Decentralization estimators for instrumental variable quantile regression models

Hiroaki Kaido
Kaspar Wüthrich

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP72/18

DECENTRALIZATION ESTIMATORS FOR INSTRUMENTAL VARIABLE QUANTILE REGRESSION MODELS

HIROAKI KAIDO* AND KASPAR WÜTHRICH†

Abstract. The instrumental variable quantile regression (IVQR) model of Chernozhukov and Hansen (2005, 2006) is a flexible and powerful tool for evaluating the impact of endogenous covariates on the whole distribution of the outcome of interest. Estimation, however, is computationally burdensome because the GMM objective function is non-smooth and non-convex. This paper shows that the IVQR estimation problem can be decomposed into a set of conventional quantile regression sub-problems, which are convex and can be solved efficiently. This allows for reformulating the original estimation problem as the problem of finding the fixed point of a low dimensional map. This reformulation leads to new identification results and, most importantly, to practical, easy to implement, and computationally tractable estimators. We explore estimation algorithms based on the contraction mapping theorem and algorithms based on root-finding methods. We prove consistency and asymptotic normality of our estimators and establish the validity of a bootstrap procedure for estimating the limiting laws. Monte Carlo simulations support the estimator's enhanced computational tractability and demonstrate desirable finite sample properties.

Keywords: instrumental variables, quantile regression, contraction mapping, fixed point estimator, bootstrap.

Date: December 28, 2018. We are grateful to Alfred Galichon for many helpful discussions and valuable inputs. We would like to thank Victor Chernozhukov, Patrik Guggenberger, Hidehiko Ichimura, Yuichiro Kamada, Hiroyuki Kasahara, Kengo Kato, Ivana Komunjer, Klaus Neusser, Quang Vuong (discussant) and conference and seminar participants at Boston College, Penn State, the University of Arizona, the University of Bern, UCLA, UCR, UCSB, UCSD, UNC, and the Workshop for Advances in Microeconometrics for helpful comments. This project was started when both authors were visiting the MIT in 2015.

1. Introduction

Quantile regression (QR), introduced by Koenker and Bassett (1978), is a very popular method for estimating the effect of regressors on the whole outcome distribution. QR is flexible, easy to interpret, and can be computed very efficiently as the solution to a convex problem. However, in many applications, the variables of interest are endogenous, rendering QR inconsistent for estimating causal quantile effects. The instrumental variable quantile regression (IVQR) model of Chernozhukov and Hansen (2004, 2005, 2006) generalizes QR to accommodate endogenous regressors. Unfortunately, in sharp contrast to QR, estimation of IVQR models is computationally burdensome because the resulting estimation problem, formulated as a generalized method of moments (GMM) problem, is non-smooth and non-convex, even for linear models. From an applied perspective, this issue is particularly troublesome since resampling methods are often used to avoid the choice of tuning parameters when estimating the asymptotic variance of the estimators.

In this paper, we propose a new class of estimators for linear IVQR models. The suggested estimators are computationally tractable, very easy to implement, and particularly suitable for settings with many exogenous, a moderate number of endogenous regressors and a large number of observations, which are ubiquitous in applied research. The key insight underlying our estimators is that the IVQR estimation problem can be decomposed into a series of (weighted) conventional QR problems, which are convex and can be solved very quickly using robust algorithms. The IVQR estimator is then characterized as a fixed point of such sub-problems. Computationally, this reformulation allows us to recast the original non-smooth and non-convex optimization problem as the problem of finding the fixed point of a low dimensional map, which leads to substantial reductions in computation times. Implementation of our preferred procedures is straightforward and only requires the availability of a routine for estimating quantile regressions and in some cases a univariate root-finder. The resulting estimation algorithms attain significant computational gains. For example, we show that in problems with two endogenous variables, a version of our estimator that uses a contraction algorithm is 110–215 times faster than the popular inverse quantile regression (IQR) estimator of Chernozhukov and Hansen (2006) with a grid search over 100×100 points. Another version that uses a nested root-finding algorithm, which is guaranteed to converge under a milder condition, is 70–125 times faster than the IQR estimator. Importantly, these computational gains do not come at a cost in terms of the finite sample performance of our procedures, which is very similar to inverse quantile regression.

The fixed point reformulation also provides a new insight into global identification in the IVQR model. In particular, it allows us to study identification and stability of the algorithms (at the population level) in the same framework. Exploiting the equivalence of global identification and uniqueness of the fixed point, we give a new identification result and population algorithms based on the contraction mapping theorem. We then compare our identification conditions to those of Chernozhukov and Hansen (2006). Further, our reformulation is shown to be useful beyond setups where the contraction mapping theorem applies as long as the parameter of interest is globally identified. For such settings, algorithms based on root-finding methods are proposed. Finally, we show that, by recursively nesting fixed point problems, it is always possible to recast the IVQR estimation problem as a univariate root-finding problem, which is particularly easy to solve.

We establish consistency and asymptotic normality of the proposed estimators. In addition, we prove validity of a bootstrap procedure for consistently estimating the limiting laws. We emphasize that the bootstrap is particularly attractive in conjunction with our efficient estimation algorithms, as it allows us to avoid the choice of tuning parameters inherent to estimating the asymptotic variance based on analytic formulas. The key technical ingredient for deriving our theoretical results is the Hadamard differentiability of the fixed point map. This result may be of independent interest.

To illustrate the usefulness of our estimation algorithms, we revisit the analysis of the impact of 401(k) plans on savings in Chernozhukov and Hansen (2004). Based on this application, we perform extensive Monte Carlo simulations, which demonstrate that our estimation and inference procedures have excellent small sample properties.

1.1. **Literature.**

We contribute to the literature on estimation and inference based on linear IVQR models. Chernozhukov and Hong (2003) have proposed a quasi-Bayesian approach. This approach can accommodate multiple endogenous variables but requires careful tuning in applications, as noted by Chernozhukov and Hansen (2013). Chernozhukov and Hansen (2006) have proposed an inverse QR algorithm that combines grid search with convex QR problems. Because the dimensionality of the grid search equals the number of endogenous variables, this approach is computationally feasible essentially only if the number of endogenous variables is very low. Chernozhukov and Hansen (2008) and Jun (2008) have studied weak instrument robust inference procedures based on the inversion of Anderson-Rubin-type tests. Chernozhukov, Hansen, and Jansson (2009) have proposed a finite sample inference approach. Andrews and Mikusheva (2016) have developed a general conditional inference approach and derived sufficient conditions for the IVQR model. Kaplan and Sun (2017) and de Castro, Galvao, Kaplan, and Liu (2018) have suggested to use smoothed estimating equations to overcome the non-smoothness of the IVQR estimation problem, although the non-convexity remains. More recently, Chen and Lee (2018) have proposed to reformulate the IVQR problem as a mixed-integer quadratic programming problem which can be solved using well-established algorithms. However, efficiently solving such a problem is still challenging even for low-dimensional settings. By replacing the ℓ_2 norm by the ℓ_∞ norm, Zhu (2018) has shown that the problem admits a reformulation as a mixed-integer linear programming problem, which can be computed much more efficiently. In addition, Zhu (2018) has proposed a k-step approach that allows for estimating models with multiple endogenous regressors based on large datasets.

Compared to existing literature, the main advantages of the proposed estimation algorithms are the following. By relying on convex QR problems, our estimators are easy to implement, robust, and computationally efficient in settings with many exogenous variables, a moderate number of endogenous variables, and a large number of observations. In addition, by exploiting the specific structure of the IVQR estimation problem, our estimators are tuning-free and do not require the availability of high-level optimization routines.

Semi- and nonparametric estimation of IVQR models has been studied by Chernozhukov, Imbens, and Newey (2007), Horowitz and Lee (2007), Chen and Pouzo (2009), Chen and Pouzo (2012), Gagliardini and Scaillet (2012) and Wüthrich (2017). Chernozhukov and Hansen (2013) and Chernozhukov, Hansen, and Wüthrich (2017) have provided surveys of the IVQR model including references to empirical applications.

Abadie, Angrist, and Imbens (2002) have proposed an alternative approach to the identification and estimation of quantile effects with binary endogenous regressors, which builds on the local average treatment effects framework of Imbens and Angrist (1994). Their approach has been extended and further developed by Frandsen, Frölich, and Melly (2012), Frölich and Melly (2013), and Belloni, Chernozhukov, Fernandez-Val, and Hansen (2017) among others. We refer to Melly and Wüthrich (2017) for a recent review of this approach and to Wüthrich (2018) for a comparison between this approach and the IVQR model. Identification and estimation in nonseparable models with continuous endogenous regressors have been studied by Chesher (2003), Ma and Koenker (2006), Lee (2007), Jun (2009), Imbens and Newey (2009), D'Haultfoeuille and Février (2015), and Torgovitsky (2015) among others.

On a broader level, our paper contributes to the literature which proposes estimation procedures that rely on decomposing computationally burdensome estimation problem into several more tractable subproblems. This type of procedure, which we call *decentralization*, has been applied in many different contexts. Examples include the estimation of single index models with unknown link function (Weisberg and Welsh, 1994), general maximum likelihood problems (Smyth, 1996), linear models with high-dimensional fixed effects (e.g., Guimaraes and Portugal, 2010, and the references therein), sample selection models (Marra and Radice, 2013), peer effects models (Arcidiacono, Foster, Goodpaster, and Kinsler, 2012), interactive fixed effects models (e.g., Chen, Fernandez-Val, and Weidner, 2014; Moon and Weidner, 2015), and random coefficient logit demand models (Lee and Seo, 2015). Most of these papers decompose a single estimation problem into two subproblems. This paper explicitly considers cases in which the number of subproblems may exceed two. Our analysis on identification, estimation, and inference can be extended beyond the IVQR model and is undertaken in ongoing work.

1.2. **Organization of the Paper.**

The remainder of the paper is structured as follows. Section 2 introduces the setup and the IVQR model. Section 3 shows that the IVQR estimation problem can be decentralized into a series of (weighted) conventional QR problems. In Section 4 we introduce population algorithms based on the contraction mapping theorem and root-finders. Section 5 discusses the corresponding sample algorithms. In Section 6 we establish the asymptotic normality of our estimators and prove the validity of the bootstrap. Section 7 presents an empirical application. In Section 8 we provide extensive simulation evidence on the finite sample properties of our methods. Section 9 concludes. All proofs and some additional results are collected in the appendix.

2. Setup and Model

Consider a setup with a continuous outcome variable Y, a $d_X \times 1$ vector of exogenous covariates X, a $d_D \times 1$ vector of endogenous treatment variables D, and a $d_Z \times 1$ vector of instruments Z. The IVQR model is developed within the standard potential outcomes framework (e.g., Rubin, 1974). Let $\{Y_d\}$ denote the (latent) potential outcomes. The object of primary interest is the conditional quantile function of the potential outcomes, which we denote by $q(d, x, \tau)$. Having conditioned on covariates $X = x$, by the Skorokhod representation of random variables, potential outcomes can be represented as

$$Y_d = q(d, x, U_d) \text{ with } U_d \sim U(0, 1).$$

This representation lies at the heart of the IVQR model. With this notation at hand, we state the main conditions of the IVQR model (Chernozhukov and Hansen, 2005, Assumptions A1-A5).

Assumption 1. *Given a common probability space (Ω, F, P), the following conditions hold jointly with probability one:*

(1) Potential outcomes: Conditional on $X = x$, for each d, $Y_d = q(d, x, U_d)$, where $q(d, x, \tau)$ is strictly increasing in τ and $U_d \sim U(0, 1)$.

(2) Independence: Conditional on $X = x$, $\{U_d\}$ are independent of Z.

(3) Selection: $D := \delta(Z, X, V)$ for some unknown function $\delta(\cdot)$ and random vector V.

(4) Rank invariance or Rank similarity: Conditional on $X = x$, $Z = z$,

(a) $\{U_d\}$ are equal to each other; or, more generally,

(b) $\{U_d\}$ are identically distributed, conditional on V.

(5) Observed variables: Observed variables consist of $Y := q(D, X, U_D)$, D, X, and Z.

We briefly discuss the most important aspects of Assumption 1 and refer the interested reader to Chernozhukov and Hansen (2005, 2006, 2013) for more comprehensive treatments. Assumption 1.1 states the Skorohod representation of Y_d and requires strict monotonicity of the potential outcome quantile function, which rules out discrete outcomes. Assumption 1.2 imposes independence between the potential outcomes and the instrument. Assumption 1.3 defines a general selection mechanism. The key restriction of the IVQR model is Assumption 1.4. Rank invariance (a) requires individual ranks U_d to be the same across treatment states. Rank similarity (b) weakens this condition, allowing for random slippages of U_d away from a common level U. Finally, Assumption 1.5 summarizes the observables.

The main implication of Assumption 1 is the following conditional moment restriction (Chernozhukov and Hansen, 2005, Theorem 1):

$$P\left(Y \leq q(D, X, \tau) \mid X, Z\right) = \tau, \quad \tau \in (0, 1). \tag{2.1}$$

In this paper, we focus on the commonly used linear-in-parameter model for $q(\cdot)$ (e.g., Chernozhukov and Hansen, 2006):

$$q(d, x, \tau) = x'\theta_1(\tau) + d_1\theta_2(\tau) + \cdots + d_{d_D}\theta_J(\tau), \tag{2.2}$$

where $J = d_D+1$, and $\theta(\tau) := (\theta_1(\tau)', \theta_2(\tau), \ldots, \theta_J(\tau))'$ is the finite dimensional parameter vector of interest. The conditional moment restriction (2.1) suggests GMM estimators based on the following unconditional population moment conditions:

$$\Psi_P\left(\theta(\tau)\right) := E_P\left[\left(1\left\{Y \leq X'\theta_1(\tau) + D_1\theta_2(\tau) + \cdots + D_{d_D}\theta_J(\tau)\right\} - \tau\right)\begin{pmatrix} X \\ Z \end{pmatrix}\right]. \tag{2.3}$$

Our primary goal is to obtain an estimator of θ^* in a computationally efficient and reliable manner. We therefore focus on just-identified moment restrictions where $d_Z = d_D$, for which the construction of an estimator is straightforward. A potential caveat of this approach is that estimators based on these restrictions do not achieve the pointwise (in τ) semiparametric efficiency bound implied by the conditional moment restrictions (2.1). Appendix A provides a discussion of overidentified GMM problems and presents a two-step approach, in which one obtains an initial estimator of the true parameter value $\theta^*(\tau)$ based on the just-identified moment restrictions. This initial estimator can then be used to construct a vector of optimal instruments.

For later use, we define

$$\Psi_P\left(\theta(\tau)\right) = \left(\Psi_{P,1}\left(\theta(\tau)\right)', \ldots, \Psi_{P,J}\left(\theta(\tau)\right)\right)',$$

where

$$\begin{aligned}\Psi_{P,1}\left(\theta(\tau)\right) &:= E_P\left[\left(1\left\{Y \leq X'\theta_1(\tau) + D_1\theta_2(\tau) + \cdots + D_{d_D}\theta_J(\tau)\right\} - \tau\right) X\right], \\ \Psi_{P,j}\left(\theta(\tau)\right) &:= E_P\left[\left(1\left\{Y \leq X'\theta_1(\tau) + D_1\theta_2(\tau) + \cdots + D_{d_D}\theta_J(\tau)\right\} - \tau\right) Z_{j-1}\right], \quad j = 2, \ldots, J.\end{aligned}$$

In what follows, we will often suppress the dependence on τ to lighten-up the exposition. We then define the true parameter value θ^* as the solution to the moment conditions, i.e.,

$$\Psi_P\left(\theta^*\right) = 0.$$

The resulting GMM objective function reads

$$Q_N\left(\theta\right) = -\frac{1}{2}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} m_i\left(\theta\right)\right)' W_N\left(\theta\right)\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} m_i\left(\theta\right)\right), \tag{2.4}$$

where $m_i\left(\theta\right) := \left(1\left\{Y_i \leq X_i'\theta_1 + D_{1i}\theta_2 + \cdots + D_{d_D i}\theta_J\right\} - \tau\right)\left(Z_i', X_i'\right)'$ and $W_N\left(\theta\right)$ is a positive definite weighting matrix. Estimation based on (2.4) is complicated by the non-smoothness and, most importantly, the non-convexity of $Q_N\left(\theta\right)$. The goal of this paper is to propose a new set of algorithms to address these challenges.

3. Decentralization

Here, we describe the basic idea behind our decentralization estimators. To simplify the exposition, we first illustrate our approach with the population problem of finding the true parameter value θ^* in the IVQR model. Our estimator then adopts the analogy principle, which will be presented in Section 5. The key insight is that the complicated nonlinear IVQR estimation problem can be "decentralized", i.e., decomposed into a set of more tractable sub-problems, each of which is solved by a "player" who best responds to other players' actions. Specifically, we first split the parameter vector θ into J subvectors $\theta_1, \ldots, \theta_J$. We then decompose the grand estimation problem into J subproblems. Each of the subproblems is allocated to a distinct player. For each j, player j's choice variable is the j-th subvector θ_j. Her problem is to find the value of θ_j such that a subset of the moment restrictions is satisfied given the other players' actions θ_{-j}. This reformulation allows us to view the estimation problem as a game of complete information and to characterize θ^* as the game's pure strategy Nash equilibrium.

We start with defining weighted population QR objective functions. For each $\theta \in \mathbb{R}^d$, let

$$\begin{aligned} Q_{P,1}(\theta) &:= E_P\left[\rho_\tau(Y - X'\theta_1 - D_1\theta_2 - \cdots - D_{d_D}\theta_J)\right], & (3.1)\\ Q_{P,j}(\theta) &:= E_P\left[\rho_\tau(Y - X'\theta_1 - D_1\theta_2 - \cdots - D_{d_D}\theta_J)(Z_{j-1}/D_{j-1})\right], \quad j = 2, \dots, J, & (3.2)\end{aligned}$$

where $\rho_\tau(u) = u(\tau - 1\{u < 0\})$ is the "check-function". We assume that the model is parametrized such that Z_ℓ/D_ℓ is positive for all $\ell = 1, \dots, d_D$. Under our assumptions, we can always reparametrize the model such that this condition is met; see Appendix B for more details.

Consider the following functions[1]

$$\begin{aligned} L_1(\theta_{-1}) &= \arg\min_{\tilde\theta_1 \in \mathbb{R}^{d_X}} Q_{P,1}\left(\tilde\theta_1, \theta_{-1}\right), & (3.3)\\ L_j(\theta_{-j}) &= \arg\min_{\tilde\theta_j \in \mathbb{R}} Q_{P,j}\left(\tilde\theta_j, \theta_{-j}\right), \quad j = 2, \dots, J. & (3.4)\end{aligned}$$

Borrowing the terminology from game theory, we refer to these functions *best response (BR) functions*. Observe that each player's problem is a weighted QR problem, which is convex in its choice variable. For the sample analogues of these problems, fast solution algorithms exist (Portnoy and Koenker, 1997). Under the conditions we specify below, the BR functions satisfy

$$\begin{aligned} 0 &= E_P\left[\left(1\left\{Y \le X'L_1(\theta_{-1}) + D'\theta_{-1}\right\} - \tau\right) X\right], & (3.5)\\ 0 &= E_P\left[\left(1\left\{Y \le (X', D'_{-(j-1)})'\theta_{-j} + D_{j-1}L_j(\theta_{-j})\right\} - \tau\right) Z_{j-1}\right], \quad j = 2, \dots, J, & (3.6)\end{aligned}$$

where $D_{-(j-1)}$ stacks as a vector all endogenous variables except D_{j-1}. Note that these are the unconditional IVQR moment conditions imposed on the true parameter value θ^*. Hence, θ^* satisfies

$$\theta_j^* = L_j(\theta_{-j}^*), \quad j = 1, \dots, J, \tag{3.7}$$

which implies that θ^* is a fixed point of the BR-maps (i.e. a Nash equilibrium of the game).

[1] Lemma 1 below ensures that these functions are well-defined on suitable domains.

We say that the IVQR estimation problem admits *decentralization* if the BR functions L_j, $j = 1, \dots, J$, are well-defined over domains for which the moment conditions can be evaluated.[2] To ensure decentralization, we make the following assumption.

Assumption 2. *The following conditions hold.*

(1) Θ *is a closed rectangle in* $\mathbb{R}^d$. θ^* *is in the interior of* Θ.

(2) $E[|Z_\ell|^2] < \infty$ *for* $\ell = 1, \dots, d_D$. $E[|X_k|^2] < \infty$ *for all* $k = 1, \dots, d_X$. *For each* $\ell = 1, \dots, d_D$, D_ℓ *has a compact support;*

(3) The conditional cdf $y \to F_{Y|D,X,Z}(y)$ *is continuously differentiable for all* $y \in \mathbb{R}$ *a.s. The conditional density* $f_{Y|D,Z,X}$ *is uniformly bounded a.s.;*

(4) For any $\theta \in \Theta$, *the matrices*

$$E_P[f_{Y|D,X,Z}\left(D'\theta_{-1} + X'\theta_1\right) XX']$$

and

$$E_P[f_{Y|D,X,Z}\left(D'\theta_{-1} + X'\theta_1\right) D_\ell Z_\ell], \quad \ell = 1, \dots, d_D,$$

are positive definite.

