econometrica*, 81, 667–737.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014), "Testing Many Moment Inequalities," working paper.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2015), "Central Limit Theorems and Bootstrap in High Dimensions," working paper.

Chetverikov, D. (2013), "Adaptive Test of Conditional Moment Inequalities," Working paper.

de Chaisemartin, C. (2015), "Tolerating Defiance? Local Average Treatment Effects without Monotonicity," Working paper.

Deaton, A. (2009), "Instruments of Development: Randomisation in the Tropics, and the Search for the Elusive Keys to Economic Development," in *Proceedings of the British Academy, Volume 162, 2008 Lectures*, British Academy, Oxford.

Eissa, N. and Liebman, J. B. (1996), "Labor Supply Response to the Earned Income Tax Credit," *The Quarterly Journal of Economics*, 111, 605–637.

Fang, Z. and Santos, A. (2014), "Inference on Directionally Differentiable Functions," Working paper.

Feng, S. and Hu, Y. (2013), "Misclassification Errors and the Underestimation of the US Unemployment Rate," *American Economic Review*, 103, 1054–70.

Frazis, H. and Loewenstein, M. A. (2003), "Estimating Linear Regressions with Mismeasured, Possibly Endogenous, Binary Explanatory Variables," *Journal of Econometrics*, 117, 151–178.

Gourieroux, C. and Monfort, A. (1995), *Statistics and Econometric Models: Volume 2*, Cambridge University Press.

Griliches, Z. (1977), "Estimating the Returns to Schooling: Some Econometric Problems," *Econometrica*, 45, 1–22.

Hausman, J. A., Newey, W. K., Ichimura, H., and Powell, J. L. (1991), "Identification and Estimation of Polynomial Errors-in-Variables Models," *Journal of Econometrics*, 50, 273–295.

Hausman, J. A., Newey, W. K., and Powell, J. L. (1995), "Nonlinear Errors in Variables Estimation of Some Engel Curves," *Journal of Econometrics*, 65, 205–233.

Heckman, J. J. and Urzúa, S. (2010), "Comparing IV with Structural Models: What Simple IV Can and Cannot Identify," *Journal of Econometrics*, 156, 27–37.

Heckman, J. J. and Vytlacil, E. (2005), "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 73, 669–738.

Heckman, J. J., Schmierer, D., and Urzua, S. (2010), "Testing the Correlated Random Coefficient Model," *Journal of Econometrics*, 158, 177–203.

Henry, M., Kitamura, Y., and Salanié, B. (2014), "Partial Identification of Finite Mixtures in Econometric Models," *Quantitative Economics*, 5, 123–144.

Henry, M., Jochmans, K., and Salanié, B. (2015), "Inference on Two-Component Mixtures under Tail Restrictions," *Econometric Theory*, forthcoming.

Hirano, K. and Porter, J. R. (2012), "Impossibility Results for Nondifferentiable Functionals," *Econometrica*, 80, 1769–1790.

Hotz, V. J. and Miller, R. A. (1993), "Conditional Choice Probabilities and the Estimation of Dynamic Models," *The Review of Economic Studies*, 60, pp. 497–529.

Hotz, V. J. and Scholz, J. K. (2003), "The Earned Income Tax Credit," in *Means-Tested Transfer Programs in the United States*, ed. R. A. Moffitt, chap. 3, pp. 141–198, University of Chicago Press.

Hotz, V. J., Miller, R. A., Sanders, S., and Smith, J. (1994), "A Simulation Estimator for Dynamic Models of Discrete Choice," *The Review of Economic Studies*, 61, 265–289.

Hotz, V. J., Mullin, C. H., and Scholz, J. K. (2010), "Examining the Effect of the Earned Income Tax Credit on the Labor Market Participation of Families on Welfare," Working paper.

Hsiao, C. (1989), "Consistent estimation for some nonlinear errors-in-variables models," *Journal of Econometrics*, 41, 159 – 185.

Hu, Y. (2008), "Identification and Estimation of Nonlinear Models with Misclassification Error using Instrumental Variables: A General Solution," *Journal of Econometrics*, 144, 27–61.