For each j, let $\Theta_{-j} \subset \mathbb{R}^{d-j}$ denote the parameter space for θ_{-j}. Assumption 2.1 ensures that Θ is compact. This assumption also ensures that each Θ_{-j} is also a closed rectangle, which we use to show that L_j is well-defined on a suitable domain. Assumption 2.2 and Assumption 2.3 impose standard regularity conditions on the conditional density and moments of the variables in the model. We assume D_ℓ has a compact support, which allows us to always reparameterize the model so that the objective function in (3.2) is well-defined and convex (cf. Appendix B). The first part of Assumption 2.4 is a standard full rank condition which is a natural extension of the local full rank condition required for local identification and decentralization (cf. Assumption 4 in the appendix). For the second part of Assumption 2.4, it suffices that the model is parametrized such that, for each $\ell \in \{1, \dots, d_D\}$, $D_\ell Z_\ell$ (and Z_ℓ / D_ℓ) is positive with probability 1.

For each j, define

$$R_{-j} := \{\theta_{-j} \in \Theta_{-j} : \Psi_{P,j}(\theta) = 0, \text{ for some } \theta = (\theta_j, \theta_{-j}) \in \Theta\}. \tag{3.8}$$

[2] In Appendix C.2, we also provide weaker conditions under which the decentralization holds on a neighborhood of θ^*. We call such a result *local decentralization*, which is sufficient for analyzing the (local) asymptotic behavior of the estimator.

This is the set of subvectors θ_{-j} for which one can find $\theta_j \in \Theta_j$ such that $\theta = (\theta_j, \theta_{-j})'$ solve the j-th moment restriction. We take this set as the domain of player j's best response map L_j.

The following lemma establishes that the IVQR model admits decentralization.

Lemma 1. *Suppose that Assumptions 1 and 2 hold. Then, there exist maps $L_j : R_{-j} \to \mathbb{R}^{d_j}, j = 1, \dots, J$ such that, for $j = 1, \dots, J$,*

$$\Psi_{P,j}\left(L_j(\theta_{-j}), \theta_{-j}\right) = 0, \quad \textit{for all } \theta_{-j} \in R_{-j}. \tag{3.9}$$

Further, L_j is continuously differentiable on the interior of R_{-j} for all $j = 1, \dots, J$.

We now introduce maps that represent all players' (joint) best responses. We consider two basic choices of such maps; one represents simultaneous responses, and the other represents sequential responses. In what follows, for any subset $a \subset \{1, \dots, J\}$, let θ_{-a} denote the subvector of θ that stacks the components of θ_j's for all $j \notin a$. If a is a singleton (i.e. $a = \{j\}$ for some j), we simply write θ_{-j}. For each j and $a \subseteq \{1, \dots, J\} \setminus \{j\}$, let $\pi_{-a} : \Theta_{-j} \to \prod_{k \in \{1,\dots,J\} \setminus (\{j\} \cup a)} \Theta_k$ be the coordinate projection of θ_{-j} to a further subvector that stacks all components of θ_{-j} except for those of θ_k with $k \in a$.

Let $D_K := \{\theta \in \Theta : \pi_{-j}\theta \in R_{-j},\ j = 1, \dots, J\}$. Let $K : D_K \to \mathbb{R}^d$ be a map defined by

$$K(\theta) = \begin{pmatrix} K_1(\theta) \\ \vdots \\ K_J(\theta) \end{pmatrix} = \begin{pmatrix} L_1(\theta_{-1}) \\ \vdots \\ L_J(\theta_{-J}) \end{pmatrix}. \tag{3.10}$$

This can be interpreted as the players' simultaneous best responses to the initial strategy $(\theta_1, \dots, \theta_J)$. With one endogenous variable, this map simplifies to

$$K(\theta) = \begin{pmatrix} L_1(\theta_2) \\ L_2(\theta_1) \end{pmatrix}. \tag{3.11}$$

Here, K maps $\theta = (\theta_1, \theta_2)$ to a new parameter value through the simultaneous best responses of players 1 and 2.

Similarly, let $D_M \subset \mathbb{R}^{d_D}$ and let $M : D_M \to \mathbb{R}^{d_D}$ be a map such that

$$M(\theta_{-1}) = \begin{pmatrix} M_1(\theta_{-1}) \\ M_2(\theta_{-1}) \\ \vdots \\ M_{d_D}(\theta_{-1}) \end{pmatrix} = \begin{pmatrix} L_2\left(L_1(\theta_{-1}), \theta_{-\{1,2\}}\right) \\ L_3\left(L_1(\theta_{-1}), L_2(L_1(\theta_{-1}), \theta_{-\{1,2\}}), \theta_{-\{1,2,3\}}\right) \\ \vdots \\ L_J\left(L_1(\theta_{-1}), L_2(L_1(\theta_{-1}), \theta_{-\{1,2\}}), \cdots\right) \end{pmatrix}, \tag{3.12}$$

which can be interpreted as the players' sequential responses (first by Player 1, then Player 2, etc.) to an initial strategy $\theta_{-1} = (\theta_2, \ldots, \theta_J)$.[3] Note that the argument of M is not the entire parameter vector. Rather, it is a subvector of θ consisting of the coefficients on the endogenous variables. In order to find a fixed point, this feature is particularly attractive when the number of endogenous variables is small. With one endogenous variable (i.e. $\theta_2 \in \mathbb{R}$ is a scalar), the map simplifies to

$$M(\theta_2) = L_2\left(L_1\left(\theta_2\right)\right),$$

which is a univariate function whose fixed point is often straightforward to compute.

Define

$$\tilde{R}_1 := \big\{\theta_{-1} \in \Theta_{-1} : \Psi_{P,1}(\theta_1, \theta_{-1}) = 0, \\ \Psi_{P,2}(\theta_1, \theta_2, \pi_{-\{1,2\}}\theta_{-1}) = 0,\ \exists(\theta_1, \theta_2) \in \Theta_1 \times \Theta_2\big\}. \tag{3.13}$$

This is the set on which the map $\theta_{-1} \to L_2\left(L_1\left(\theta_{-1}\right), \pi_{-\{1,2\}}\theta_{-1}\right)$, the first component of M, is well-defined. We then recursively define $\tilde{R}_j$ for $j = 2, \ldots, d_D$ in a similar manner. A precise definition of these sets is given in Appendix C. Now define

$$D_M := \bigcap_{j=1}^{d_D} \tilde{R}_j = \tilde{R}_{d_D}, \tag{3.14}$$

where the second equality follows because $\tilde{R}_{d_D}$ turns out to be a subset of $\tilde{R}_j$ for all $j \leq d_D$. The following corollary ensures that K and M are well-defined on D_K and D_M respectively.

Corollary 1. *The maps $K : D_K \to \mathbb{R}^d$ and $M : D_M \to \mathbb{R}^{d_D}$ exist and are continuously differentiable on the interior of their domains.*

The key insight that we exploit is that, by construction of the BR maps, the problem of finding a solution to $\Psi_P(\theta) = 0$ is equivalent to the problem of finding a fixed-point of K (or M). The following proposition states the formal result.

Proposition 1. *Suppose Assumptions 1 and 2 hold. Then,*

(i) $\Psi_P(\theta^*) = 0$ *if and only if* $K\left(\theta^*\right) = \theta^*$

(ii) $\Psi_P(\theta^*) = 0$ *if and only if* $M\left(\theta^*_{-1}\right) = \theta^*_{-1}$ *and* $\theta^*_1 = L_1(\theta^*_{-1})$.

In view of Proposition 1, the original IVQR estimation problem can be reformulated as the problem of finding the fixed point of K (or M). This reformulation naturally leads to discrete

[3] One may define M by changing the order of responses as well. For theoretical analysis, it suffices to consider only one of them. Once the fixed point θ^*_{-1} of M is found, one may also obtain θ^*_1 using $\theta^*_1 = L_1(\theta^*_{-1})$.

dynamical systems associated with these maps, which in turn provides straightforward iterative algorithms for computing θ^*.

(1) Simultaneous dynamical system:[4]

$$\theta^{(s+1)} = K\left(\theta^{(s)}\right), \; s = 0, 1, 2, \ldots, \; \theta^{(0)} \text{ given.} \tag{3.15}$$

(2) Sequential dynamical system:[5]

$$\theta_{-1}^{(s+1)} = M\left(\theta_{-1}^{(s)}\right), \; s = 0, 1, 2, \ldots, \; \theta_{-1}^{(0)} \text{ given.} \tag{3.16}$$

where $\theta_1^{(s+1)} = L_1\left(\theta_{-1}^{(s)}\right)$.

These discrete dynamical systems will be the starting point for our estimation algorithms.[6]

4. Population Algorithms

In this section, we explore the implications of the fixed point reformulation for constructing population-level algorithms for computing fixed points.

4.1. Contraction-based Algorithms.

We first consider conditions under which K and M are contraction mappings. They ensure that the discrete dynamical systems induced by K and M are convergent to unique fixed points. Moreover, in view of Proposition 1, (point) identification is equivalent to the uniqueness of the fixed point of K (or M). Therefore, the conditions we provide below are also sufficient for the point identification of θ^*. We will discuss the relationship between our conditions and existing ones in the next section.

For any vector-valued map E, let $J_E(x)$ denote its Jacobian matrix evaluated at its argument x. We provide conditions in terms of the Jacobian matrices of K and M, which are well-defined by Corollary 1.

Assumption 3. *There exist open strictly convex sets $\tilde{D}_K \subseteq D_K$ and $\tilde{D}_M \subseteq D_M$ such that*

(1) $\|J_K(\theta)\| \leq \lambda$ for some $\lambda < 1$ for all $\theta \in \tilde{D}_K$;

(2) $\|J_M(\theta_{-1})\| \leq \lambda$ for some $\lambda < 1$ for all $\theta_{-1} \in \tilde{D}_M$.

[4] This algorithm is akin to the Jacobi computational procedure.

[5] Smyth (1996) considers this type of algorithm for $J = 2$ and calls it "zigzag" algorithm. It is akin to a Gauss-Seidel procedure.

[6] These discrete dynamical systems can also be viewed as learning dynamics in a game (Li and Basar, 1987; Fudenberg and Levine, 2007).

Under this additional assumption, the iterative algorithms are guaranteed to converge to the fixed point. We summarize this result below.

Proposition 2. *Suppose Assumptions 1, 2, and 3 hold. Then,*

(i) K is a contraction on the closure of $\tilde{D}_K$. The fixed point $\theta^* \in cl(\tilde{D}_K)$ of K is unique. For any $\theta^{(0)} \in \tilde{D}_K$, the sequence $\{\theta^{(s)}\}_{s=0}^{\infty}$ defined in (3.15) satisfies $\theta^{(s)} \to \theta^*$ as $s \to \infty$.

(ii) M is a contraction on the closure of $\tilde{D}_M$. The fixed point $\theta^*_{-1} \in cl(\tilde{D}_M)$ of M is unique. For any $\theta^{(0)}_{-1} \in \tilde{D}_M$, the sequence $\{\theta^{(s)}_{-1}\}_{s=0}^{\infty}$ defined in (3.16) satisfies $\theta^{(s)}_{-1} \to \theta^*_{-1}$ as $s \to \infty$.

In the case of a single endogenous variable, the Jacobian matrices of K and M are given by

$$J_K(\theta) = \begin{pmatrix} 0 & J_{L_1}(\theta_2) \\ J_{L_2}(\theta_1) & 0 \end{pmatrix}, \qquad \text{and} \qquad J_M(\theta_2) = J_{L_2}\left(L_1(\theta_2)\right) J_{L_1}\left(\theta_2\right),$$

where

$$J_{L_{-j}}\left(\theta_j\right) = -\left(\left.\frac{\partial \Psi_{P,-j}(\theta_j, \theta_{-j})}{\partial \theta'_{-j}}\right|_{\theta=(\theta_j, L_{-j}(\theta_j))} \right)^{-1} \left.\frac{\partial \Psi_{P,-j}(\theta_j, \theta_{-j})}{\partial \theta'_j}\right|_{\theta=(\theta_j, L_{-j}(\theta_j))}, \text{ for } j = 1, 2.$$

One may therefore check the high-level condition through the Jacobians of the original moment restrictions. In Appendix C.2.1, we illustrate a simple primitive condition for the local version of Assumption 3.

4.2. **Connections to Identification Conditions.** In view of Proposition 1, identification of θ^* is equivalent to uniqueness of the fixed points of K and M, which is ensured by Proposition 2. Here, we discuss how the conditions required by Proposition 2 relate to the ones in the literature.

We first start with local identification. The parameter vector θ^* is said to be locally identified if there is a neighborhood $\mathcal{N}$ of θ^* such that $\Psi_P(\theta) \neq 0$ for all $\theta \neq \theta^*$ in the neighborhood. Local identification in the IVQR model follows from standard results (e.g., Rothenberg, 1971; Chen, Chernozhukov, Lee, and Newey, 2014). For example, if $\Psi_P(\theta)$ is differentiable, Chen, Chernozhukov, Lee, and Newey (2014, Section 2.1) show that full rank of $J_{\Psi_P}(\theta)$ at θ^* is sufficient for local identification.

It is interesting to compare this full rank condition to Assumption 5.1 in the appendix, which is a local version of Assumption 3.1. Assumption 5.1 requires that $\rho\left(J_K(\theta^*\right) < 1$, where $\rho(A)$ denotes the spectral radius of a square matrix A. We highlight the connection in the case with a single endogenous variable. Full rank of $J_{\Psi_P}(\theta^*)$ is equivalent to $\det\left(J_{\Psi_P}(\theta^*)\right) \neq 0$. Observe that, for any θ,

$$\begin{aligned}
\det\left(J_{\Psi_P}(\theta)\right) &= \det\begin{pmatrix} \partial\Psi_{P,1}(\theta_1,\theta_2)/\partial\theta_1' & \partial\Psi_{P,1}(\theta_1,\theta_2)/\partial\theta_2' \\ \partial\Psi_{P,2}(\theta_1,\theta_2)/\partial\theta_1' & \partial\Psi_{P,2}(\theta_1,\theta_2)/\partial\theta_2' \end{pmatrix} \\
&= \det\left(\begin{pmatrix} \partial\Psi_{P,1}(\theta_1,\theta_2)/\partial\theta_1' & 0 \\ 0 & \partial\Psi_{P,2}(\theta_1,\theta_2)/\partial\theta_2' \end{pmatrix}\begin{pmatrix} I_{d_1} & -J_{L_1}(\theta_2) \\ -J_{L_2}(\theta_1) & I_{d_2} \end{pmatrix}\right) \\
&= \det\begin{pmatrix} \partial\Psi_{P,1}(\theta_1,\theta_2)/\partial\theta_1' & 0 \\ 0 & \partial\Psi_{P,2}(\theta_1,\theta_2)/\partial\theta_2' \end{pmatrix}\det\begin{pmatrix} I_{d_1} & -J_{L_1}(\theta_2) \\ -J_{L_2}(\theta_1) & I_{d_2} \end{pmatrix}.
\end{aligned}$$

If $\partial\Psi_{P,j}(\theta)/\partial\theta_j'|_{\theta=\theta^*}$ is invertible for $j=1,2$ (which is true under Assumption 2.4), $J_{\Psi_P}(\theta^*)$ is full rank if and only if

$$0 \neq \det\begin{pmatrix} I_{d_1} & -J_{L_1}(\theta_2^*) \\ -J_{L_2}(\theta_1^*) & I_{d_2} \end{pmatrix} = \det(I_d - J_K(\theta^*)). \tag{4.1}$$

That is, it requires that none of the eigenvalues has modulus one. Therefore, Assumption 5.1 is sufficient but not necessary for condition (4.1) to hold. Specifically, Assumption 5.1 requires all eigenvalues of $J_K(\theta^*)$ to lie strictly within the unit circle, while local identification only requires all eigenvalues not to be on the unit circle. In terms of the dynamical systems induced by K, the former ensures that the dynamical system has a unique *asymptotically stable fixed point*, while the latter ensures that the system has a unique *hyperbolic fixed point*, which is a more general class of fixed points (e.g. Galor, 2007).[7] Under the former condition, iteratively applying the contraction map induces convergence, while the latter generally requires a root finding method to obtain the fixed point.

Now we turn to global identification and compare Proposition 2 to the global identification result in Chernozhukov and Hansen (2006).

Lemma 2 (Theorem 2 in Chernozhukov and Hansen (2006))**.** *Suppose that Assumption 1 holds. Moreover, suppose that (i) Θ is compact and convex and θ^* is in the interior of Θ; (ii) $f_{Y|D,Z,X}$ is uniformly bounded a.s.; (iii) $J_\Psi(\theta)$ is continuous and has full rank uniformly over Θ; and (iv) the image of Θ under the mapping $\theta \to \Psi(\theta)$ is simply connected. Then, θ^* uniquely solves $\Psi(\theta) = 0$ over Θ.*

[7] The argument above also applies to settings with multiple endogenous variables. A similar result can also be shown for M.

Under Conditions (i)–(iv), which are substantially stronger than the local identification conditions discussed above, the result in Lemma 2 follows from an application of Hadamard's global univalence theorem (e.g. Theorem 1.8 in Ambrosetti and Prodi (1995)).

Comparing Lemma 2 to Proposition 2, we can see that the result in Lemma 2 establishes identification over the whole parameter space Θ, while Proposition 2 establishes identification over the sets $\tilde{D}_K$ and $\tilde{D}_M$, which will generally be subsets of Θ. Regarding the underlying assumptions, Conditions (i) and (ii) in Lemma 2 correspond to our Assumptions 2.1 and 2.3. Moreover, our Assumption 2.3 constitutes an easy-to-interpret sufficient condition for continuity of J_{Ψ_P} as required in Condition (iii). To apply Hadamard's global univalence theorem, Chernozhukov and Hansen (2006) assume the simple connectedness of the image of Ψ (Condition (iv)). By contrast, we use a different univalence theorem by Gale and Nikaido (1965) (applied to the map Ξ defined in (D.3) that arises from each subsystem), which does not require further conditions. However, when establishing global identification based on the contraction mapping theorem, we need to impose an additional condition on the Jacobian (Assumption 3). In sum, our conditions are somewhat stronger in terms of restrictions on the Jacobian, but they are relatively easy to check and allow us to dispense with an abstract condition (simple connectedness of the image of a certain map) to apply a global univalence theorem.

4.3. **Root-Finding Algorithms and Nesting.** Note that Assumption 3 is a sufficient condition for the uniqueness of the fixed point and the convergence of the contraction-based algorithm. Even in cases this assumption fails to hold, one may still identify θ^* and design an algorithm that is able to find it under weaker conditions on the Jacobian. This is the case under the assumptions in the general (global) identification result of Chernozhukov and Hansen (2006); see Lemma 2.

Note that, for the simultaneous dynamical system, θ^* solves

$$(I_d - K)(\theta^*) = 0, \tag{4.2}$$

where I_d is the identity map. Similarly, in the sequential dynamical system, θ^*_{-1} solves

$$(I_{d_D} - M)(\theta^*_{-1}) = 0. \tag{4.3}$$

Therefore, standard root-finding algorithms can be used to compute the fixed point.

For implementing root-finding algorithms, we find that reducing the dimension of the fixed point problem is often helpful. Toward this end, we briefly discuss another class of dynamical systems and associated population algorithms which can be used for the purpose of dimension reduction. Namely, with more than two players, one can construct *nested* dynamical systems, which induce nested fixed point algorithms. Nesting is useful as it allows transforming any setup with more than two players into a two-player system.

To fix ideas, consider the case of three players ($J = 3$). Fix player 3's action $\theta_3 \in \Theta_3 \subset \mathbb{R}$ and consider the associated "sub-game" between players 1 and 2. To describe the subgame, define $M_{1,2|3}(\cdot \mid \theta_3) : \Theta_2 \to \Theta_2$ pointwise by

$$M_{1,2|3}(\theta_2 \mid \theta_3) := L_2\left(L_1\left(\theta_2, \theta_3\right), \theta_3\right). \tag{4.4}$$

This map gives the the sequential best responses of players 1 and 2 while taking player 3's strategy given. Define the fixed point $L_{12} : \Theta_3 \to \Theta_1 \times \Theta_2$ of the subgame by

$$L_{12}(\theta_3) := \begin{pmatrix} \bar{\theta}_1(\theta_3) \\ \bar{\theta}_2(\theta_3) \end{pmatrix} = \begin{pmatrix} L_1(\bar{\theta}_2(\theta_3), \theta_3) \\ M_{1,2|3}(\bar{\theta}_2(\theta_3) \mid \theta_3). \end{pmatrix} \tag{4.5}$$

This map then defines a new "best response" map. Here, given θ_3, the players in the subgame (i.e. players 1 and 2) collectively respond by choosing the Nash equilibrium of the subgame. The overall dynamical system induced by the nested decentralization is then given by

$$M_3(\theta_3) = L_3\left(L_{12}(\theta_3)\right). \tag{4.6}$$

Hence, we can interpret the nested algorithm as a two-player dynamical system where one player solves an internal fixed point problem. This nesting procedure is generic and can be extended to more than three players by sequentially adding additional layers of nesting.[8] It follows that any decentralized estimation problem with more than two players can be reformulated as a nested dynamical system with two players: player J and all others $-J$. The resulting dynamical system $M_J(\theta_J) = L_J\left(L_{-J}(\theta_J)\right)$ is particularly useful when M_J is not necessarily a contraction map but θ_J is a scalar (which is the case in our IVQR model). As we see below, its fixed point can efficiently be computed using univariate root-finding algorithms.

5. Sample Estimation Algorithms

Let $\{(Y_i, D_i', X_i', Z_i')\}_{i=1}^N$ be a sample generated from the IVQR model. Our estimators are constructed using the analogy principle. For this, define the sample payoff functions for the players

[8] In the current example, consider adding player 4 and letting players 1-3 best respond by returning the fixed point of the subgame through M_3 given θ_4. One can repeat this for additional players. This procedure can also be applied to the simultaneous dynamical system induced by K.

as

$$Q_{N,1}(\theta) := \frac{1}{N}\sum_{i=1}^{N}\rho_\tau(Y_i - X_i'\theta_1 - D_{1,i}\theta_2 - \cdots - D_{d_D,i}\theta_J), \tag{5.1}$$

$$Q_{N,j}(\theta) := \frac{1}{N}\sum_{i=1}^{N}\rho_\tau(Y_i - X_i'\theta_1 - D_{1,i}\theta_2 - \cdots - D_{d_D,i}\theta_J)(Z_{j-1,i}/D_{j-1,i}), \quad j = 2, \ldots, J. \tag{5.2}$$

The sample BR functions are defined as

$$\hat{L}_1(\theta_{-1}) := \arg\min_{\tilde{\theta}_1 \in \mathbb{R}^{d_1}} Q_{N,1}(\tilde{\theta}_1, \theta_{-1}), \tag{5.3}$$

$$\hat{L}_j(\theta_{-j}) := \arg\min_{\tilde{\theta}_j \in \mathbb{R}} Q_{N,j}(\tilde{\theta}_j, \theta_{-j}), \quad j = 2, \ldots, J. \tag{5.4}$$

Assuming that the model is parametrized in such a way that $Z_{\ell,i}/D_{\ell,i}$, $\ell = 1, \ldots, d_D$, is positive, these are convex (weighted) QR problems for which fast solution algorithms exist. In our empirical applications and simulations, we use the R-package `quantreg` to estimate the QRs (Koenker, 2018). For example, $\hat{L}_2(\theta_{-2})$ can be computed by running a QR with weights Z_{1i}/D_{1i} in which one regresses $Y_i - X_i'\theta_1 - D_{2,i}\theta_3 - \cdots - D_{d_D,i}\theta_J$ on D_{1i} without a constant.

Remark 5.1. The proposed estimators rely on decentralizing the original non-smooth and non-convex IVQR GMM problem into a series of convex QR problems. The quality and the computational performance of our procedures therefore crucially depends on the choice of the underlying QR estimation approach, which deserves some further discussion. The interested reader is referred to Koenker (2017) for an excellent overview over the computational aspects of quantile regression. In this paper, we use the Barrodale and Roberts algorithm which is implemented as the default in the `quantreg` package and described in detail in Koenker and D'Orey (1987, 1994). This algorithm is computationally tractable for problems up several thousand observations. For larger problems, we recommend using interior point methods, potentially after preprocessing; see Portnoy and Koenker (1997) for a detailed description. These methods are conveniently implemented in the `quantreg` package. For very large problems, one can resort to first-order gradient descent methods, which are amenable to modern parallelized computation; see Section 5.5 in Koenker (2017) for an excellent introduction and simulation evidence on the performance of such methods.

We construct estimation algorithms by mimicking the population algorithms. Let $\hat{K}$ and $\hat{M}$ denote sample analogs of K and M:

$$\hat{K}(\theta) := \begin{pmatrix} \hat{L}_1(\theta_{-1}) \\ \vdots \\ \hat{L}_J(\theta_{J-1}) \end{pmatrix} \tag{5.5}$$

and

$$\hat{M}\left(\theta_{-1}\right) := \begin{pmatrix} \hat{M}_1(\theta_{-1}) \\ \hat{M}_2(\theta_{-1}) \\ \vdots \\ \hat{M}_{d_D}(\theta_{-1}) \end{pmatrix} = \begin{pmatrix} \hat{L}_2\left(\hat{L}_1(\theta_{-1}), \theta_{-\{1,2\}}\right) \\ \hat{L}_3\left(\hat{L}_1(\theta_{-1}), \hat{L}_2(\hat{L}_1(\theta_{-1}), \theta_{-\{1,2\}}), \theta_{-\{1,2,3\}}\right) \\ \vdots \\ \hat{L}_J\left(\hat{L}_1(\theta_{-1}), \hat{L}_2(\hat{L}_1(\theta_{-1}), \theta_{-\{1,2\}}), \cdots\right) \end{pmatrix}, \tag{5.6}$$

where $\theta_1 = \hat{L}_1\left(\theta_{-1}\right)$. These maps induce sample analogs of the dynamical systems in Section 3.

(1) Sample simultaneous dynamical system:

$$\theta^{(s+1)} = \hat{K}\left(\theta^{(s)}\right), \ s = 0, 1, 2, \ldots, \ \theta^{(0)} \text{ given.} \tag{5.7}$$

(2) Sample sequential dynamical system:

$$\theta_{-1}^{(s+1)} = \hat{M}\left(\theta_{-1}^{(s)}\right), \ s = 0, 1, 2, \ldots, \ \theta_{-1}^{(0)} \text{ given,} \tag{5.8}$$

where $\theta_1^{(s+1)} = \hat{L}_1\left(\theta_{-1}^{(s)}\right)$.

5.1. **Contraction-based Algorithms.** The first set of algorithms exploits that, under Assumption 3, $\hat{K}$ and $\hat{M}$ are contraction mappings with probability approaching one. In this case, we iterate the dynamical systems (5.7) or (5.8) until $\|\theta^{(s)} - \hat{K}\left(\theta^{(s)}\right)\|$ (or $\|\theta_{-1}^{(s)} - \hat{M}(\theta_{-1}^{(s)})\|$) is within a numerical tolerance e_N.[9] This iterative algorithm is known to converge at least linearly. The approximate sample fixed point $\hat{\theta}_N$ that meets the convergence criterion then serves as an estimator for θ.

5.2. **Algorithms based on Root-Finders and Optimizers.** As discussed in Section 4.3, for root-finding algorithms, the sequential dynamical system (induced by M) is particularly useful because it leads to a substantial dimension reduction. The original $(d_X + d_D)$-dimensional GMM estimation problem can be reduced to a d_D-dimensional root-finding problem. An estimator $\hat{\theta}_N$ of θ^* can be constructed as an approximate fixed point to the sample problem:

$$\|\hat{\theta}_{N,-1} - \hat{M}\left(\hat{\theta}_{N,-1}\right)\| \leq e_N, \tag{5.9}$$

where $\hat{\theta}_{N,1} = \hat{L}_1\left(\hat{\theta}_{N,-1}\right)$ and e_N is a numerical tolerance. This problem can be solved efficiently using well-established root-finding algorithms since $\hat{M}$ is easy to evaluate as the composition of standard QRs. When $d_D = 1$, one may use Brent's method (Brent, 1971) whose convergence is superlinear. When $d_D > 1$, one could apply the Newton-Raphson method, which achieves quadratic

[9] In the next section, we require $e_N = o(N^{-1/2})$, which ensures that the numerical error does not affect the asymptotic distribution.

convergence but requires an estimate or a finite difference approximation of the derivative. The corresponding approximation error may affect the performance. Alternatively, on can compute the fixed point by minimizing $\|\hat{M}(\theta) - \theta\|^2$. The potential issue with this approach is that translating the root-finding problem into a minimization problem can lead to local minima in the objective function. Therefore, it is important to use global optimization strategies.

As described in Section 4.3, nesting can be used to reduce the dimensionality even further. In particular, the problem can be reformulated as a one-dimensional fixed point problem, which can be solved efficiently using existing methods. We found that Brent's method works very well in our context.

6. Asymptotic Theory

6.1. **Estimators.** We define an estimator $\hat{\theta}_N$ of θ^* as an approximate fixed point of $\hat{K}$ in the following sense:

$$\|\hat{\theta}_N - \hat{K}(\hat{\theta}_N)\| \leq \inf_{\theta' \in \Theta} \|\theta' - \hat{K}(\theta')\| + o_p(N^{-1/2}). \tag{6.1}$$

In what follows, we call $\hat{\theta}_N$ the *fixed point estimator* or θ^*. Alternatively, using $\hat{M}$, one may define an estimator $\hat{\theta}_{N,-1}$ of θ_{-1} as

$$\|\hat{\theta}_{N,-1} - \hat{M}(\hat{\theta}_{N,-1})\| \leq \inf_{\theta'_{-1} \in \Theta_{-1}} \|\theta'_{-1} - \hat{M}(\theta'_{-1})\| + o_p(N^{-1/2}). \tag{6.2}$$

An estimator of θ_1^* can be constructed by setting

$$\hat{\theta}_{N,1} := \hat{L}_1(\hat{\theta}_{N,-1}). \tag{6.3}$$

Under the conditions we introduce below, the definitions in (6.1) and (6.2)–(6.3) are asymptotically equivalent; see Lemma 5 in the appendix for a proof. Therefore, we mostly focus on the definition based on $\hat{K}$ below. $\hat{K}$ (or $\hat{M}$) is defined similarly for the nested dynamical system in which one player solves a fixed-point problem in a subgame.

Consistency and parametric convergence rates of $\hat{\theta}_N$ can be established using existing results. When $\hat{K}$ (or $\hat{M}$) is asymptotically a contraction map, one may construct an estimator $\hat{\theta}_N$ satisfying (6.1) using the contraction algorithm in Section 5.1 with tolerance $e_N = o(N^{-1/2})$. One may then apply the result of Dominitz and Sherman (2005) to obtain the root-N consistency of the estimator.[10] For completeness, this result is summarized in Appendix G.

[10] Satisfying $e_N = o(N^{-1/2})$ requires the number of iterations to increase as the sample size tends to infinity, which in turn satisfies requirement (ii) in Theorem 2 (Dominitz and Sherman, 2005).

More generally, if $\hat{K}$ is not guaranteed to be a contraction, one may use root-finding algorithms that solve $\theta - \hat{K}(\theta) = 0$ or $\theta_{-1} - \hat{M}(\theta_{-1}) = 0$ up to approximation errors of $o(N^{-1/2})$. The root-N consistency of $\hat{\theta}_N$ then follows from the standard argument for extremum estimators, in which we take $\mathcal{L}_N(\theta) = \|\theta - \hat{K}(\theta)\|$ as a criterion function.[11] Since these results are standard, we omit details and focus below on the asymptotic distribution and bootstrap validity of the fixed point estimators. Our contributions are two-fold. First, we establish the asymptotic distribution of the fixed point estimator without assuming that $\hat{K}$ or $\hat{M}$ is an asymptotic contraction map, which therefore allows the practitioner to conduct inference using the estimator based on the general root-finding algorithm and complements the result of Dominitz and Sherman (2005). Second, to our knowledge, the bootstrap validity of the fixed point estimators is new. These results are established by showing that, under regularity conditions, the population fixed point is Hadamard-differentiable and hence admits the use of the functional δ-method, which may be of independent theoretical interest.

Remark 6.1. To establish the asymptotic properties, one could try to reformulate our estimator as an estimator that approximately solves a GMM problem. Here, instead of relying on another reformulation, which would require establishing a sample analog version of Proposition 1, we develop and directly apply an asymptotic theory for fixed point estimators. The theory itself contains generic results (Theorem 1 and Lemmas 6–7) surrounding the Hadamard-differentiability of fixed points, which can potentially be used to analyze decentralized estimators outside the IVQR class.

6.2. **Asymptotic Theory and Bootstrap Validity.** The following theorem gives the limiting distribution of our estimator. For each $w = (y, d', x', z')'$ and $\theta \in \Theta$, let $f(w;\theta) \in \mathbb{R}^{d_X + d_D}$ be a vector whose sub-vectors are given by

$$f_1(w;\theta) = (1\{y \le d'\theta_{-1} + x'\theta_1\} - \tau)x,$$
$$f_j(w;\theta) = (1\{y \le d'\theta_{-1} + x'\theta_1\} - \tau)z_{j-1}, \quad j = 2, \dots, J,$$

and let $g(w;\theta) = (g_1(w;\theta)', \dots, g_J(w;\theta))'$ be a vector such that

$$g_j(w;\theta) = \frac{\partial^2}{\partial\theta_j \partial\theta_j'} Q_{P,j}(L_j(\theta_{-j}), \theta_{-j})^{-1} f_j(w; L_j(\theta_{-j}), \theta_{-j}),\ j = 1, \dots, J. \tag{6.4}$$

[11] The key conditions for these results, uniform convergence (in probability) of $\hat{K}$ and its stochastic equicontinuity, are established in Lemma 10.

Theorem 1. *Suppose that Assumptions 1 and 2 hold. Let $\{W_i\}_{i=1}^N$ be an i.i.d. sample generated from the IVQR model, where $W_i = (Y_i, D_i', X_i', Z_i')$. Then,*

$$\sqrt{N}(\hat{\theta}_N - \theta^*) \xrightarrow{L} N(0, V) \ , \tag{6.5}$$

with

$$V = (I_d - J_K(\theta^*))^{-1} E[\mathbb{W}(\theta^*)\mathbb{W}(\theta^*)'](I_d - J_K(\theta^*))^{-1}, \tag{6.6}$$

where $\mathbb{W}$ is a tight Gaussian process in $\ell^\infty(\Theta)^d$ with the covariance kernel

$$\mathrm{Cov}(\mathbb{W}(\theta), \mathbb{W}(\tilde{\theta})) = E_P\big[(g(W;\theta) - E_P[g(W;\theta)])(g(w;\tilde{\theta}) - E_P[g(w;\tilde{\theta})])'\big]. \tag{6.7}$$

To conduct inference on θ^*, one may employ a natural bootstrap procedure. For this, use in (5.5) and (6.1) the bootstrap sample instead of the original sample to define the bootstrap analogs $\hat{K}^*$ and $\hat{\theta}_N^*$ of $\hat{K}$ and $\hat{\theta}_N$. In practice, the bootstrap can be implemented using the following steps.

(1) Compute the fixed point estimator $\hat{\theta}_N$ using the original sample.

(2) Draw a bootstrap sample $\{W_i^*\}_{i=1}^N$ randomly with replacement from P_N. Use the simultaneous (or sequential) dynamical system based on $\hat{K}^*$ (or $\hat{M}^*$) combined with a contraction or root-finding algorithm to compute $\hat{\theta}_N^*$.

(3) Repeat Step 2 across bootstrap replications $b = 1, \dots, B$. Let

$$F_B(x) := \frac{1}{B}\sum_{b=1}^{B} 1\Big\{\sqrt{N}(\hat{\theta}_N^{*,b} - \hat{\theta}_N) \leq x\Big\}, \ x \in \mathbb{R}. \tag{6.8}$$

Use F_B as an approximation to the sampling distribution of the root $\sqrt{N}(\hat{\theta}_N - \theta^*)$.

We would like to emphasize that the bootstrap is particularly attractive in conjunction with our new and computationally efficient estimation algorithms. By contrast, directly bootstrapping for instance the IQR estimator of Chernozhukov and Hansen (2006) is computationally very costly. Alternative methods (either an asymptotic approximation or a score-based bootstrap) require estimation of the influence function, which involves nonparametric estimation of a certain conditional density. Directly bootstrapping our fixed point estimators avoids the use of any smoothing parameter.[12]

[12] The use of the bootstrap here is for consistently estimating the law of the estimator. Whether one may obtain higher-order refinements through a version of the bootstrap, e.g., the m out of n bootstrap with extrapolation (Sakov and Bickel, 2000), is an interesting question which we leave for future research.

The following theorem establishes the consistency of the bootstrap procedure. For this, let $\overset{L^*}{\rightsquigarrow}$ denote the weak convergence of the bootstrap law in outer probability, conditional on the sample path $\{W_i\}_{i=1}^{\infty}$.

Theorem 2. *Suppose that Assumptions 1 and 2 hold. Let $\{W_i\}_{i=1}^N$ be an i.i.d. sample generated from the IVQR model. Then,*

$$\sqrt{N}(\hat{\theta}_N^* - \hat{\theta}_N) \overset{L^*}{\rightsquigarrow} N(0, V),$$

where V is as in (6.6).

7. Empirical Example

In this section, we illustrate the proposed estimators by reanalyzing the effect of 401(k) plans on savings behavior as in Chernozhukov and Hansen (2004). This empirical example constitutes the basis for our Monte Carlo simulations in Section 8. As explained by Chernozhukov and Hansen (2004), 401(k) plans are tax-deferred savings options that allow deducting contributions from taxable income and accruing tax-free interest. These plans are provided by employers and were introduced in the United States in the early 1980s to increase individual savings. To estimate the effect of 401(k) plans (D) on accumulated assets (Y) on has to deal with the potential endogeneity of the actual participation status. Chernozhukov and Hansen (2004) propose an instrumental variables approach to overcome this problem. They use 401(k) eligibility as an instrument (Z) for the participation in 401(k) plans. The argument behind this strategy, which is due to Poterba, Venti, and Wise (1994, 1995, 1998) and Benjamin (2003), is that eligibility is exogenous after conditioning on income and other observable factors. We use the same identification strategy here but note that there are also papers which argue that 401(k) eligibility is not conditionally exogenous (e.g., Engen, Gale, and Scholz, 1996).

We use the same dataset as in Chernozhukov and Hansen (2004). The dataset contains information about 9913 observations from a sample of households from the 1991 Survey of Income and Program Participation.[13] We refer to Chernozhukov and Hansen (2004) for more information about the data and to their Tables 1 and 2 for descriptive statistics. Here we focus on net financial assets as our outcome of interest.[14]

[13] The dataset analyzed by Chernozhukov and Hansen (2004) has 9,915 observations. Here we delete the two observations with negative income.

[14] Chernozhukov and Hansen (2004) also consider total wealth and net non-financial assets.

We consider the following linear model for the conditional potential outcome quantiles

$$q(D, X, \tau) \quad = \quad D\theta_2(\tau) + X'\theta_1(\tau). \tag{7.1}$$

The vector of covariates X includes seven dummies for income categories, five dummies for age categories, family size, four dummies for education categories, indicators for marital status indicator, two-earner status, defined benefit pension status, individual retirement account participation status and homeownership, and a constant. Because $P(D = 0) > 0$, we re-parametrize the model by replacing D by $D^\star = D + 1$ to ensure that $Z/D^\star$ is well-defined and positive.

We found that, in this empirical setting (and simulations based on it), contraction algorithms based on $\hat{K}$ can be rather sensitive to the choice of starting values. We therefore focus on algorithms based on $\hat{M}$. Figure 1 graphically illustrates our fixed point algorithms. It displays $\hat{M}$ at three different quantile levels $\tau \in \{0.25, 0.50, 0.75\}$. Our theoretical results show that, under appropriate conditions, the intersection between $\hat{M}$ and the 45-degree line provides an estimate of θ_2. Figure 1 further provides a straightforward graphical way to check the validity of the sample analog of Assumption 3. We can see that the sample analog of J_M (i.e. the slope of M) is smaller than one. This suggests that the contraction-based sequential algorithm converges at all three quantile levels, which is indeed what we find.

[Figure 1 about here.]

We consider two different algorithms based on $\hat{M}$: a contraction algorithm and a root-finding algorithm based on Brent's method implemented by the R-package `uniroot`. We compare our estimators to the IQR estimator of Chernozhukov and Hansen (2006) with 500 grid points which provides a slow but very robust benchmark.

Figure 2 displays the estimates of $\theta_2(\tau)$ for $\tau \in \{0.15, 0.20, \dots, 0.85\}$. We can see that all estimation algorithms yield very similar results. We also note that the contraction-based algorithm converges for all quantile levels considered.

[Figure 2 about here.]

Figures 3 depicts pointwise 95% confidence intervals for the proposed estimators obtained using on the empirical bootstrap described in Section 6.2 with 500 replications. We can see that the resulting confidence intervals are very similar for both algorithms and do not include zero at all quantile levels considered.

[Figure 3 about here.]

8. Simulation Study

In this section, we assess and compare the finite sample performance of our estimation algorithms. We first discuss the competing algorithms and then introduce the DGPs.

8.1. Estimation Algorithms.

In this section we assess and compare the performance of several different algorithms all of which are based on the dynamical system $\hat{M}$. We do not explore contraction algorithms based on $\hat{K}$ because we found them to be less robust than the corresponding algorithms based on $\hat{M}$ and somewhat sensitive to the choice of starting values. For the root-finding algorithms, using $\hat{K}$ will typically be less attractive than using $\hat{M}$ because the dimensionality of the root-finding problem is much larger when using $\hat{K}$ $(d_D + d_X)$ than when using $\hat{M}$ (d_D).

For the models with one endogenous variable, we consider a contraction algorithm and a root-finding algorithm based on Brent's method. For models with two endogenous variables, we analyze a contraction algorithm, a root-finding algorithm implemented as a minimization problem based on simulated annealing (SA), and a nested root-finding algorithm based on Brent's method.[15] For all estimators, we use two-stage least squares estimates as starting values. We compare the results of our algorithms to those obtained from IQR, which serves as a slow but very robust benchmark. We use 500 (one endogenous variable) and 1600 (two endogenous variables) grid points for IQR.[16] Table 1 presents more details about the algorithms.

[Table 1 about here.]