Hu, Y., Shiu, J.-L., and Woutersen, T. (2015), "Identification and Estimation of Single Index Models with Measurement Error and Endogeneity," *Econometrics Journal*, forthcoming.

Huber, M. and Mellace, G. (2015), "Testing Instrument Validity for LATE Identification Based on Inequality Moment Constraints," *Review of Economics and Statistics*, 97, 398–411.

Hurd, M. D., McFadden, D., Chand, H., Gan, L., Menill, A., and Roberts, M. (1998), "Consumption and Savings Balances of the Elderly: Experimental Evidence on Survey Response Bias," in *Frontiers in the Economics of Aging*, ed. D. A. Wise, pp. 353–392, University of Chicago Press, Chicago.

Imai, K. and Yamamoto, T. (2010), "Causal Inference with Differential Measurement Error: Nonparametric Identification and Sensitivity Analysis," *American Journal of Political Science*, 54, 543–560.

Imbens, G. W. (2010), "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)," *Journal of Economic Literature*, 48, 399–423.

Imbens, G. W. (2014), "Instrumental Variables: An Econometrician's Perspective," *Statistical Science*, 29, 323–358.

Imbens, G. W. and Angrist, J. D. (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–75.

Imbens, G. W. and Manski, C. F. (2004), "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 72, 1845–1857.

Jones, S. R. G. and Riddell, W. C. (1999), "The Measurement of Unemployment: An Empirical Approach," *Econometrica*, 67, 147–161.

Kaestner, R., Joyce, T., and Wehbeh, H. (1996), "The Effect of Maternal Drug Use on Birth Weight: Measurement Error in Binary Variables," *Economic Inquiry*, 34, 617–629.

Kane, T. J. and Rouse, C. E. (1995), "Labor-Market Returns to Two- and Four-Year College," *American Economic Review*, 85, 600–614.

Kane, T. J., Rouse, C. E., and Staiger, D. (1999), "Estimating Returns to Schooling When Schooling is Misreported," NBER Working Paper No. 7235.

Kim, K. I. (2009), "Set Estimation and Inference with Models Characterized by Conditional Moment Inequalities," Working paper.

Kingdon, G. and Knight, J. (2006), "The Measurement of Unemployment When Unemployment Is High," *Labour Economics*, 13, 291–315.

Kitagawa, T. (2010), "Testing for Instrument Independence in the Selection Model," working paper.

Kitagawa, T. (2014), "A Test for Instrument Validity," *Econometrica*, forthcoming.

Kreider, B. and Pepper, J. V. (2007), "Disability and Employment: Reevaluating the Evidence in Light of Reporting Errors," *Journal of the American Statistical Association*, 102, 432–441.

Kreider, B., Pepper, J. V., Gundersen, C., and Jolliffe, D. (2012), "Identifying the Effects of SNAP (Food Stamps) on Child Health Outcomes when Participation is Endogenous and Misreported," *Journal of the American Statistical Association*, 107, 958–975.

Lewbel, A. (1998), "Semiparametric Latent Variable Model Estimation with Endogenous or Mismeasured Regressors," *Econometrica*, 66, 105–121.

Lewbel, A. (2007), "Estimation of Average Treatment Effects with Misclassification," *Econometrica*, 75, 537–551.

Magnac, T. and Thesmar, D. (2002), "Identifying Dynamic Discrete Decision Processes," *Econometrica*, 70, 801–816.

Mahajan, A. (2006), "Identification and Estimation of Regression Models with Misclassification," *Econometrica*, 74, 631–665.

Manski, C. F. (1995), *Identification Problems in the Social Sciences*, Harvard University Press, Cambridge, MA.

Manski, C. F. (2003), *Partial Identification of Probability Distributions*, Springer-Verlag, New York.

Manski, C. F. and Tamer, E. (2002), "Inference on Regressions with Interval Data on a Regressor or Outcome," *Econometrica*, 70, 519–546.

McFadden, D. and Newey, W. K. (1994), "Large sample estimation and hypothesis testing," in *Handbook of Econometrics*, eds. R. F. Engle and D. L. McFadden, vol. 4 of *Handbook of Econometrics*, pp. 2111–2245, Elsevier.

Menzel, K. (2014), "Consistent Estimation with Many Moment Inequalities," *Journal of Econometrics*, 182, 329–350.