8.2. An Application-Based DGP.

Here we consider DGPs which are based on the empirical application of Section 7.[17] We focus on a simplified setting with only two covariates: income and age. The covariates are drawn from their joint empirical distribution. The instrument Z_i is generated as Bernoulli $\left(\bar{Z}\right)$, where $\bar{Z}$ is the mean of the instrument in the empirical application. We then generate the endogenous variable as $D_i = Z_i \cdot 1\left\{0.6 \cdot V_i < U_i\right\}$, where $U_i \sim \text{Uniform}(0,1)$ and $V_i \sim \text{Uniform}(0,1)$ are independent disturbances. The DGP for D_i is chosen to roughly match

[15] We have also explored algorithms based on Newton-Raphson-type root-finders. These algorithms are, in theory, up to an order of magnitude faster than the contraction algorithm and the nested algorithm, but, unlike the other algorithms considered here, require an approximation to the Jacobian and are not very robust to the choice of starting values. We therefore do not report the results here.

[16] We note that one could of course always improve the performance of IQR by increasing the number of grid points. However, as we document below, IQR becomes computationally prohibitive in this case.

[17] The construction of our DGPs is inspired by the construction of the application-based DGPs in Kaplan and Sun (2017).

the joint empirical distribution of (D_i, Z_i). The outcome variable Y_i is generated as

$$Y_i = X_i'\theta_1(U_i) + D_i\theta_2(U_i) + G^{-1}(U_i). \tag{8.1}$$

The coefficient $\theta_1(U_i)$ is constant and equal to the IQR median estimate in the empirical application. $\theta_2(U_i) = 5000 + U_i \cdot 10000$ is chosen to match the increasing shape of the estimated conditional quantile treatment effects in Figure 2. $G^{-1}(\cdot)$ is the quantile function of a re-centered Gamma distribution, estimated to match the distribution of the IQR residuals at the median. To investigate the performance of our procedure with more than one endogenous variable, we add a second endogenous regressor:

$$Y_i = X_i'\theta_1(U_i) + D_i\theta_2(U_i) + D_{2i}\theta_3(U_i) + G^{-1}(U_i), \tag{8.2}$$

where we set $\theta_3(U_i) = 10000$. The second endogenous variable is generated as

$$D_{2i} = 0.8 \cdot Z_{2i} + 0.2 \cdot \Phi^{-1}(U_i)$$

and the second instrument is generated as $Z_{2i} \sim N(0,1)$. We set $N = 9913$ as in the empirical application.

First, we investigate the finite sample bias and root mean squared error (RMSE) of the different methods. Tables 2 and 3 present the results. With one endogenous regressor, all three methods perform well and exhibit very a similar bias and RMSE. Turning to the results with two endogenous regressors, we can see that the nested algorithm exhibits the best overall performance, while the performance of our other algorithms is only slightly worse. The finite sample properties of the proposed algorithms are comparable to IQR.

[Table 2 about here.]

[Table 3 about here.]

Next, we analyze the finite sample properties of our bootstrap inference procedure. Table 4 shows the empirical coverage probabilities for the contraction-based algorithm and the root-finding algorithm based on Brent's method. Both methods exhibit coverage rates which are very close to the respective nominal level.

[Table 4 about here.]

Finally, we investigate the computational performance of the different procedures. Tables 5 and 6 show the average computation time (in seconds) for estimating the model with one and two endogenous variables for different sample sizes. We compare our procedures to the IQR algorithm with a grid search over 500 points (one endogenous regressor) and 100×100 points (two endogenous regressors). Note that we choose a higher (and arguably more practically relevant) number of grid points for the model with two endogenous regressors than in the simulations.[18] All computations were carried out on a standard desktop computer with a 3.2 GHz Intel Core i5 processor and 8GB RAM.

With one endogenous regressor, both of our algorithms are computationally much more efficient than IQR. Specifically, the root-finding algorithm based on Brent's method is about 10 to 30 times as fast as IQR, and the contraction algorithm is 1.5 to 12 times as fast. Among our algorithms, the root-finding method is almost twice as fast as the contraction-based iterative algorithm. However, it is important to note that the computational speed of the contraction algorithm depends on $|\hat{J}_M|$, which is rather close to one in this application (cf. Figure 1). This implies that the contraction algorithm will be rather slow here and can be expected to be faster in other applications.

The computational gain of our algorithms becomes more pronounced with two endogenous variables. Looking at the results in Table 6, IQR's average computation times are around two orders of magnitude slower than those of our procedures. Specifically, the nested root-finding algorithm is 70 to 125 times as fast as the IQR, while the contraction algorithm is 110 to 215 times as fast. This is as expected since, due to the use of grids, IQR's computational cost increases exponentially as the number of endogenous variables increases.[19] Among our algorithms, the contraction algorithm is almost twice as fast as the nested algorithm. However, we would like to emphasize that both of these procedures are computationally very efficient even for large samples. By contrast, the minimization-based algorithm based on SA is about an order of magnitude slower that the contraction algorithm and the nested algorithm.

[Table 5 about here.]

[Table 6 about here.]

[18] We found that using the same number of grid points for IQR in the simulations reported in Tables 2-4 was computationally prohibitive.

[19] Our implementation of IQR with two endogenous variables is inherently slower than the implementation with one endogenous variable, even when the number of grid points is the same. First, there is an additional covariate in the underlying QRs (the second instrument). Second, with one endogenous variable, we choose the grid value that minimizes the absolute value of the coefficient on the instrument. By contrast, with two endogenous regressors, we choose the grid point which minimizes a quadratic form based on the inverse of the estimated QR variance covariance matrix as suggested in Chernozhukov, Hansen, and Wüthrich (2017), which requires an additional computational step.

8.3. **Additional Simulations.** This section presents some additional simulation evidence based on the following location-scale shift model:

$$Y_i = \gamma_1 + \gamma_2 X_i + \gamma_3 D_{1i} + \gamma_4 D_{2i} + (\gamma_5 + \gamma_6 D_{1i} + \gamma_7 D_{2i}) U_i \tag{8.3}$$

Here D_{1i} and D_{2i} are the endogenous variables of interest and X_i is an exogenous covariate. In addition, we have access to two instruments Z_{1i} and Z_{2i}. For $\gamma_2 = \gamma_4 = \gamma_7 = 0$, this model reduces to the model considered in Section 6.1 of Andrews and Mikusheva (2016). We set $\gamma_1 = \cdots = \gamma_7 = 1$. To evaluate the performance of our algorithms with one endogenous variable, we set $\gamma_4 = \gamma_7 = 0$ and use Z_{1i} as the instrument. Following Andrews and Mikusheva (2016), we consider a symmetric as well as an asymmetric DGP for $(U_i, D_{1i}, D_{2i}, Z_{1i}, Z_{2i}, X_i)$:

$$(U_i, D_{1i}, D_{2i}, Z_{1i}, Z_{2i}, X_i) = (\Phi(\xi_{U,i}), \Phi(\xi_{D_1,i}), \Phi(\xi_{D_2,i}), \Phi(\xi_{Z_1,i}), \Phi(\xi_{Z_2,i}), \Phi(\xi_{X,i})), \qquad \text{(symmetric)}$$

$$(U_i, D_{1i}, D_{2i}, Z_{1i}, Z_{2i}, X_i) = (\xi_{U,i}, \exp(2\xi_{D_1,i}), \xi_{D_2,i}, \xi_{Z_1,i}, \xi_{Z_2,i}, \xi_{X,i}), \qquad \text{(asymmetric)}$$

where $(\xi_{U,i}, \xi_{D_1,i}, \xi_{D_2,i}, \xi_{Z_1,i}, \xi_{Z_2,i}, \xi_{X,i})$ is a Gaussian vector with mean zero, all variances are set equal to one, $Cov(\xi_U, \xi_{D_1}) = Cov(\xi_U, \xi_{D_2}) = 0.5$, $Cov(\xi_{D_1}, \xi_{Z_1}) = 0.8$, $Cov(\xi_{D_2}, \xi_{Z_2}) = 0.4$, which allows us to investigate the impact of instrument strength, all other covariances are equal to zero, and Φ is the cumulative distribution function of the standard normal distribution.

We first investigate the bias and RMSE of the different methods. Tables 7–10 present the results. With one endogenous variable, the performances of the root-finding algorithm using Brent's method and IQR are very similar both in terms of bias and RMSE. The contraction algorithm performs well but exhibits some bias at the tail quantiles. Turning to the results with two endogenous variables, we can see that the nested algorithm exhibits the best overall performance, both in terms of bias and RMSE. The performances of the SA-based optimization algorithm and IQR are similar and only slightly worse than that of the nested algorithm. The contraction algorithm tends to exhibit some bias at the tail quantiles. However, this bias decreases substantially as the sample size gets larger. Finally, comparing the results for the coefficients on D_1 and D_2, we can see that the instrument strength matters for the performance of all the estimator (including IQR), suggesting that weak identification can have implications for the estimation of IVQR models.

[Table 7 about here.]

[Table 8 about here.]

[Table 9 about here.]

[Table 10 about here.]

Table 11 displays the empirical coverage probabilities of the bootstrap confidence intervals. The results show that the our bootstrap procedure exhibits very good size properties. The confidence intervals based on the contraction algorithm tend to be somewhat closer to the nominal level than those based on Brent's method, which exhibits some over-coverage, especially for $N = 500$ and $\alpha = 0.1$.

[Table 11 about here.]

9. Conclusion

The main contribution of this paper is to develop computationally convenient and easy-to-implement estimation algorithms for IVQR models. Our key insight is that the non-smooth and non-convex IVQR estimation problem can be decomposed into a sequence of much more tractable convex QR problems, which can be solved very quickly using well-established methods. The proposed algorithms are particularly well-suited if the number of exogenous variables is large and the number of endogenous variables is moderate as in many empirical applications.

An avenue for further research is to investigate weak identification robust inference within the decentralized model. One may, for example, write the (re-scaled) sample fixed point restriction as $\sqrt{N}(I - \hat{K})(\theta) = s_N(\theta) + \mathbb{W}(\theta) + r_N(\theta)$, where $s_N(\theta) = \sqrt{N}(I - K)(\theta)$, $\mathbb{W}$ is a Gaussian process, and r_N is an error that tends 0 uniformly. This paper assumes that $s_N(\theta^*) = 0$ uniquely, and outside $N^{-1/2}$-neighborhoods of θ^*, $s_N(\theta)$ diverges and dominates $\mathbb{W}$. For a one-dimensional FP problem, this requires the BR map to be bounded away from the 45-degree line outside any $N^{-1/2}$-neighborhood of the fixed point. However if s_N fails to dominate $\mathbb{W}$ over a substantial part of the parameter space, one would end up with weak identification.[20] How to conduct robust inference in such settings is an interesting question, which we leave for future research.

Finally, we note that while we study the performance of the proposed algorithms separately, our reformulation and the resulting algorithms are potentially very useful when combined with other existing procedures. For instance, one could choose starting values using an initial grid search over a coarse grid and then apply the contraction algorithm.

[*] Economics Department, Boston University, hkaido@bu.edu

[†] Economics Department, UC San Diego, kwuthrich@ucsd.edu

[20] Andrews and Mikusheva (2016) study weak identification robust inference methods in models characterized by moment restrictions.

References

Abadie, A., J. Angrist, and G. Imbens (2002): "Instrumental variable estimates of the effect of subsidized training on the quantile of trainee earnings," *Econometrica*, 70(1), pp. 91–117.

Ambrosetti, A., and G. Prodi (1995): *A primer of nonlinear analysis.* Cambridge Studies in Advanced Mathematics, vol. 34. Cambridge University Press, Cambridge.

Andrews, D. W. (1994): "Empirical process methods in econometrics," *Handbook of econometrics*, 4, 2247–2294.

Andrews, I., and A. Mikusheva (2016): "Conditional Inference With a Functional Nuisance Parameter," *Econometrica*, 84(4), 1571–1612.

Arcidiacono, P., G. Foster, N. Goodpaster, and J. Kinsler (2012): "Estimating spillovers using panel data, with an application to the classroom," *Quantitative Economics*, 3(3), 421–470.

Belloni, A., V. Chernozhukov, I. Fernandez-Val, and C. Hansen (2017): "Program Evaluation and Causal Inference With High-Dimensional Data," *Econometrica*, 85(1), 233–298.

Benjamin, D. (2003): "Does 401(k) eligibility increase saving?: Evidence from propensity score subclassification," *Journal of Public Economics*, 87, pp. 1259–1290.

Bertsekas, D. P., and J. N. Tsitsiklis (1989): *Parallel and distributed computation: numerical methods*, vol. 23. Prentice hall Englewood Cliffs, NJ.

Brent, R. P. (1971): "An algorithm with guaranteed convergence for finding a zero of a function," *The Computer Journal*, 14(4), 422–425.

Chamberlain, G. (1987): "Asymptotic efficiency in estimation with conditional moment restrictions," *Journal of Econometrics*, 34(3), 305 – 334.

Chen, L.-Y., and S. Lee (2018): "Exact computation of GMM estimators for instrumental variable quantile regression models," *Journal of Applied Econometrics*, 33(4), 553–567.

Chen, M., I. Fernandez-Val, and M. Weidner (2014): "Nonlinear Panel Models with Interactive Effects," arXiv:1412.5647.

Chen, X., V. Chernozhukov, S. Lee, and W. K. Newey (2014): "Local Identification of Nonparametric and Semiparametric Models," *Econometrica*, 82(2), 785–809.

Chen, X., and D. Pouzo (2009): "Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals," *Journal of Econometrics*, 152(1), pp. 46–60.

——— (2012): "Estimation of Nonparametric Conditional Moment Models With Possibly Nonsmooth Generalized Residuals," *Econometrica*, 80(1), pp. 277–321.

Chernozhukov, V., I. Fernandez-Val, and B. Melly (2013): "Inference on Counterfactual Distributions," *Econometrica*, 81(6), pp. 2205–2268.

Chernozhukov, V., and C. Hansen (2004): “The Effects of 401(k) Participation on the Wealth Distribution: An Instrumental Quantile Regression Analysis,” *The Review of Economics and Statistics*, 86(3), pp. 735–751.

——— (2005): “An IV Model of Quantile Treatment Effects,” *Econometrica*, 73(1), pp. 245–261.

Chernozhukov, V., and C. Hansen (2006): “Instrumental quantile regression inference for structural and treatment effects models,” *Journal of Econometrics*, 132, pp. 491–525.

Chernozhukov, V., and C. Hansen (2008): “Instrumental variable quantile regression: A robust inference approach,” *Journal of Econometrics*, 142(1), pp. 379–398.

Chernozhukov, V., and C. Hansen (2013): “Quantile Models with Endogeneity,” *Annual Review of Economics*, 5(1), pp. 57–81.

Chernozhukov, V., C. Hansen, and M. Jansson (2009): “Finite sample inference for quantile regression models,” *Journal of Econometrics*, 152(2), pp. 93–103.

Chernozhukov, V., C. Hansen, and K. Wüthrich (2017): “Instrumental Variable Quantile Regression,” in *Handbook of Quantile Regression*, ed. by V. Chernozhukov, X. He, R. Koenker, and L. Peng. CRC Chapman-Hall, forthcoming.

Chernozhukov, V., and H. Hong (2003): “An MCMC approach to classical estimation,” *Journal of Econometrics*, 115(2), pp. 293–346.

Chernozhukov, V., G. W. Imbens, and W. K. Newey (2007): “Instrumental variable estimation of nonseparable models,” *Journal of Econometrics*, 139(1), pp. 4–14.

Chesher, A. (2003): “Identification in Nonseparable Models,” *Econometrica*, 71(5), pp. 1405–1441.

de Castro, L., A. F. Galvao, D. M. Kaplan, and X. Liu (2018): “Smoothed GMM for quantile models,” Journal of Econometrics, accepted.

D’Haultfoeuille, X., and P. Février (2015): “Identification of Nonseparable Triangular Models With Discrete Instruments,” *Econometrica*, 83(3), pp. 1199–1210.

Dominitz, J., and R. P. Sherman (2005): “Some convergence theory for iterative estimation procedures with an application to semiparametric estimation,” *Econometric Theory*, 21(04), 838–863.

Engen, E. M., W. G. Gale, and J. K. Scholz (1996): “The Illusory Effects of Saving Incentives on Saving,” *Journal of Economic Perspectives*, 10(4), pp. 113–138.

Frandsen, B. R., M. Frölich, and B. Melly (2012): “Quantile treatment effects in the regression discontinuity design,” *Journal of Econometrics*, 168(2), pp. 382–395.

Frölich, M., and B. Melly (2013): "Unconditional Quantile Treatment Effects Under Endogeneity," *Journal of Business & Economic Statistics*, 31(3), pp. 346–357.

Fudenberg, D., and D. K. Levine (2007): *The theory of learning in games.* Cambridge: MIT Press.

Gagliardini, P., and O. Scaillet (2012): "Nonparametric Instrumental Variable Estimation of Structural Quantile Effects," *Econometrica*, 80(4), pp. 1533–1562.

Gale, D., and H. Nikaido (1965): "The Jacobian matrix and global univalence of mappings," *Mathematische Annalen*, 159(2), 81–93.

Galor, O. (2007): *Discrete dynamical systems.* Springer Science & Business Media.

Guimaraes, P., and P. Portugal (2010): "A simple feasible procedure to fit models with high-dimensional fixed effects," *Stata Journal*, 10(4), pp. 628–649.

Hasselblatt, B., and A. Katok (2003): *A First Course in Dynamics with a Panorama of Recent Developments.* Cambridge University Press.

Horn, R. A., and C. R. Johnson (1990): *Matrix analysis.* Cambridge university press.

Horowitz, J. L., and S. Lee (2007): "Nonparametric Instrumental Variables Estimation of a Quantile Regression Model," *Econometrica*, 75(4), pp. 1191–1208.

Husmann, K., A. Lange, and E. Spiegel (2017): *The R Package optimization: Flexible Global Optimization with Simulated-Annealing.*

Imbens, G. W., and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62(2), pp. 467–475.

Imbens, G. W., and W. K. Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations Models without Additivity," *Econometrica*, 77(5), pp. 1481–1512.

Jun, S. J. (2008): "Weak identification robust tests in an instrumental quantile model," *Journal of Econometrics*, 144(1), 118 – 138.

——— (2009): "Local structural quantile effects in a model with a nonseparable control variable," *Journal of Econometrics*, 151(1), 82 – 97.

Kaplan, D. M., and Y. Sun (2017): "Smoothed Estimation Equations for Instrumental Variables Quantile Regression," *Econometric Theory*, 33(1), 105–157.

Koenker, R. (2017): "Computational Methods for Quantile Regression," in *Handbook of Quantile Regression*, ed. by V. Chernozhukov, X. He, R. Koenker, and L. Peng, pp. pp. 55–67. CRC Chapman-Hall.

——— (2018): *quantreg: Quantile Regression* R package version 5.36.

KOENKER, R., AND J. BASSETT, GILBERT (1978): "Regression Quantiles," *Econometrica*, 46(1), pp. 33–50.

KOENKER, R. W., AND V. D'OREY (1987): "Algorithm AS 229: Computing Regression Quantiles," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 36(3), 383–393.

——— (1994): "Remark AS R92: A Remark on Algorithm AS 229: Computing Dual Regression Quantiles and Regression Rank Scores," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 43(2), 410–414.

KRANTZ, S. G., AND H. R. PARKS (2012): *The implicit function theorem: history, theory, and applications*. Springer Science & Business Media.

LEE, J., AND K. SEO (2015): "A computationally fast estimator for random coefficients logit demand models using aggregate data," *The RAND Journal of Economics*, 46(1), 86–102.

LEE, S. (2007): "Endogeneity in quantile regression models: A control function approach," *Journal of Econometrics*, 141(2), pp. 1131–1158.

LI, S., AND T. BASAR (1987): "Distributed algorithms for the computation of noncooperative equilibria," *Automatica*, 23(4), 523 – 533.

MA, L., AND R. KOENKER (2006): "Quantile regression methods for recursive structural equation models," *Journal of Econometrics*, 134(2), pp. 471 – 506.

MARRA, G., AND R. RADICE (2013): "Estimation of a regression spline sample selection model," *Computational Statistics & Data Analysis*, 61, 158 – 173.

MELLY, B., AND K. WÜTHRICH (2017): "Local quantile treatment effects," in *Handbook of Quantile Regression*, ed. by V. Chernozhukov, X. He, R. Koenker, and L. Peng, pp. pp. 145–164. CRC Chapman-Hall.

MOON, H. R., AND M. WEIDNER (2015): "Linear Regression for Panel With Unknown Number of Factors as Interactive Fixed Effects," *Econometrica*, 83(4), 1543–1579.

PORTNOY, S., AND R. KOENKER (1997): "The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators," *Statistical Science*, 12(4), 279–300.

POTERBA, J. M., S. F. VENTI, AND D. A. WISE (1994): "401(k) Plans and Tax-Deferred Saving," in *Studies in the Economics of Aging*, ed. by D. A. Wise. University of Chicago Press.

——— (1995): "Do 401(k) contributions crowd out other personal saving?," *Journal of Public Economics*, 58(1), pp. 1–32.

——— (1998): "Personal Retirement Saving Programs and Asset Accumulation: Reconciling the Evidence," in *Frontiers in the Economics of Aging*, ed. by D. A. Wise. University of Chicago Press.

R Core Team (2018): *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing, Vienna, Austria.

Rothenberg, T. J. (1971): "Identification in Parametric Models," *Econometrica*, 39(3), 577–591.

Rubin, D. B. (1974): "Estimating Causal Effects of Treatment in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66(5), pp. 688–701.

Sakov, A., and P. J. Bickel (2000): "An Edgeworth expansion for the m out of n bootstrapped median," *Statistics & Probability Letters*, 49(3), 217 – 223.

Smyth, G. K. (1996): "Partitioned algorithms for maximum likelihood and other non-linear estimation," *Statistics and Computing*, 6(3), 201–216.

Torgovitsky, A. (2015): "Identification of Nonseparable Models Using Instruments With Small Support," *Econometrica*, 83(3), pp. 1185–1197.

Van der Vaart, A., and J. Wellner (1996): *Weak Convergence and Empirical Processes: With Application to Statistics.* Springer-Verlag.

Weisberg, S., and A. H. Welsh (1994): "Adapting for the Missing Link," *The Annals of Statistics*, 22(4), 1674–1700.

Wüthrich, K. (2017): "A closed-form estimator for quantile treatment effects with endogeneity," Working Paper, UCSD.

——— (2018): "A comparison of two quantile models with endogeneity," Journal of Business & Economic Statistics.

Zhu, Y. (2018): "k-step correction for mixed integer linear programming: a new approach for instrumental variable quantile regressions and related problems," arXiv:1805.06855.

Appendix A. Overidentification

In the main text, we focus on just-identified moment restrictions with $d_Z = d_D$, for which the construction of an estimator is straightforward. If the model is overidentified (i.e. if $d_Z > d_D$), we can transform the original moment conditions

$$E_P\left[(1\left\{Y \leq (X', D')\theta(\tau)\right\} - \tau)\begin{pmatrix} X \\ Z \end{pmatrix}\right] = 0$$

into a set of just-identified moment conditions

$$E_P\left[(1\left\{Y \leq X'\theta_1(\tau) + D_1\theta_2(\tau) + \cdots + D_{d_D}\theta_J(\tau)\right\} - \tau)\begin{pmatrix} X \\ \tilde{Z} \end{pmatrix}\right] = 0, \tag{A.1}$$

where $\tilde{Z}$ is a $d_D \times 1$ vector of transformations of (X, Z). A practical choice is to construct $\tilde{Z}$ using a least squares projection of D on Z and X.

To achieve pointwise (in τ) efficiency, we can employ the following two-step procedure (e.g., Chernozhukov and Hansen, 2006, Remark 5):

Step 1: We first obtain an initial consistent estimate of θ^* using one of our estimators based on a set of just-identified moment conditions such as (A.1). We then use nonparametric estimators to estimate the conditional densities $V(\tau) = f_{\varepsilon(\tau)|X,Z}(0)$ and $v(\tau) = f_{\varepsilon(\tau)|D,X,Z}(0)$, where $\varepsilon(\tau) = Y_i - X_i'\theta_1^*(\tau) - D_1\theta_2^*(\tau) - \cdots - D_{d_D}\theta_J^*(\tau)$, and the conditional expectation function $E_P\left[Dv(\tau) \mid X, Z\right]$.

Step 2: We apply our procedure to obtain a solution to following moment conditions:

$$E_P\left[(1\left\{Y \leq X'\theta_1(\tau) + D_1\theta_2(\tau) + \cdots + D_{d_D}\theta_J(\tau)\right\} - \tau)\begin{pmatrix} V(\tau)X \\ E_P\left[Dv(\tau) \mid X, Z\right] \end{pmatrix}\right] = 0.$$

This can be achieved by defining the BR maps as follows:

$$\begin{aligned} L_1\left(\theta_{-1}(\tau)\right) &:= \arg\min_{\tilde{\theta}_1 \in \mathbb{R}^{d_X}} Q_{P,1}\left(\tilde{\theta}_1, \theta_{-1}\right) \\ L_j\left(\theta_{-j}(\tau)\right) &:= \arg\min_{\tilde{\theta}_j \in \mathbb{R}} Q_{P,j}\left(\tilde{\theta}_j, \theta_{-j}\right), \ j = 2, \ldots, J, \end{aligned}$$

where

$$\begin{aligned} Q_{P,1}\left(\theta(\tau)\right) &:= E_P\left[\rho_\tau(Y - X'\theta_1(\tau) - D_1\theta_2(\tau) - \cdots - D_{d_D}\theta_J(\tau))V(\tau)\right], \\ Q_{P,j}\left(\theta(\tau)\right) &:= E_P\left[\rho_\tau(Y - X'\theta_1(\tau) - D_1\theta_2(\tau) - \cdots - D_{d_D}\theta_J(\tau))(E_P\left[Dv(\tau) \mid X, Z\right]_{j-1}/D_{j-1})\right], \ j = 2, \ldots, J, \end{aligned}$$

where $E_P\left[Dv(\tau) \mid X, Z\right]_{j-1}$ is the j-th element of $E_P\left[Dv(\tau) \mid X, Z\right]$. These are convex population QR problems provided that the model is parametrized such that $E_P\left[Dv(\tau) \mid X, Z\right]_{j-1}/D_{j-1}$, $j = 2, \ldots, J$, is positive. Estimation can then proceed by replacing the population QR problems by their sample analogues and applying one of the estimation algorithms discussed in the main text. The resulting estimator uses the optimal instrumental variables and thus achieves pointwise (in τ) efficiency (Chamberlain, 1987).

Appendix B. Reparametrization

In the main text, we assume that the model is reparametrized such that Z_ℓ/D_ℓ is positive for all $\ell = 1, \dots, d_D$. This ensures that the weights are well-defined and that the weighted QR problems are convex. However, in empirical applications, the weights may not be well-defined (e.g., if D_ℓ is an indicator variable with $P(D_\ell = 0) > 0$) or negative in some instances. Assuming that Z_ℓ is positive, a simple way to reparametrize the model is to add a large enough constant c to D_ℓ.[21] This transformation is theoretically justified by the compactness of the support of D_ℓ (Assumption 2.2). To fix ideas, suppose that one is interested in estimating the following linear-in-parameters model with a single endogenous variable:

$$q(D, X, \tau) = \theta_{11} + \tilde{X}'\theta_{12} + D\theta_2,$$

where $\theta_1 = (\theta_{11}, \theta_{12}')'$ and $X = \left(1, \tilde{X}'\right)'$. Suppose further that the support of D is a compact interval, $[d_{\min}, d_{\max}] \subset \mathbb{R}$, with $d_{\min} < 0$. In this case, we can apply the transformation $D^\star = D + c$, where $c > |d_{\min}|$. The transformed model reads

$$q(D, X, \tau) = \theta_{11}^\star + \tilde{X}'\theta_{12} + D^\star\theta_2,$$

where $\theta_{11}^\star = \theta_{11} - c\theta_2$. Importantly, one can always back out the original parameters, $\theta = (\theta_{11}, \theta_{12}', \theta_2)'$, from the parameters in the reparametrized model, $\theta^\star = (\theta_{11}^\star, \theta_{12}', \theta_2)'$.

Appendix C. Decentralization

C.1. **The domains of M_j-maps.** Recall that we defined the set

$$\begin{aligned}\tilde{R}_1 := \big\{\theta_{-1} \in \Theta_{-1} : &\Psi_{P,1}(\theta_1, \theta_{-1}) = 0,\\ &\Psi_{P,2}(\theta_1, \theta_2, \pi_{-\{1,2\}}\theta_{-1}) = 0,\ \exists(\theta_1, \theta_2) \in \Theta_1 \times \Theta_2\big\}.\end{aligned}$$

Similarly for $k = 2, \dots, d_D - 1$, define

$$\begin{aligned}\tilde{R}_k := \big\{\theta_{-1} \in \Theta_{-1} : &\Psi_{P,1}(\theta_1, \theta_{-1}) = 0,\\ &\Psi_{P,2}(\theta_1, \theta_2, \pi_{-\{1,2\}}\theta_{-1}) = 0,\\ &\vdots\\ &\Psi_{P,k}(\theta_1, \dots, \theta_k, \pi_{-\{1,\dots,k\}}\theta_{-1}) = 0,\ \exists(\theta_1, \dots, \theta_k) \in \prod_{j=1}^{k} \Theta_j\big\}.\end{aligned}$$

[21] Since the unconditional moment conditions Ψ_P are derived from a conditional moment restriction, we can use a positive transformation of Z_ℓ instead of Z_ℓ itself in case Z_ℓ is not positive.

For $k = d_D$, let

$$\begin{aligned}\tilde{R}_{d_D} := \big\{\theta_{-1} \in \Theta_{-1} :& \Psi_{P,1}(\theta_1, \theta_{-1}) = 0,\\ & \Psi_{P,2}(\theta_1, \theta_2, \pi_{-\{1,2\}}\theta_{-1}) = 0,\\ & \vdots\\ & \Psi_{P,J}(\theta_1, \ldots, \theta_J) = 0,\ \exists(\theta_1, \ldots, \theta_J) \in \prod_{j=1}^{J} \Theta_j\big\}.\end{aligned}$$

Note that $\tilde{R}_{d_D} \subset \tilde{R}_j$ for all $j \leq d_D$.

C.2. **Local Decentralization and Local Contractions.** We say that an estimation problem admits *local decentralization* if the BR functions L_j, $j = 1, \ldots, J$, and the maps K and M are well-defined over a local neighborhood of θ^*. The following weak conditions are sufficient for local decentralization of the IVQR estimation problem.

Assumption 4. *The following conditions hold.*

(1) The conditional cdf $y \mapsto F_{Y|D,X,Z}(y)$ is continuously differentiable at $y^* = d'\theta^*_{-1} + x'\theta^*_1$ for almost all (d, x, z). The conditional density $f_{Y|D,Z,X}$ is bounded on a neighborhood of y^* a.s.;

(2) The matrices

$$E_P[f_{Y|D,X,Z}\left(D'\theta^*_{-1} + X'\theta^*_1\right) XX']$$

and

$$E_P[f_{Y|D,X,Z}\left(D'\theta^*_{-1} + X'\theta^*_1\right) D_\ell Z_\ell], \quad \ell = 1, \ldots, d_D,$$

are positive definite.

Assumption 4 is weaker than Assumption 2.3–2.4. Under this condition, we can study the local properties of our population algorithms. For this, the following lemma ensures that the BR maps are well-defined locally.

Lemma 3. *Suppose that Assumptions 1, 2.1–2.2, and 4 hold. Then, there exist open neighborhoods $\mathcal{N}_{L_{-j}}, j = 1 \ldots J$, $\mathcal{N}_K$, $\mathcal{N}_M$ of θ^*_{-j}, θ^*, and θ^*_{-1} such that*

(i) There exist maps $L_j : \mathcal{N}_{-j} \to \mathbb{R}^{d_j}$, $j = 1, \ldots, J$ such that, for $j = 1, \ldots, J$,

$$\Psi_{P,j}\left(L_j(\theta_{-j}), \theta_{-j}\right) = 0, \quad \textit{for all } \theta_{-j} \in \mathcal{N}_{-j}$$

Further, L_j is continuously differentiable for all $j = 1, \ldots, J$.

(ii) The maps $K : \mathcal{N}_K \to \mathbb{R}^d$ and $M : \mathcal{N}_M \to \mathbb{R}^{d_D}$ are continuously differentiable.

Proof. (i) The proof is similar to that of Lemma 1. Therefore, we sketch the argument below for $j = 1$. By Assumptions 2.2 and 4.1, $\Psi_{P,1}$ is continuously differentiable on a neighborhood V of θ^*. By Assumption 4.2 and the continuity of $\det(\partial\Psi_{P,1}(\theta)/\partial\theta_1')$, one may choose V so that $\det(\partial\Psi_{P,1}(\theta)/\partial\theta_1') \neq 0$ for all $\theta = (\theta_1, \theta_{-1}) \in V$. By the implicit function theorem, there is a continuously differentiable function L_1 and an open set $\mathcal{N}_{-1}$ containing θ_{-1} such that

$$\Psi_{P,1}(L_1(\theta_{-1}), \theta_{-1}) = 0, \text{ for all } \theta_{-1} \in \mathcal{N}_{-1}.$$

The arguments for L_j, $j \neq 1$ are similar.

(ii) Let $\mathcal{N}_K = \{\theta \in \Theta : \pi_{-j}\theta \in \mathcal{N}_{-j},\ j = 1, \dots, J\}$ and let $\mathcal{N}_M$ be defined by mimicking (3.13), while replacing Θ_j with $\mathcal{N}_j$ in the definition of $\tilde{R}_j$ for $j = 1, \dots, J$. The continuous differentiability of K and M follows from that of L_j, $j = 1, \dots, J$. □

C.2.1. *Local Contractions.* The following assumption ensures that K and M are local contractions.

Assumption 5.

(1) $\rho\left(J_K\left(\theta^*\right)\right) < 1$*;*

(2) $\rho\left(J_M\left(\theta_2^*\right)\right) < 1$

Here, we illustrate a primitive condition for Assumption 5. Consider a simple setup without covariates (i.e. $X = 1$), a binary D, and a binary Z. We only analyze Assumption 5.1. A similar result can be derived for Assumption 5.2. In this setting, the Jacobian of K evaluated at θ^* is given by

$$J_K(\theta^*) = \begin{pmatrix} 0 & -\frac{E_P\left[f_{Y|D,Z}(D\theta_2^*+\theta_1^*)D\right]}{E_P\left[f_{Y|D,Z}\left(D\theta_2^*+\theta_1^*\right)\right]} \\ -\frac{E_P\left[f_{Y|D,Z}(D\theta_2^*+\theta_1^*)Z\right]}{E_P\left[f_{Y|D,Z}\left(D\theta_2^*+\theta_1^*\right)ZD\right]} & 0 \end{pmatrix}.$$

The characteristic polynomial is then given by

$$p_K(\lambda) \quad = \quad \lambda^2 - \frac{E_P\left[f_{Y|D,Z}\left(D\theta_2^*+\theta_1^*\right)D\right]}{E_P\left[f_{Y|D,Z}\left(D\theta_2^*+\theta_1^*\right)\right]}\frac{E_P\left[f_{Y|D,Z}\left(D\theta_2^*+\theta_1^*\right)Z\right]}{E_P\left[f_{Y|D,Z}\left(D\theta_2^*+\theta_1^*\right)ZD\right]}.$$

Hence, Assumption 3.1 holds if all eigenvalues (i.e. the roots λ_K of $p_K(\lambda) = 0$) have modulus less than one, which holds when

$$\left|\frac{E_P\left[f_{Y|D,Z}\left(D\theta_2^*+\theta_1^*\right)D\right]}{E_P\left[f_{Y|D,Z}\left(D\theta_2^*+\theta_1^*\right)\right]}\frac{E_P\left[f_{Y|D,Z}\left(D\theta_2^*+\theta_1^*\right)Z\right]}{E_P\left[f_{Y|D,Z}\left(D\theta_2^*+\theta_1^*\right)ZD\right]}\right| < 1.$$

This condition can be simplified to

$$f_{Y|0,1}(\theta_1^*)p(0|1)f_{Y|1,0}(\theta_2^*+\theta_1^*)p(1|0) < f_{Y|1,1}(\theta_2^*+\theta_1^*)p(1|1)f_{Y|0,0}(\theta_1^*)p(0|0), \tag{C.1}$$

where $f_{Y|d,z}(y) := f_{Y|D=d,Z=z}(y)$ and $p(d|z) := P(D = d \mid Z = z)$. It is instructive to interpret condition (C.1) under the local average treatment effects framework of Imbens and Angrist (1994). Specifically, condition (C.1) holds (i) if their monotonicity assumption is such that there are compliers but no defiers and (ii) the complier potential outcome density functions are strictly positive. Conversely, the condition is violated if there are defiers but no compliers.

Proposition 3. *Suppose that Assumptions 1, 2.1, 2.2, 4, and 5 hold. Then:*

(i) There exists a closed neighborhood $\bar{\mathcal{N}}_K$ of θ^* such that $K(\bar{\mathcal{N}}_K) \subset \bar{\mathcal{N}}_K$ and K is a contraction on $\bar{\mathcal{N}}_K$ with respect to an adapted norm.

(ii) There exists a closed neighborhood $\bar{\mathcal{N}}_M$ of θ_2^* such that $M(\bar{\mathcal{N}}_K) \subset \bar{\mathcal{N}}_M$ and M is a contraction on $\bar{\mathcal{N}}_M$ with respect to an adapted norm.

Proof. We only prove the result for K, the proof for M is similar. By Lemma 3, L_j is continuously differentiable at θ^*. Note that J_K is given by

$$J_K(\theta) = \begin{bmatrix} 0 & \frac{\partial L_1(\theta_{-1})}{\partial \theta_2'} & \cdots & \cdots & \frac{\partial L_1(\theta_{-1})}{\partial \theta_J'} \\ \frac{\partial L_2(\theta_{-2})}{\partial \theta_1'} & 0 & \frac{\partial L_2(\theta_{-2})}{\partial \theta_3'} & \cdots & \frac{\partial L_2(\theta_{-2})}{\partial \theta_J'} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\partial L_J(\theta_{-J})}{\partial \theta_1'} & \cdots & \cdots & \frac{\partial L_J(\theta_{-J})}{\partial \theta_{J-1}'} & 0 \end{bmatrix}, \tag{C.2}$$

which is continuous at θ^*. The desired result now follows, for instance, from Proposition 2.2.19 in (Hasselblatt and Katok, 2003). □

Appendix D. Proofs of Theoretical Results in Section 3

Proof of Lemma 1. (i) We first show that L_1 is well-defined. For a given $\theta_{-1} \in \mathbb{R}^{d-d_X}$, let $\theta_1^* \in \arg\min_{\tilde{\theta}_1 \in \mathbb{R}^{d_X}} Q_{P,1}(\tilde{\theta}_1, \theta_{-1})$. Under Assumption 2, the objective function is convex and differentiable with respect to $\tilde{\theta}_1$. Therefore, by the necessary and sufficient condition of minimization, θ_1^* solves

$$E_P[(1\{Y \le D'\theta_{-1} + X'\theta_1^*\})X] = 0.$$

In what follows, we show that the map $L_1 : \theta_{-1} \mapsto \theta_1^*$ is well-defined on R_{-1} using a global inverse function theorem. Recall that

$$\Psi_{P,1}(\theta) = E_P[(1\{Y \le D'\theta_{-1} + X'\theta_1\})X]. \tag{D.1}$$

This function is continuously differentiable with respect to θ. The Jacobian is given by

$$J_{\Psi_{P,1}}(\theta) = \frac{\partial}{\partial \theta'} E_P[F_{Y|D,X,Z}(D'\theta_{-1} + X'\theta_1)X] = E_P[f_{Y|D,X,Z}(D'\theta_{-1} + X'\theta_1)X(X', D')], \tag{D.2}$$

where the second equality follows from Assumption 2 and the dominated convergence theorem. Define a transform $\Xi : \Theta \to \mathbb{R}^d$ by

$$\Xi(\theta) := (\Psi_{P,1}(\theta)', \theta_{-1}{}')'. \tag{D.3}$$

We follow Krantz and Parks (2012) (Section 3.3) to obtain an implicit function L_1 on a suitable domain such that $\theta_1 = L_1(\theta_2)$ if and only if $\Psi_{P,1}(\theta) = 0$. The key is to apply a global inverse function theorem to Ξ. Toward this end, we analyze the Jacobian of Ξ, which is given as

$$\begin{aligned} J_\Xi(\theta) &= \begin{bmatrix} \partial\Psi_{P,1}(\theta_1,\theta_{-1})/\partial\theta_1' & \partial\Psi_{P,1}(\theta_1,\theta_{-1})/\partial\theta_{-1}' \\ 0_{d_{-1}\times d_1} & I_{d_{-1}} \end{bmatrix} \\ &= \begin{bmatrix} E_P[f_{Y|D,X,Z}(D'\theta_{-1}+X'\theta_1)XX'] & E_P[f_{Y|D,X,Z}(D'\theta_{-1}+X'\theta_1)XD'] \\ 0_{d_{-1}\times d_1} & I_{d_{-1}} \end{bmatrix}, \end{aligned} \tag{D.4}$$

where I_d denotes the $d \times d$ identity matrix. Let $I \subset \{1,\dots d\}$.

For any matrix A, let $[A]_{I,I}$ denote a principal minor of A, which collects the rows and columns of A whose indices belong to the index set I. By (D.4), if $I \subset \{1,\dots,d_1\}$,

$$[J_\Xi(\theta)]_{I,I} = E_P[f_{Y|D,X,Z}(D'\theta_{-1}+X'\theta_1)\tilde{X}\tilde{X}'] \tag{D.5}$$

for a subvector $\tilde{X}$ of X, which is positive definite by Assumption 2 and Lemma 4. If $I \subset \{d_1+1,\dots,d\}$, $[J_\Xi(\theta)]_{I,I} = I_\ell$ for some $1 \le \ell \le d-d_1$ and is hence positive definite. Otherwise, any principal minor is of the following form:

$$[J_\Xi(\theta)]_{I,I} = \begin{bmatrix} E_P[f_{Y|D,X,Z}(D'\theta_{-1}+X'\theta_1)\tilde{X}\tilde{X}'] & B \\ 0_{\ell\times m} & I_\ell \end{bmatrix} \tag{D.6}$$

for some subvector $\tilde{X}$ of X and a $m \times \ell$ matrix B. Note that

$$\begin{aligned} \det([J_\Xi(\theta)]_{I,I}) &= \det(E_P[f_{Y|D,X,Z}(D'\theta_{-1}+X'\theta_1)\tilde{X}\tilde{X}'] - BI_\ell^{-1}\times 0_{\ell\times m})\det(I_\ell) \\ &= \det(E_P[f_{Y|D,X,Z}(D'\theta_{-1}+X'\theta_1)\tilde{X}\tilde{X}']) > 0, \end{aligned} \tag{D.7}$$

where the last inequality follows again from Assumption 2 and Lemma 4. Hence, $J_\Xi(\theta)$ is a P-matrix. Note that Θ is a closed rectangle. By Theorem 4 in Gale and Nikaido (1965), Ξ is univalent, and hence the inverse map Ξ^{-1} is well defined.