Meyer, B. D. and Rosenbaum, D. T. (2001), "Welfare, the Earned Income Tax Credit, and the Labor Supply of Single Mothers," *The Quarterly Journal of Economics*, 116, 1063–1114.

Moffitt, R. (1983), "An Economic Model of Welfare Stigma," *American Economic Review*, 73, 1023–1035.

Molinari, F. (2008), "Partial Identification of Probability Distributions with Misclassified Data," *Journal of Econometrics*, 144, 81–117.

Mourifié, I. Y. and Wan, Y. (2014), "Testing Local Average Treatment Effect Assumptions," working paper.

Newey, W. K. (1985a), "Generalized method of moments specification testing," *Journal of Econometrics*, 29, 229–256.

Newey, W. K. (1985b), "Maximum Likelihood Specification Testing and Conditional Moment Tests," *Econometrica*, 53, 1047–1070.

Pepper, J. V. (2000), "The Intergenerational Transmission Of Welfare Receipt: A Nonparametric Bounds Analysis," *The Review of Economics and Statistics*, 82, 472–488.

Pesendorfer, M. and Schmidt-Dengler, P. (2008), "Asymptotic Least Squares Estimators for Dynamic Games," *Review of Economic Studies*, 75, pp. 901–928.

Ponomareva, M. (2010), "Inference in Models Defined by Conditional Moment Inequalities with Contin- uous Covariates," Working paper.

Romano, J. P. and Shaikh, A. M. (2008), "Inference for identifiable parameters in partially identified econometric models," *Journal of Statistical Planning and Inference*, 138, 2786–2807.

Romano, J. P. and Shaikh, A. M. (2010), "Inference for the Identified Set in Partially Identified Econometric Models," *Econometrica*, 78, 169–211.

Rosen, A. M. (2008), "Confidence Sets for Partially Identified Parameters That Satisfy a Finite Number of Moment Inequalities," *Journal of Econometrics*, 146, 107–117.

Rothenberg, T. J. (1971), "Identification in Parametric Models," *Econometrica*, 39, 577–591.

Royalty, A. B. (1996), "The effects of job turnover on the training of men and women," *Industrial & Labor Relations Review*, 49, 506–521.

Royden, H. L. (1988), *Real Analysis*, Prentice-Hall.

Rust, J. (1987), "Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher," *Econometrica*, 55, 999–1033.

Rust, J. (1988), "Maximum Likelihood Estimation of Discrete Control Processes," *SIAM Journal on Control and Optimization*, 26, 1006–1024.

Schennach, S. M. (2013), "Measurement Error in Nonlinear Models - A Review," in *Advances in Economics and Econometrics: Economic theory*, eds. D. Acemoglu, M. Arellano, and E. Dekel, vol. 3, pp. 296–337, Cambridge University Press.

Schorfheide, F. (2005), "VAR Forecasting under Misspecification," *Journal of Econometrics*, 128, 99–136.

Seber, G. A. F. (2008), *A Matrix Handbook for Statisticians*, John Wiley and Sons, Inc., Hoboken.

Song, S. (2015), "Semiparametric Estimation of Models with Conditional Moment Restrictions in the Presence of Nonclassical Measurement Errors," *Journal of Econometrics*, 185, 95–109.

Song, S., Schennach, S. M., and White, H. (2015), "Estimating Nonseparable Models with Mismeasured Endogenous Variablesonseparable models with mismeasured endogenous variables," *Quantitative Economics*, 6, 749–794.

Stoye, J. (2009), "More on Confidence Intervals for Partially Identified Parameters," *Econometrica*, 77, 1299–1315.

White, H. (1982), "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 681–700.

White, H. (1996), *Estimation, Inference and Specification Analysis*, no. 22 in Econometric Society Monographs, Cambridge university press, Cambridge, U.K.

Biography

Takuya Ura was born on July 31, 1985, in Yokohama, Japan. He graduated from Keio University in 2008 with a B.A. in Economics and from the University of Tokyo in 2011 with a M.A. in Economics. He is earning a Ph.D. in Economics from Duke University in 2016. He is going to start his academic career as an Assistant Professor of Economics at the University of California, Davis.