Let

$$R_{-1} = \{\theta_{-1}\in\mathbb{R}^{d_{-1}} : (0,\theta_{-1}) \in \Xi(\Theta)\} = \{\theta_{-1}\in\mathbb{R}^{d_{-1}} : \Psi_{P,1}(\theta_1,\theta_{-1}) = 0, \text{ for some } (\theta_1,\theta_{-1})\in\Theta\},$$

which coincides with the definition in (3.8) with $j=1$. Let $F_1 = [I_{d_1}, 0_{d_1\times d_{-1}}]$. For each $\theta_{-1}\in R_{-1}$, define

$$L_1(\theta_{-1}) := F_1\Xi^{-1}(0,\theta_{-1}).$$

Then, for any $\theta\in\Theta$, $\Psi_{P,1}(\theta) = 0$ if and only if $\theta_{-1}\in R_{-1}$ and $\Xi(\theta) = (0,\theta_{-1})$. By the univalence of Ξ, this is true if and only if $\theta = \Xi^{-1}(0,\theta_{-1})$, and the first d_1 components extracted by applying F_1 is θ_1. This ensures L_1 is well-defined on R_{-1}.

Below, for any set A, let A^o denote the interior of A. Let $R^o_{-1} = \{\theta_{-1} \in \mathbb{R}^{d_{-1}} : (0, \theta_{-1}) \in \Xi(\Theta^o)\}$. Note that $\Psi_{P,1}$ is $\mathcal{C}^1$ on Θ^o and, for each $\theta = (\theta_1, \theta_{-1}) \in \Theta$ with $\theta_{-1} \in R^o_{d_{-1}}$, $\det(\partial \Psi_{P,1}(\theta)/\partial \theta_1') \neq 0$. Therefore, by the implicit function theorem, there is a $\mathcal{C}^1$-function $\tilde{L}_1$ and an open set V containing θ_{-1} such that

$$\Psi_{P,1}(\tilde{L}_1(\theta_{-1}), \theta_{-1}) = 0, \text{ for all } \theta_{-1} \in V.$$

However, such a local implicit function must coincide with the unique global map L_1 on V. Hence, $L_1|_V = \tilde{L}_1$, and therefore L_1 is continuously differentiable at θ_{-1}. Since the choice of θ_{-1} is arbitrary, L_1 is continuously differentiable for all $\theta_{-1} \in R^o_2$.

Showing that the conclusion holds for any other L_j for $j = 2, \ldots, J$ is similar, and hence we omit the proof. □

Lemma 4. *Suppose $E_P[f_{Y|D,X,Z}(D'\theta_{-1} + X'\theta_1) XX']$ is positive definite. Then, for any subvector $\tilde{X}$ of X with dimension $\tilde{d}_X \leq d_X$, $E_P[f_{Y|D,X,Z}(D'\theta_{-1} + X'\theta_1) \tilde{X}\tilde{X}']$ is positive definite.*

Proof. In what follows, let $W = f_{Y|D,X,Z}(D'\theta_{-1} + X'\theta_1)$ and let

$$A := E_P[f_{Y|D,X,Z}(D'\theta_{-1} + X'\theta_1) XX'] = E[WXX']. \tag{D.8}$$

Let $\tilde{X}$ be a subvector of X with $\tilde{d}_X$ components. Then, there exists a $d_X \times d_X$ permutation matrix P_π such that the first $\tilde{d}_X$ components of $P_\pi X$ is $\tilde{X}$.

Let $B := E[WP_\pi XX'P_\pi']$ and note that

$$B = P_\pi E[WXX']P_\pi' = P_\pi A P_\pi', \tag{D.9}$$

by the linearity of the expectation operator and W being a scalar. Let λ be an eigenvalue of B such that

$$Bz = \lambda z, \tag{D.10}$$

for the corresponding eigenvector $z \in \mathbb{R}^{d_X}$. By (D.9)-(D.10),

$$P_\pi A P_\pi' z = \lambda z \iff A P_\pi' z = \lambda P_\pi^{-1} z. \tag{D.11}$$

Note that $P_\pi^{-1} = P_\pi'$ due to P_π being a permutation matrix. Letting $y := P_\pi' z$ then yields

$$Ay = \lambda y, \tag{D.12}$$

which in turn shows that λ is an eigenvalue of A. For any eigenvalue of A, the argument above can be reversed to show that it is also an eigenvalue of B. Since the choice of the eigenvalue is arbitrary, A and B share the same eigenvalues.

Now let $C := E[W\tilde{X}\tilde{X}']$ and note that it is a leading principal submatrix of B. Then, by the eigenvalue inclusion principle (Horn and Johnson, 1990, Theorem 4.3.28),

$$\lambda_{\min}(C) \geq \lambda_{\min}(B) = \lambda_{\min}(A) > 0, \tag{D.13}$$

where the last inequality follows from the positive definiteness of A. This completes the claim of the lemma. □

Proof of Corollary 1. The existence of K and its continuous differentiability follows immediately from Lemma 1. For M, by the definition of $\tilde{R}_1$, for any $\theta_{-1} \in \tilde{R}_j$, there exists $(\theta_1, \theta_2) \in \Theta_1 \times \Theta_2$ such that

$$\Psi_{P,1}(\theta_1, \theta_{-1}) = 0, \tag{D.14}$$

$$\Psi_{P,2}(\theta_1, \theta_2, \pi_{-\{1,2\}}\theta_{-1}) = 0, \tag{D.15}$$

By (i), one may then write $\theta_1 = L_1(\theta_{-1})$ and $\theta_2 = L_2(L_1(\theta_{-1}), \pi_{-\{1,2\}}\theta_{-1})$. Hence, the map $M_1 : \tilde{R}_1 \to \Theta_2$ defined below is well-

$$M_1(\theta_{-1}) = L_2\big(L_1(\theta_{-1}), \pi_{-\{1,2\}}\theta_{-1}\big). \tag{D.16}$$

Recursively, arguing in the same way, the maps

$$\begin{aligned}
M_2(\theta_{-1}) &= L_3\big(L_1(\theta_{-1}), M_1(\theta_{-1}), \pi_{-\{1,2,3\}}\theta_{-1}\big) && \text{(D.17)}\\
&\vdots\\
M_j(\theta_{-1}) &= L_{j+1}\big(L_1(\theta_{-1}), M_1(\theta_{-1}), \ldots, M_{j-1}(\theta_{-1}), \pi_{-\{1,\ldots,j+1\}}\theta_{-1}\big) && \text{(D.18)}\\
&\vdots\\
M_{d_D}(\theta_{-1}) &= L_J\big(L_1(\theta_{-1}), M_1(\theta_{-1}), \ldots, M_{d_D-1}(\theta_{-1})\big) && \text{(D.19)}
\end{aligned}$$

are well-defined on $\tilde{R}_2, \cdots, \tilde{R}_{d_D}$ respectively. The continuous differentiability of M follows from that of L_js and the chain rule. □

Proof of Proposition 1. $\Rightarrow$: For every solution, $\Psi_P(\theta^*) = 0$, $\theta^*_j = L_j\left(\theta^*_{-j}\right)$ by construction under Assumptions 1 and 2. It follows that $K\left(\theta^*\right) = \theta^*$ and $M\left(\theta^*_{-1}\right) = \theta^*_{-1}$.

$\Leftarrow$: For the simultaneous response note that $K\left(\bar{\theta}\right) = \bar{\theta}$ implies that $\bar{\theta}_j = L_j\left(\bar{\theta}_{-j}\right)$ for all $j \in \{1, \cdots, J\}$. Thus, $\bar{\theta}$ solves $\Psi_P(\bar{\theta}) = 0$ by Lemma 1. Consider next the sequential response. Let $\tilde{\theta}, \bar{\theta} \in \Theta$ be such that $\tilde{\theta}_j = L_j(\bar{\theta}_{-j})$ for $j = 1, \ldots, J$. By Lemma 1, they satisfy

$$\begin{aligned}
\Psi_{P,1}\left(\tilde{\theta}_1, \bar{\theta}_2, \cdots, \bar{\theta}_J\right) &= 0\\
\Psi_{P,2}\left(\tilde{\theta}_1, \tilde{\theta}_2, \cdots, \bar{\theta}_J\right) &= 0\\
&\vdots\\
\Psi_{P,J}\left(\tilde{\theta}_1, \tilde{\theta}_2, \cdots, \tilde{\theta}_J\right) &= 0
\end{aligned}$$

Thus, a fixed point $\tilde{\theta} = \bar{\theta}$ satisfies $\Psi_P\left(\bar{\theta}\right) = 0$. □

Appendix E. Proofs of Theoretical Results in Section 4

Proof of Proposition 2. We prove the result for K. By Assumption 3, there exists a strictly convex set $\tilde{D}_K$ on which the spectral norm of the Jacobian of K is uniformly bounded by 1. This ensures that K is a contraction map on $cl(\tilde{D}_K)$, and the claim of the proposition now follows from Theorem 2.2.16 in Hasselblatt and Katok (2003). □

Appendix F. Proofs of Theoretical Results in Section 6

Proof of Theorem 1. Let $H := I_d - K$. A fixed point θ^* of K then satisfies

$$H(\theta^*) = 0.$$

Similarly, let $\hat{H} := I_d - \hat{K}$. The estimator $\hat{\theta}$ satisfies

$$\|\hat{H}(\hat{\theta})\| \le \inf_{\theta' \in \Theta} \|\hat{H}(\theta)\| + r_N, \tag{F.1}$$

where $r_N = o_p(N^{-1/2})$. Let $\varphi : \ell^\infty(\Theta)^d \times \mathbb{R} \to \mathbb{R}^d$ be a map such that, for each $(H, r) \in \ell^\infty(\Theta)^d \times \mathbb{R}$, $\tilde{\theta} = \varphi(H, r)$ is an r-approximate solution, which satisfies

$$\|H(\tilde{\theta})\| \le \inf_{\theta' \in \Theta} \|H(\theta')\| + r. \tag{F.2}$$

One may then write

$$\sqrt{N}(\hat{\theta}_N - \theta^*) = \sqrt{N}(\varphi(\hat{H}, \hat{r}) - \varphi(H, 0)). \tag{F.3}$$

By Corollary 2, $\sqrt{N}(\hat{K} - K) \rightsquigarrow \mathbb{W}$ in $\ell^\infty(\Theta)^d$, where $\mathbb{W}$ is a Gaussian process defined in Corollary 2. By Lemmas 6-7, Condition Z in CFM holds, which in turn ensures that one may apply Lemmas E.2 and E.3 in CFM. This ensures

$$\sqrt{N}(\varphi(\hat{H}, \hat{r}) - \varphi(H, 0)) \rightsquigarrow \varphi'_{H,0}(\mathbb{W}, 0) = -\dot{H}_{\theta^*}^{-1}\mathbb{W}(\theta^*). \tag{F.4}$$

Hence, we obtain (6.5) with

$$V = \dot{H}_{\theta^*}^{-1} E[\mathbb{W}(\theta^*)\mathbb{W}(\theta^*)']\dot{H}_{\theta^*}^{-1}. \tag{F.5}$$

Finally, note that $\dot{H}_{\theta^*} = I_d - J_K(\theta^*)$ by Lemma 7. This establishes the theorem. □

Proof of Theorem 2. Recall that $\hat{H} = I_d - \hat{K}$. The estimator $\hat{\theta}_N$ satisfies

$$\|\hat{H}(\hat{\theta}_N)\|^2 \le \inf_{\theta' \in \Theta} \|\hat{H}(\theta')\|^2 + r_N, \tag{F.6}$$

where $r_N = o_p(N^{-1/2})$. Similarly, let $\hat{H}^* = I_d - \hat{K}^*$. Let P^* denote the law of $\hat{H}^*$ conditional on $\{W_i\}_{i=1}^\infty$. The bootstrap estimator $\hat{\theta}_N^*$ satisfies

$$\|\hat{H}^*(\hat{\theta}_N^*)\|^2 \le \inf_{\theta' \in \Theta} \|\hat{H}^*(\theta')\|^2 + r_N^*, \tag{F.7}$$

where $r_N^* = o_{P^*}(N^{-1/2})$ conditional on $\{W_i\}_{i=1}^\infty$.

Using the r-approximation, one may therefore write

$$\sqrt{N}(\hat\theta_N^* - \hat\theta_N) = \sqrt{N}(\varphi(\hat H^*, r_N^*) - \varphi(\hat H, r_N)). \tag{F.8}$$

Let E_{P^*} denote the conditional expectation with respect to P^*. Let BL_1 denote the space of bounded Lipschitz functions on $\mathbb{R}^d$ with Lipschitz constant 1. Then, for any $\epsilon > 0$,

$$\begin{aligned}\sup_{h\in BL_1}\Big|E_{P^*}h\big(\sqrt{N}\big[\varphi(\hat H^*, r_N^*) - \varphi(\hat H, r_N)\big]\big) - E_{P^*}h\big(\varphi'_{H,0}\big(\sqrt{N}\big[(\hat H^*, r_N^*)' - (\hat H, r_N)'\big]\big)\big)\Big|& \\ \le \epsilon + 2P^*\Big(\big\|\sqrt{N}\big[\varphi(\hat H^*, r_N^*) - \varphi(\hat H, r_N)\big] - \varphi'_{H,0}\big(\sqrt{N}\big[(\hat H^*, r_N^*) - (\hat H, r_N)\big]\big)\big\| > \epsilon\Big).& \quad \text{(F.9)}\end{aligned}$$

By Corollary 2, $\sqrt{N}(\hat H^* - \hat H) = -\sqrt{N}(\hat K^* - \hat K) \overset{L^*}{\rightsquigarrow} -\mathbb{W} \overset{d}{=} \mathbb{W}$. Noting that $h\circ\varphi'_{H,0} \in BL_1(\ell^\infty(\Theta)\times\mathbb{R})$ and $r_N = o_p(N^{-1/2})$, it follows that

$$\sup_{h\in BL_1}\Big|E_{P^*}h\big(\varphi'_{H,0}\big(\sqrt{N}\big[(\hat H^*, r_N^*) - (\hat H, r_N)\big]\big)\big) - E_{P^*}h\circ\varphi'_{H,0}(\mathbb{W}, 0)\Big| \to 0, \tag{F.10}$$

with probability approaching 1 due to $r_N = o_P(N^{-1/2})$. Hence, for the conclusion of the theorem, it suffices to show that the right hand side of (F.9) tends to 0 in probability.

For this, as shown in the proof of Theorem 1, φ is Hadamard differentiable at $(H, 0)$. Hence, by Theorem 3.9.4 in Van der Vaart and Wellner (1996),

$$\begin{aligned}\sqrt{N}\big[\varphi(\hat H^*, r_N^*) - \varphi(H, 0)\big] &= \varphi'_{H,0}(\sqrt{N}[(\hat H^*, r_N^*) - (H, 0)]) + o_{P^*}(1)\\ \sqrt{N}\big[\varphi(\hat H, r_N) - \varphi(H, 0)\big] &= \varphi'_{H,0}(\sqrt{N}[(\hat H, r_N) - (H, 0)]) + o_P(1),\end{aligned}$$

Take the difference of the left and right hand sides respectively and note that $\varphi'_{H,0}$ is linear. This implies the right hand side of (F.9) tends to 0 in probability. This ensures

$$\sqrt{N}(\varphi(\hat H, r_N^*) - \varphi(\hat H, r_N)) \overset{L^*}{\rightsquigarrow} \varphi'_{H,0}(\mathbb{W}, 0) = -\dot H_{\theta^*}^{-1}\mathbb{W}(\theta^*). \tag{F.11}$$

□

Lemma 5. *Suppose Assumptions 1-2 hold. (i) Let $\hat\theta_N$ be an estimator of θ^* that satisfies (6.1). Then, it also satisfies (6.2)-(6.3); (ii) Let $\hat\theta_N$ be an estimator of θ^* that satisfies (6.2)-(6.3). Then, it also satisfies (6.1).*

Proof. (i) Consider the case $j = 2$. Note that, by (6.1),

$$\begin{aligned}\hat\theta_{N,2} - \hat L_2(\hat L_1(\hat\theta_{N,-1}), \hat\theta_{N,3}, \dots, \hat\theta_{N,J}) &= \hat\theta_{N,2} - \hat L_2(\hat\theta_{N,1} + r_{N,1}, \hat\theta_{N,3}, \dots, \hat\theta_{N,J}) & \text{(F.12)}\\ &= \hat L_2(\hat\theta_{N,1}, \hat\theta_{N,3}, \dots, \hat\theta_{N,J}) - \hat L_2(\hat\theta_{N,1} + r_{N,1}, \hat\theta_{N,3}, \dots, \hat\theta_{N,J}), & \text{(F.13)}\end{aligned}$$

where $r_{N,1} = o_p(N^{-1/2})$, and the second equality follows from the definition of $\hat{\theta}_{N,2}$. (F.13) can be written as

$$
\begin{aligned}
\hat{L}_2(\hat{\theta}_{N,1}, \hat{\theta}_{N,3}, \ldots, \hat{\theta}_{N,J}) &- \hat{L}_2(\hat{\theta}_{N,1} + r_{N,1}, \hat{\theta}_{N,3}, \ldots, \hat{\theta}_{N,J}) \\
&= \Big([\hat{L}_2(\hat{\theta}_{N,1}, \hat{\theta}_{N,3}, \ldots, \hat{\theta}_{N,J}) - L_2(\hat{\theta}_{N,1}, \hat{\theta}_{N,3}, \ldots, \hat{\theta}_{N,J})] \\
&\quad - [\hat{L}_2(\hat{\theta}_{N,1} + r_{N,1}, \hat{\theta}_{N,3}, \ldots, \hat{\theta}_{N,J}) - L_2(\hat{\theta}_{N,1} + r_{N,1}, \hat{\theta}_{N,3}, \ldots, \hat{\theta}_{N,J})] \Big) \\
&\quad + [L_2(\hat{\theta}_{N,1} + r_{N,1}, \hat{\theta}_{N,3}, \ldots, \hat{\theta}_{N,J}) - L_2(\hat{\theta}_{N,1}, \hat{\theta}_{N,3}, \ldots, \hat{\theta}_{N,J})] \\
&= o_p(N^{-1/2}) + O_P(r_{N,1}), \qquad \text{(F.14)}
\end{aligned}
$$

where the last equality follows from the stochastic equicontinuity of $\mathcal{L}_N$ shown in the proof of Lemma 10 and L_2 being Lipschitz since L_2 is continuously differentiable with a derivative that is uniformly bounded on the compact set Θ. By (F.12)-(F.14), it holds that $\hat{\theta}_{N,j} = M_j(\hat{\theta}_{N,-1}) + o_p(N^{-1/2})$ for $j = 2$. Repeat the same argument sequentially for $j = 3, \ldots, J$. The first conclusion of the lemma then follows.

(ii) Suppose now that $r_{N,1} := \hat{\theta}_{N,1} - \hat{L}_1(\hat{\theta}_{N,-1}) \neq o_P(N^{-1/2})$. Then, there is a subsequence k_N along which, for any $\eta > 0$, $\sqrt{k}_N r_{k_N,1} > \eta$ for all k_N with positive probability. Then, the $O_P(r_{k_N,1})$-term in (F.14) is not $o_p(k_N^{-1/2})$, which therefore implies $\hat{\theta}_{N,j} \neq M_j(\hat{\theta}_{N,-1}) + o_p(N^{-1/2})$ for $j = 2$. The second conclusion of the lemma then follows. □

Lemma 6. *Let $\Lambda \subset \mathbb{R}^p$ be a compact set, and let $K : \Lambda \to \mathbb{R}^p$ be a map that has a unique fixed point $\lambda_0 \in \Lambda$. let $H : \Lambda \to \mathbb{R}^p$ be defined by $H(\lambda) := \lambda - K(\lambda)$. Then $H^{-1}(x) = \{\lambda \in \Lambda : H(\lambda) = x\}$ is continuous at $x = 0$ in Hausdorff distance.*

Proof. For any x, write

$$H^{-1}(x) = \{\lambda : \lambda - K(\lambda) = x\}.$$

Let $x_n \to 0$. Since λ_0 is the unique fixed point of K, $H^{-1}(0) = \{\lambda_0\}$. Therefore,

$$
\begin{aligned}
d_H(H^{-1}(0), H^{-1}(x_n)) &= \max\left\{ \inf_{\lambda \in H^{-1}(x_n)} \|\lambda - \lambda_0\|, \sup_{\lambda \in H^{-1}(x_n)} \|\lambda - \lambda_0\| \right\} \\
&= \sup_{\lambda \in H^{-1}(x_n)} \|\lambda - \lambda_0\|.
\end{aligned}
$$

Hence, it suffices to show that $\sup_{\lambda \in H^{-1}(x_n)} \|\lambda - \lambda_0\| = o(1)$. We show this by contradiction. Suppose that there is a sequence $\{\lambda_n\} \subset \Lambda$ and $\delta > 0$ such that $\lambda_n \in H^{-1}(x_n)$ for all n and $\{\lambda_n\}$ has a subsequence $\{\lambda_{k_n}\}$ such that $\|\lambda_{k_n} - \lambda_0\| > \delta$ for all n. $\lambda_{k_n} \in \Lambda$ is a sequence in a compact space, and hence there is a further subsequence λ_{h_n} such that $\lambda_{h_n} \to \lambda^*$ for some $\lambda^* \in \Lambda$ with $\lambda^* \neq \lambda_0$. By the continuity of K, one then has

$$\lambda_{h_n} - K(\lambda_{h_n}) \to \lambda^* - K(\lambda^*).$$

By $\lambda_{h_n} - K(\lambda_{h_n}) = x_n$ and $x_n \to 0$, it must hold that

$$\lambda^* - K(\lambda^*) = 0.$$

However this contradicts the fact that λ_0 is the unique fixed point, and hence the conclusion follows. □

Lemma 7. *Suppose $H = I - K$ and $K : \mathbb{R}^p \to \mathbb{R}^p$ is continuously differentiable at λ_0. Suppose further that* $\det(I - J_K(\lambda_0)) \neq 0$. *Let $\dot{H}_{\lambda_0} := I - J_K(\lambda_0)$. Then,*

$$\lim_{t \downarrow 0} \sup_{h:\|h\|=1} \|t^{-1}[H(\lambda_0 + th) - H(\lambda_0)] - \dot{H}_{\lambda_0} h\| = 0,$$

and

$$\inf_{h:\|h\|=1} \|\dot{H}_{\lambda_0} h\| > 0.$$

Proof. Let $\{h_n\} \subset \mathbb{S}^p$ be a sequence on the unit sphere. Then,

$$\begin{aligned} t^{-1}[H(\lambda_0 + th_n) - H(\lambda_0)] - \dot{H}_{\lambda_0} h_n &= t^{-1}[\lambda_0 + th_n + K(\lambda_0 + th_n) - \lambda_0 - K(\lambda_0)] - h_n - J_K(\lambda_0)h_n \\ &= t^{-1}[K(\lambda_0 + th_n) - K(\lambda_0)] - J_K(\lambda_0)h_n \\ &= (J_K(\bar{\lambda}_n) - J_K(\lambda_0))h_n, \end{aligned}$$

where $\bar{\lambda}_n$ is a mean value between $\lambda_0 + th_n$ and λ_0. Therefore, by the Cauchy-Schwarz inequality,

$$\|(J_K(\bar{\lambda}_n) - J_K(\lambda_0))h_n\| \leq \|J_K(\bar{\lambda}_n) - J_K(\lambda_0)\|\|h_n\| \to 0,$$

where we used $\|h_n\| = 1$, $\bar{\lambda}_n \to \lambda_0$, and the continuity of the Jacobian.

For the second claim, note that

$$\|\dot{H}_{\lambda_0} h\| = \|(I - J_K(\lambda_0))h\|,$$

and $h \mapsto \|(I - J_K(\lambda_0))h\|$ is continuous. Since the domain of h is compact, there is $h^* \in \mathbb{S}^p$ such that $\inf_{\|h\|=1} \|\dot{H}_{\lambda_0} h\| = \|(I - J_K(\lambda_0))h^*\|$. Let $q = (I - J_K(\lambda_0))h^*$ and note that $I - J_K(\lambda_0)$ is linearly independent (due to $\det(I - J_K(\lambda_0)) \neq 0$), and hence $q \neq 0$. Hence $\inf_{\|h\|=1} \|\dot{H}_{\lambda_0} h\| = \|q\| > 0$. Hence, the second conclusion follows. □

The following result is a slight extension of Lemma E.1 in CFM.

Lemma 8. *Suppose that $\Lambda \subset \mathbb{R}^p$ and $\mathcal{U}$ is a compact and convex set in $\mathbb{R}^q$. Let $\mathcal{I}$ be an open set containing $\mathcal{U}$. Suppose that $\Psi : \Lambda \times \mathcal{I} \to \mathbb{R}^p$ is continuous and $\lambda \mapsto \Psi(\lambda, u)$ is the gradient of a convex function in λ for each $u \in \mathcal{U}$; (b) for each $u \in \mathcal{U}$, $\Psi(\lambda_0(u), u) = 0$; (c) $\frac{\partial}{\partial(\lambda', u')}\Psi(\lambda, u)$ exists at $(\lambda_0(u), u)$ and is continuous at $(\lambda_0(u), u)$ for each $u \in \mathcal{U}$ and $\dot{\Psi}_{\lambda_0(u),u} := \frac{\partial}{\partial \lambda'}\Psi(\lambda, u)|_{\lambda_0(u)}$ obeys* $\inf_{u \in \mathcal{U}} \inf_{\|h\|=1} \|\dot{\Psi}_{\lambda_0(u),u} h\| > c_0 > 0$. *Then, Condition Z in CFM holds and $u \mapsto \lambda_0(u)$ is continuously differentiable with derivative $J_{\lambda_0}(u) = -\dot{\Psi}^{-1}_{\lambda_0(u)u} \frac{\partial}{\partial u'}\Psi(\lambda_0(u), u)$.*

Proof. The proof is the same as that of Lemma E.1 in CFM, in which $\mathcal{U}$ is a compact interval in $\mathbb{R}$. A slight modification is needed when one computes the derivative of $\lambda_0(u)$ with respect to u. Since u is allowed to be multidimensional, the implicit function theorem gives

$$J_{\lambda_0}(u) = -\dot{\Psi}^{-1}_{\lambda_0(u)u}\frac{\partial}{\partial u'}\Psi(\lambda_0(u), u), \tag{F.15}$$

which is uniformly bounded and continuous in u by condition (c), which ensures continuous differentiability of $u \mapsto \lambda_0(u)$. Note that for any $\delta > 0$ and $\lambda \in B_\delta(\lambda_0(u))$, there is $\eta > 0$ and u' such that $\|u' - u\| \le \eta$ so that

$$\|\lambda - \lambda_0(u')\| \le \|\lambda - \lambda_0(u)\| + \|\lambda_0(u) - \lambda_0(u')\| \le 2\delta. \tag{F.16}$$

Since $\mathcal{U}$ is compact (and hence totally bounded), there is a finite set $\{u_j\}_{j=1}^J \subset \mathcal{U}$ such that $\mathcal{U} \subset \bigcup_j B_\eta(u_j)$. The argument above then shows that $\mathcal{N} = \bigcup_{u\in\mathcal{U}} B_\delta(\lambda_0(u)) \subset \bigcup_j B_{2\delta}(\lambda_0(u_j))$, which ensures that $\mathcal{N}$ is totally bounded. Since $\mathcal{N}$ is a subset of a Euclidean space (equipped with a complete metric), it follows that $\mathcal{N}$ is compact. This ensures condition Z (i) in CFM. The rest of the proof is essentially the same as the case, in which $\mathcal{U}$ being a compact interval. □

Lemma 9. *Suppose Assumption 2 holds. Let $w = (y, d', x', z')$ and let $\tau \in (0, 1)$. Define*

$$\mathcal{M} := \Big\{f : f(w;\theta) = \big((1\{y \le d'\theta_{-1} + x'\theta_1\} - \tau)x, \\ (1\{y \le d'\theta_{-1} + x'\theta_1\} - \tau)z_1, \ldots, (1\{u \le d'\theta_{-1} + x'\theta_1\} - \tau)z_{d_D}\big), \theta \in \Theta\Big\}. \tag{F.17}$$

Then, $\mathcal{M}$ is a Donsker-class.

Proof. The proof is standard, and hence we give a brief sketch for the first component of f, $f_1(w;\theta) = (1\{y \le d'\theta_{-1} + x'\theta_1\} - \tau)x$. Note that $w \mapsto 1\{y \le d'\theta_{-1} + x'\theta_1\} - \tau$ belongs to Type II-class in Andrews (1994), and the map $w \mapsto x$ does not depend on the parameter. By Theorems 2 and 3 in Andrews (1994), this function then satisfies the uniform entropy condition with the envelope function $\bar{M}(w) = x$, which is square integrable by assumption. Similar arguments apply to the other components of f. By Theorem 1 in Andrews (1994), the empirical process: $\mathbb{G}_n f$ is stochastically equicontinuous, and $\mathbb{G}_n f(\cdot, \theta)$ obeys the classical central limit theorem for each $\theta \in \Theta$. Hence, we conclude that $\mathcal{M}$ is Donsker. □

Below, let $g(w;\theta) = (g_1(w;\theta)', \ldots, g_J(w;\theta))'$ be a vector such that

$$g_j(w;\theta) = \frac{\partial^2}{\partial\theta_j\partial\theta_j'}Q_{P,j}(L_j(\theta_{-j}), \theta_{-j})^{-1} f_j(w; L_j(\theta_{-j}), \theta_{-j}),\ j = 1, \ldots, J. \tag{F.18}$$

Let $\rho(\theta, \tilde{\theta}) := \big\|\mathrm{diag}\big(E_P\big[(g(W;\theta) - E_P[g(W;\theta)])(g(w;\tilde{\theta}) - E_P[g(w;\tilde{\theta})])'\big]\big)\big\|$ be the variance semimetric. Let $W_i = (Y_i, D_i', X_i', Z_i'), i = 1, \ldots, N$ be an i.i.d. sample generated from the IVQR model. Define

$$\mathcal{L}_{N,j}(\theta_{-j}) := \sqrt{N}(\hat{L}_j(\theta_{-j}) - L_j(\theta_{-j}))\ , j = 1, \ldots, J. \tag{F.19}$$

Similarly, let $W_i^* = (Y_i^*, D_i^{*\prime}, X_i^{*\prime}, Z_i^{*\prime})', i = 1, \ldots, N$ be an bootstrap sample from the empirical distribution $\hat{P}_N$ of $\{W_i\}$. Define

$$\mathcal{L}^*_{N,j}(\theta_{-j}) := \sqrt{N}(\hat{L}^*_j(\theta_{-j}) - \hat{L}_j(\theta_{-j}))\ , j = 1, \ldots, J, \tag{F.20}$$

where $\hat{L}_j^*$ is the sample best response map of player j, which is defined as in (5.3)-(5.4) while replacing W_i with the bootstrap sample W_i^* in (5.1)-(5.2).

Lemma 10. *Suppose that Assumptions 1 and 2 hold. Then, (i) $\mathcal{L}_N := (\mathcal{L}_{N,1}, \dots, \mathcal{L}_{N,J})$ satisfies*

$$\mathcal{L}_N(\cdot) \rightsquigarrow \mathbb{W}, \tag{F.21}$$

where $\mathbb{W}$ is a tight Gaussian process in $\ell^\infty(\Theta)^d$ with the covariance kernel

$$\mathrm{Cov}(\mathbb{W}(\theta), \mathbb{W}(\tilde{\theta})) = E_P\big[(g(W;\theta) - E_P[g(W;\theta)])(g(W;\tilde{\theta}) - E_P[g(W;\tilde{\theta})])'\big]; \tag{F.22}$$

$\mathcal{L}_N$ is stochastically equicontinuous with respect to the variance semimetric ρ; (ii) $\mathcal{L}_N^* := (\mathcal{L}_{N,1}^*, \dots, \mathcal{L}_{N,J}^*)$ satisfies

$$\mathcal{L}_N^*(\cdot) \overset{L^*}{\rightsquigarrow} \mathbb{W}; \tag{F.23}$$

(iii) ρ satisfies $\lim_{\delta \downarrow 0} \sup_{\|\theta - \tilde{\theta}\| < \delta} \rho(\theta, \tilde{\theta}) \to 0$.

Proof. (i) We first work with $\mathcal{L}_{N,1}$. For this, we establish that L_1 is Hadamard differentiable. Note that $\theta_1 = L_1(\theta_{-1})$ solves

$$E_P[(1\{Y \le D'\theta_{-1} + X'\theta_1\} - \tau)X] = 0. \tag{F.24}$$

Take $\mathcal{U} = \Theta_{-1}$, $\Xi = \Theta_1$, $\psi(\lambda, u) = E_P[(1\{Y \le Du + X'\lambda\} - \tau)X]$. Define $\phi : \ell^\infty(\Xi \times \mathcal{U})^{k_b} \times \ell^\infty(\mathcal{U}) \to \ell^\infty(\mathcal{U})$, which maps (ψ, r) to a solution $\phi(\psi, r) = \lambda(\cdot)$ such that

$$\|\psi(\lambda(u), u)\|^2 \le \inf_{\lambda' \in \Theta} \|\psi(\lambda', u)\|^2 + r(u)^2. \tag{F.25}$$

Then, one may write $L_1(\cdot) = \phi(\psi, 0)$. We then show that ψ satisfies the conditions of Lemma 8. Note first that $\mathcal{U}$ and Ξ are compact. ψ is continuous and $\lambda \mapsto \psi(\lambda, u)$ is the gradient of the convex function $\lambda \mapsto E_P[\rho_\tau(Y - Du - X'\lambda)]$. The function $L_1(u) = \lambda_0(u)$ is defined as the exact solution of $\psi(\lambda, u) = 0$. Note also that, by Assumption 2,

$$\begin{aligned}
\frac{\partial^2}{\partial\theta_1 \partial\theta_1'} Q_{P,1}(\theta_1, \theta_{-1}) &= \frac{\partial}{\partial\theta_1'} E_P[(1\{Y \le D'\theta_{-1} + X'\theta_1\} - \tau)X] \\
&= E_P[\frac{\partial}{\partial\theta_1'}(F_{Y|D,X,Z}(D'\theta_{-1} + X'\theta_1) - \tau)X] \\
&= E_P[f_{Y|D,X,Z}(D'\theta_{-1} + X'\theta_1)XX'],
\end{aligned} \tag{F.26}$$

where the second equality follows from the dominated convergence theorem, and the last display is well-defined by the square integrability of X. Similarly,

$$\frac{\partial^2}{\partial\theta_1 \partial\theta_{-1}'} Q_{P,1}(\theta_1, \theta_{-1}) = E_P[f_{Y|D,X,Z}(D'\theta_{-1} + X'\theta_1)XD']. \tag{F.27}$$

Hence, the derivative

$$\frac{\partial}{\partial(\lambda', u')}\Psi(\lambda, u) = \Big(\frac{\partial^2}{\partial\theta_1 \partial\theta_1'} Q_{P,1}(\theta_1, \theta_{-1}), \frac{\partial^2}{\partial\theta_1 \partial\theta_{-1}'} Q_{P,1}(\theta_1, \theta_{-1})\Big)$$

exists and is continuous by Assumption 2. By Assumption 2.4, $\dot{\Psi}_{\lambda_0(u),u} = \frac{\partial^2}{\partial\theta_1\partial\theta_1'}Q_{P,1}(L_1(\theta_{-1}),\theta_{-1})$ obeys

$$\inf_{u\in\mathcal{U}}\inf_{\|h\|=1}\|\dot{\Psi}_{\lambda_0(u),u}h\| = \inf_{\theta_{-1}\in\Theta_{-1}}\inf_{\|h\|=1}\|E_P[f_{Y|D,X,Z}(D'\theta_{-1}+X'\theta_1)XX']h\| > 0. \tag{F.28}$$

Then, by Lemma 8 and Lemma E.2 in Chernozhukov, Fernandez-Val, and Melly (2013), ϕ is Hadamard differentiable tangentially to $\mathcal{C}(\mathcal{N}\times\mathcal{U})^K\times\{0\}$ with the Hadamard derivative (of L_1)

$$\phi'_{\Psi,0}(z,0) = -\frac{\partial^2}{\partial\theta_1\partial\theta_1'}Q_{P,1}(L_1(\cdot),\cdot)^{-1}z(L_1(\cdot),\cdot), \tag{F.29}$$

where $(z,0)\mapsto\phi'_{\Psi,0}(z,0)$ is continuous over $z\in\ell^\infty(\Theta)^K$.

For $j\neq 1$, the argument is similar. For example, for $\ell=2$, one may take $\mathcal{U}=\Theta_{-2}$, $\Xi=\Theta_2$ and $\psi(\lambda,u)=E_P[(1\{Y\leq D_2\theta_2+(D_1,X)'u\}-\tau)Z_2]$ and write $L_2(\cdot)=\phi(\psi,0)$. The rest of the argument is the same.

By Lemma 9 and arguing as in (F.24)-(F.29) and applying the δ-method (as in Lemma E.3 in CFM), we obtain

$$\mathcal{L}_N(\cdot)\rightsquigarrow\mathbb{W}, \tag{F.30}$$

where $\mathbb{W}=(\mathbb{W}_1',\dots,\mathbb{W}_J')'$ is a tight Gaussian process in $\ell^\infty(\Theta)^d$, where for each j, $\mathbb{W}_j\in\ell^\infty(\Theta_{-j})^{d_j}$ is given pointwise by

$$\mathbb{W}_j(\theta_{-j}) = -\frac{\partial^2}{\partial\theta_j\partial\theta_j'}Q_{P,j}(L_j(\theta_{-j}),\theta_{-j})^{-1}\mathbb{G}f_j(w;L_j(\theta_{-j}),\theta_{-j}),\ j=1,\cdots,J; \tag{F.31}$$

Hence, its covariance kernel is as given in (F.22). By Lemma 1.3.8. in Van der Vaart and Wellner (1996), $\{\mathcal{L}_N\}$ is asymptotically tight, which in turn means that $\{\mathcal{L}_N\}$ is stochastically equicontinuous with respect to ρ by Theorem 1.5.7 in Van der Vaart and Wellner (1996).

(ii) For each j, let $\mathcal{L}^*_{N,j}\in\ell^\infty(\Theta_{-j})^{d_j}$ be defined pointwise by

$$\mathcal{L}^*_{N,j}(\theta_{-j}) = \sqrt{N}(\hat{L}^*_j(\theta_{-j}))-\hat{L}_j(\theta_{-j})). \tag{F.32}$$

Below, again we work with the case $j=1$. Using ϕ (the solution to (F.25)), we may write

$$\mathcal{L}^*_{N,1}(\theta_{-1}) = \sqrt{N}(\phi(\hat{\psi}^*_N,r^*_N)-\phi(\hat{\psi}_N,r_N)), \tag{F.33}$$

where $\hat{\psi}_N(\lambda,u) = N^{-1}\sum_{i=1}^N(1\{Y_i\leq D_iu+X_i'\lambda\}-\tau)X_i$, and $\hat{\psi}^*_N$ is defined similarly for the bootstrap sample. Let E_{P^*} denote the conditional expectation with respect to P^*, the law of $\{W_i^*\}_{i=1}^N$ conditional on the sample path. Let BL_1 denote the space of bounded Lipschitz functions on $\mathbb{R}^{d_1}$ with Lipschitz constant 1. Then, for any $\epsilon>0$,

$$\begin{aligned}\sup_{h\in BL_1}&\Big|E_{P^*}h\big(\sqrt{N}\big[\phi(\hat{\psi}^*_N,r^*_N)-\phi(\hat{\psi}_N,r_N)\big]\big)-E_{P^*}h\big(\phi'_{\Psi,0}\big(\sqrt{N}\big[(\hat{\psi}^*_N,r^*_N)-(\hat{\psi}_N,r_N)\big]\big)\big)\Big|\\ &\leq\epsilon+2P^*\Big(\big\|\sqrt{N}\big[\phi(\hat{\psi}^*_N,r^*_N)-\phi(\hat{\psi}_N,r_N)\big]-\phi'_{\Psi,0}\big(\sqrt{N}\big[(\hat{\psi}^*_N,r^*_N)-(\hat{\psi}_N,r_N)\big]\big)\big\|>\epsilon\Big).\end{aligned} \tag{F.34}$$

By Lemma 9 and Theorem 3.6.2 in Van der Vaart and Wellner (1996), $\sqrt{N}(\hat{\psi}_N^* - \hat{\psi}_N) \overset{L^*}{\rightsquigarrow} \mathbb{G}f_1$. Noting that $h \circ \phi'_{\Psi,0} \in BL_1(\ell^\infty(\Theta_{-1})^{d_1} \times \mathbb{R})$ and $r_N = o_p(N^{-1/2})$, it follows that

$$\sup_{h \in BL_1} \Big| E_{P^*} h\big(\phi'_{\Psi,0}\big(\sqrt{N}\big[(\hat{\psi}_N^*, r_N^*) - (\hat{\psi}_N, r_N)\big]\big)\big) - E_{P^*} h \circ \phi'_{\Psi,0}(\mathbb{G}f_1, 0)\Big| \to 0, \tag{F.35}$$

with probability approaching 1 due to $r_N = o_P(N^{-1/2})$. Hence, for the conclusion of the theorem, it suffices to show that the second term on the right hand side of (F.34) tends to 0.

As shown in the proof of (i), ϕ is Hadamard differentiable at $(\psi, 0)$. Hence, by Theorem 3.9.4 in Van der Vaart and Wellner (1996),

$$\begin{aligned}\sqrt{N}\big[\phi(\hat{\psi}_N^*, r_N^*) - \phi(\psi, 0)\big] &= \phi'_{\Psi,0}(\sqrt{N}[(\hat{\psi}_N^*, r_N^*) - (\psi, 0)]) + o_{P^*}(1)\\ \sqrt{N}\big[\phi(\hat{\psi}_N, r_N) - \phi(\psi, 0)\big] &= \phi'_{\Psi,0}(\sqrt{N}[(\hat{\psi}_N, r_N) - (\psi, 0)]) + o_P(1),\end{aligned}$$

Take the difference of the left and right hand sides respectively and note that $\phi'_{\Psi,0}$ is linear. This implies the right hand side of (F.34) tends to 0 in probability. This, together with (F.34)-(F.35), ensures

$$\mathcal{L}_{N,1}^* \overset{L^*}{\rightsquigarrow} \mathbb{W}_1, \tag{F.36}$$

where $\mathbb{W}_1(\theta_{-1}) = -\frac{\partial^2}{\partial\theta_1\partial\theta_1'} Q_{P,1}(L_1(\theta_{-1}), \theta_{-1})^{-1} \mathbb{G} f_j(\cdot; L_1(\theta_{-1}), \theta_{-1})$. The analysis for any $j \neq 1$ is similar, and one may apply the arguments above jointly across $j = 1, \ldots, J$, which yields the second claim of the lemma.

(iii) Consider the first submatrix of $E_P[(g(W;\theta) - E_P[g(W;\theta)])(g(w;\tilde{\theta}) - E_P[g(w;\tilde{\theta})])']$. It is given by

$$\begin{aligned}&\mathrm{Var}\Big(-\frac{\partial^2}{\partial\theta_1\partial\theta_1'} Q_{P,1}(L_1(\theta_{-1}), \theta_{-1})^{-1} f_1(w; L_1(\theta_{-1}), \theta_{-1})\Big)\\ &\qquad - \mathrm{Var}\Big(-\frac{\partial^2}{\partial\theta_1\partial\theta_1'} Q_{P,1}(L_1(\tilde{\theta}_{-1}), \tilde{\theta}_{-1})^{-1} f_1(w; L_1(\tilde{\theta}_{-1}), \tilde{\theta}_{-1})\Big)\\ &\quad = \frac{\partial^2}{\partial\theta_1\partial\theta_1'} Q_{P,1}(L_1(\theta_{-1}), \theta_{-1})^{-1} \mathrm{Var}(f_1(w; L_1(\theta_{-1}), \theta_{-1})) \frac{\partial^2}{\partial\theta_1\partial\theta_1'} Q_{P,1}(L_1(\theta_{-1}), \theta_{-1})^{-1}\\ &\qquad - \frac{\partial^2}{\partial\theta_1\partial\theta_1'} Q_{P,1}(L_1(\tilde{\theta}_{-1}), \tilde{\theta}_{-1})^{-1} \mathrm{Var}(f_1(w; L_1(\tilde{\theta}_{-1}), \tilde{\theta}_{-1})) \frac{\partial^2}{\partial\theta_1\partial\theta_1'} Q_{P,1}(L_1(\tilde{\theta}_{-1}), \tilde{\theta}_{-1})^{-1}.\end{aligned} \tag{F.37}$$

Note that Θ is compact and $\theta_{-1} \mapsto \frac{\partial^2}{\partial\theta_1\partial\theta_1'} Q_{P,1}(L_1(\theta_{-1}), \theta_{-1})^{-1}$ is continuous by Lemma 1, which implies that this map is uniformly continuous. Therefore, it remains to show the uniform continuity of $\theta \mapsto \mathrm{Var}(f_1(w;\theta))$. Note that

$$\begin{aligned}\mathrm{Var}(f_1(w; L_1(\theta_{-1}), \theta_{-1})) &= E_P[(1\{Y \leq D'\theta_{-1} + X'L_1(\theta_{-1})\} - \tau)XX']\\ &\quad - E_P[(1\{Y \leq D'\theta_{-1} + X'L_1(\theta_{-1})\} - \tau)X] E_P[(1\{Y \leq D'\theta_{-1} + X'L_1(\theta_{-1})\} - \tau)X]'.\end{aligned} \tag{F.38}$$

The right hand side of the display above is continuous on the compact domain Θ, and hence it is uniformly continuous. One can argue the same way for the other subcomponents of $\mathrm{diag}\big(E_P[(g(W;\theta) - E_P[g(W;\theta)])(g(w;\tilde{\theta}) - E_P[g(w;\tilde{\theta})])']\big)$. This completes the proof. □

Corollary 2. *Suppose that Assumptions 1 and 2 hold. (i) Let $W_i = (Y_i, D_i', X_i', Z_i')', i = 1, \dots, N$ be an i.i.d. sample generated from the IVQR model. Then,*

$$\sqrt{N}\big(\hat{K} - K\big) \rightsquigarrow \mathbb{W}. \tag{F.39}$$

(ii) Let $W_i^* = (Y_i^*, D_i^{*\prime}, X_i^{*\prime}, Z_i^{*\prime})', i = 1, \dots, N$ be an bootstrap sample from the empirical distribution $\hat{P}_N$ of $\{W_i\}_{i=1}^N$. Then,

$$\sqrt{N}\big(\hat{K}^* - \hat{K}\big) \overset{L^*}{\rightsquigarrow} \mathbb{W}.$$

Proof. (i) By Lemma 10, it follows that

$$\sqrt{N}\big(\hat{L}_1(\cdot) - L_1(\cdot), \dots, \hat{L}_J(\cdot) - L_J(\cdot)\big)' \rightsquigarrow \mathbb{W}.$$

Note that, by the definition of $\hat{L}$ and L, one has

$$\sqrt{N}(\hat{K}_j(\theta) - K_j(\theta)) = \sqrt{N}(\hat{L}_j(\theta_{-j}) - L_j(\theta_{-j})), j = 1, \cdots, J.$$

The conclusion of the lemma then follows. The proof of (ii) is similar, and is therefore omitted. □

Appendix G. Consistency of the Contraction Estimator

Below, we adopt the framework of Dominitz and Sherman (2005) Let $(\mathcal{X}, d)$ be a metric space. For a contraction map $F : \mathcal{X} \to \mathcal{X}$, let c_F be the modulus of contraction such that

$$d(F(x), F(x')) \le c_F d(x, x'),$$

for any $x, x' \in \mathcal{X}$.

Lemma 11. *Suppose Assumptions 1, 2, and 3 hold. Let $\hat{\theta}_N$ be an estimator constructed by iterating the dynamical system in (5.7) or (in (5.8)) s_N times, where $s_N \ge -\frac{1}{2}\ln N / \ln c_K$. Then,*

$$\hat{\theta}_N - \theta^* = O_p(N^{-1/2}).$$

Proof. We show the result by applying Theorem 1 in Dominitz and Sherman (2005) to the estimator obtained from the simultaneous dynamical system. The argument for the sequential system is similar.

By Assumption 3, K is a contraction map on D_K. Let $\theta^{(s)}$ be obtained from iterating s-times the population dynamical system in (3.15). The iteration on the dynamical system is covergent at least linearly (Bertsekas and Tsitsiklis, 1989, Proposition 1.1). Under the condition on s_N, arguing as in (Dominitz and Sherman, 2005, p.842), it follows that $N^{1/2}\|\theta^{(s_N)} - \theta^*\| \le \|\theta^{(0)} - \theta^*\|$. Finally, by Corollary 2 and tightness of $\mathbb{W}$, $N^{1/2} \sup_{\theta \in D_K} \|\hat{K}(\theta) - K(\theta)\| = O_p(1)$. These imply the conditions of Theorem 1 in Dominitz and Sherman (2005) with $\delta = 1/2$. The claim of the lemma then follows. □

Tables

Table 1. Algorithms

One endogenous variable		
Algorithm	R-Package	Comments
Contraction algorithm		
Root-finding algorithm	`uniroot` (R Core Team, 2018)	
IQR		500 gridpoints

Two endogenous variables		
Algorithm	R-Package	Comments
Contraction algorithm		
Root-finding algorithm	`optim_sa` (Husmann, Lange, and Spiegel, 2017)	implemented as optimizer
Nested root-finding algorithm	`uniroot` (R Core Team, 2018)	
IQR		40×40 gridpoints, implementation: p.132 in Chernozhukov, Hansen, and Wüthrich (2017)

Table 2. Bias and RMSE, 401(k) DGP with one endogenous regressor

	Bias/10^2			RMSE/10^3		
τ	Contr.	Brent	IQR	Contr.	Brent	IQR
0.15	-6.66	-6.52	-8.65	8.08	7.43	7.88
0.25	-1.77	-3.17	-3.14	3.89	3.97	3.97
0.50	0.88	0.54	0.74	1.99	1.99	2.00
0.75	-1.41	-1.10	-0.91	1.96	1.96	1.96
0.85	0.05	0.65	0.74	2.10	2.11	2.11

Notes: Monte Carlo simulation with 500 repetitions as described in the main text. Contr: contraction algorithm; Brent: root-finding algorithm based on Brent's method; IQR: inverse quantile regression.

Table 3. Bias and RMSE, 401(k) DGP with two endogenous regressors

	Bias/10^2				RMSE/10^3			
τ	Contr.	SA	Nested	IQR	Contr.	SA	Nested	IQR
Coefficient on binary endogenous variable								
0.15	-9.12	-1.73	-7.42	-9.36	8.03	6.92	7.63	8.19
0.25	-5.74	-5.84	-6.11	-6.52	4.46	4.40	4.45	4.54
0.50	-0.25	-0.36	-0.43	-0.42	1.94	1.96	1.95	2.00
0.75	0.24	0.26	0.21	0.36	1.81	1.82	1.82	1.87
0.85	-0.31	0.07	0.07	0.06	2.20	2.21	2.21	2.26
Coefficient on continuous endogenous variable								
0.15	2.14	4.66	0.54	0.48	1.07	2.12	1.04	1.13
0.25	2.26	0.90	0.33	-0.03	0.97	1.25	0.97	1.04
0.50	1.12	0.16	0.03	0.01	0.89	0.96	0.95	1.07
0.75	-1.40	0.01	-0.26	0.00	0.98	1.06	1.07	1.16
0.85	-3.28	-1.08	-1.23	-1.12	1.11	1.25	1.26	1.33

Notes: Monte Carlo simulation with 500 repetitions as described in the main text. Contr: contraction algorithm; SA: simulated annealing based optimization algorithm; Nested: nested algorithm based Brent's method; IQR: inverse quantile regression.

Table 4. Size, 401(k) DGP with one endogenous regressor

	$1-\alpha=0.95$		$1-\alpha=0.9$	
τ	Contr.	Brent	Contr.	Brent
0.15	0.95	0.95	0.91	0.88
0.25	0.96	0.96	0.93	0.93
0.50	0.96	0.96	0.91	0.91
0.75	0.94	0.94	0.89	0.89
0.85	0.94	0.95	0.90	0.90

Notes: Monte Carlo simulation with 1000 repetitions as described in the main text. Contr: contraction algorithm; Brent: root-finding algorithm based on Brent's method.

Table 5. Computation time, 401(k) DGP with one endogenous regressor

N	Contr.	Brent	IQR
1000	0.28	0.04	0.42
5000	0.49	0.18	4.00
10000	1.10	0.54	13.40
20000	1.95	0.77	23.24

Notes: The table reports average computation time in seconds at $\tau = 0.5$ over 50 simulation repetitions based on the DGP described in the main text. Contr: contraction algorithm; Brent: root-finding algorithm based on Brent's method; IQR: inverse quantile regression with grid search over 500 grid points.

Table 6. Computation time, 401(k) DGP with two endogenous regressor

N	Contr.	SA	Nested	IQR
1000	0.29	3.70	0.51	62.21
5000	1.54	20.10	2.95	322.78
10000	4.45	59.44	9.24	730.20
20000	20.09	193.63	31.62	2152.55

Notes: The table reports average computation time in seconds at $\tau = 0.5$ over 50 simulation repetitions based on the DGP described in the main text. Contr: contraction algorithm; SA: simulated annealing based optimization algorithm; Nested: nested algorithm based Brent's method; IQR: inverse quantile regression with grid search over 100×100 grid points.

Table 7. Bias and RMSE, symmetric design with one endogenous regressor

	$N = 500$					
	Bias			RMSE		
τ	Contr.	Brent	IQR	Contr.	Brent	IQR
0.15	0.03	-0.00	-0.00	0.10	0.10	0.10
0.25	0.03	0.00	0.00	0.12	0.12	0.12
0.50	-0.00	-0.00	-0.00	0.12	0.14	0.14
0.75	-0.04	-0.01	-0.01	0.13	0.12	0.12
0.85	-0.04	-0.00	-0.00	0.11	0.11	0.11

	$N = 1000$					
	Bias			RMSE		
τ	Contr.	Brent	IQR	Contr.	Brent	IQR
0.15	0.02	0.00	0.00	0.07	0.07	0.07
0.25	0.01	-0.00	-0.00	0.08	0.08	0.08
0.50	-0.01	-0.01	-0.01	0.09	0.10	0.10
0.75	-0.02	-0.00	-0.00	0.09	0.08	0.08
0.85	-0.02	-0.00	-0.00	0.08	0.08	0.08

Notes: Monte Carlo simulation with 500 repetitions as described in the main text. Contr: contraction algorithm; Brent: root-finding algorithm based on Brent's method; IQR: inverse quantile regression.

Table 8. Bias and RMSE, asymmetric design with one endogenous regressor

	$N = 500$					
	Bias			RMSE		
τ	Contr.	Brent	IQR	Contr.	Brent	IQR
0.15	0.12	0.01	-0.00	0.22	0.20	0.20
0.25	0.07	0.00	-0.00	0.17	0.16	0.16
0.50	0.04	-0.00	-0.00	0.13	0.12	0.12
0.75	0.03	0.00	0.00	0.11	0.11	0.11
0.85	-0.03	-0.01	-0.00	0.12	0.11	0.11

	$N = 1000$					
	Bias			RMSE		
τ	Contr.	Brent	IQR	Contr.	Brent	IQR
0.15	0.05	-0.01	-0.01	0.16	0.15	0.15
0.25	0.04	0.00	0.00	0.11	0.11	0.11
0.50	0.03	0.00	0.00	0.08	0.08	0.08
0.75	0.01	-0.01	-0.01	0.08	0.08	0.08
0.85	-0.03	-0.01	-0.01	0.09	0.09	0.09

Notes: Monte Carlo simulation with 500 repetitions as described in the main text. Contr: contraction algorithm; Brent: root-finding algorithm based on Brent's method; IQR: inverse quantile regression.

Table 9. Bias and RMSE, symmetric design with two endogenous regressors

	$N = 500$							
	Bias				RMSE			
τ	Contr.	SA	Nested	IQR	Contr.	SA	Nested	IQR
	Coefficient on D_1							
0.15	0.00	0.00	-0.00	-0.01	0.11	0.14	0.12	0.13
0.25	0.01	-0.01	-0.00	-0.01	0.15	0.17	0.16	0.16
0.50	-0.02	-0.02	-0.02	-0.02	0.17	0.19	0.19	0.20
0.75	-0.04	-0.03	-0.03	-0.03	0.21	0.21	0.20	0.20
0.85	-0.05	-0.03	-0.03	-0.03	0.18	0.18	0.17	0.17
	Coefficient on D_2							
0.15	0.10	-0.01	-0.01	-0.02	0.27	0.29	0.27	0.31
0.25	0.10	-0.02	-0.00	-0.02	0.29	0.30	0.29	0.30
0.50	-0.01	-0.02	-0.02	-0.02	0.33	0.39	0.38	0.39
0.75	-0.15	-0.06	-0.04	-0.05	0.40	0.41	0.40	0.41
0.85	-0.19	-0.06	-0.05	-0.07	0.39	0.40	0.36	0.43

	$N = 1000$							
	Bias				RMSE			
τ	Contr.	SA	Nested	IQR	Contr.	SA	Nested	IQR
	Coefficient on D_1							
0.15	-0.00	-0.01	-0.00	-0.00	0.08	0.10	0.09	0.10
0.25	-0.00	-0.01	-0.00	-0.01	0.10	0.12	0.11	0.13
0.50	-0.01	-0.01	-0.01	-0.01	0.12	0.13	0.13	0.16
0.75	-0.01	-0.01	-0.01	-0.00	0.13	0.14	0.13	0.14
0.85	-0.02	-0.02	-0.01	-0.02	0.12	0.13	0.12	0.13
	Coefficient on D_2							
0.15	0.05	-0.01	-0.01	-0.02	0.19	0.21	0.19	0.20
0.25	0.05	-0.01	-0.00	-0.01	0.22	0.23	0.21	0.23
0.50	-0.02	-0.02	-0.02	-0.03	0.25	0.27	0.27	0.29
0.75	-0.09	-0.02	-0.02	-0.03	0.27	0.28	0.25	0.26
0.85	-0.09	-0.03	-0.01	-0.03	0.26	0.25	0.23	0.24

Notes: Monte Carlo simulation with 500 repetitions as described in the main text. Contr: contraction algorithm; SA: simulated annealing based optimization algorithm; Nested: nested algorithm based Brent's method; IQR: inverse quantile regression.

Table 10. Bias and RMSE, asymmetric design with two endogenous regressors

	$N = 500$							
	Bias				RMSE			
τ	Contr.	SA	Nested	IQR	Contr.	SA	Nested	IQR
				Coefficient on D_1				
0.15	-0.02	0.00	0.02	0.01	0.25	0.28	0.26	0.26
0.25	-0.05	0.00	0.01	-0.00	0.20	0.21	0.20	0.21
0.50	-0.04	0.00	-0.00	0.00	0.16	0.21	0.17	0.19
0.75	-0.02	-0.01	-0.02	-0.02	0.17	0.18	0.17	0.19
0.85	-0.01	-0.02	-0.01	-0.02	0.20	0.19	0.19	0.19
				Coefficient on D_2				
0.15	0.26	-0.11	-0.06	-0.13	0.57	0.58	0.52	0.59
0.25	0.23	-0.02	-0.01	-0.01	0.45	0.43	0.41	0.44
0.50	0.12	-0.04	-0.03	-0.07	0.34	0.48	0.32	0.73
0.75	0.04	-0.06	-0.05	-0.05	0.32	0.34	0.31	0.34
0.85	-0.13	-0.01	-0.03	0.01	0.40	0.38	0.34	0.36

	$N = 1000$							
	Bias				RMSE			
τ	Contr.	SA	Nested	IQR	Contr.	SA	Nested	IQR
				Coefficient on D_1				
0.15	-0.03	-0.00	0.01	-0.01	0.18	0.19	0.19	0.19
0.25	-0.04	-0.01	-0.00	-0.01	0.15	0.16	0.15	0.16
0.50	-0.03	-0.01	-0.01	-0.01	0.13	0.14	0.13	0.14
0.75	-0.03	-0.01	-0.01	-0.01	0.12	0.13	0.12	0.14
0.85	0.01	0.00	0.00	-0.00	0.14	0.15	0.13	0.15
				Coefficient on D_2				
0.15	0.15	-0.03	-0.03	-0.04	0.37	0.38	0.37	0.39
0.25	0.10	-0.01	-0.01	-0.02	0.28	0.30	0.28	0.28
0.50	0.05	-0.03	-0.02	-0.03	0.22	0.23	0.22	0.24
0.75	0.06	-0.02	-0.01	-0.01	0.24	0.24	0.22	0.24
0.85	-0.08	-0.04	-0.03	-0.03	0.27	0.26	0.24	0.24

Notes: Monte Carlo simulation with 500 repetitions as described in the main text. Contr: contraction algorithm; SA: simulated annealing based optimization algorithm; Nested: nested algorithm based Brent's method; IQR: inverse quantile regression.

Table 11. Size, location-scale DGP with one endogenous regressor

	$N = 500$							
	Symmetric Design				Asymmetric Design			
	$1-\alpha=0.95$		$1-\alpha=0.9$		$1-\alpha=0.95$		$1-\alpha=0.9$	
τ	Contr.	Brent	Contr.	Brent	Contr.	Brent	Contr.	Brent
0.15	0.95	0.97	0.91	0.93	0.92	0.97	0.87	0.94
0.25	0.95	0.97	0.91	0.92	0.93	0.96	0.89	0.93
0.50	0.96	0.97	0.90	0.91	0.94	0.96	0.90	0.92
0.75	0.95	0.96	0.90	0.92	0.96	0.96	0.93	0.92
0.85	0.96	0.97	0.91	0.93	0.95	0.95	0.93	0.92

	$N = 1000$							
	Symmetric Design				Asymmetric Design			
	$1-\alpha=0.95$		$1-\alpha=0.9$		$1-\alpha=0.95$		$1-\alpha=0.9$	
τ	Contr.	Brent	Contr.	Brent	Contr.	Brent	Contr.	Brent
0.15	0.96	0.96	0.90	0.91	0.93	0.96	0.87	0.93
0.25	0.94	0.94	0.90	0.89	0.93	0.95	0.88	0.91
0.50	0.96	0.96	0.90	0.91	0.93	0.94	0.89	0.89
0.75	0.95	0.95	0.91	0.92	0.95	0.94	0.90	0.90
0.85	0.96	0.95	0.91	0.92	0.96	0.95	0.92	0.90

Notes: Monte Carlo simulation with 1000 repetitions as described in the main text. Contr: contraction algorithm; Brent: root-finding algorithm based on Brent's method.

Figures

Figure 1. Illustration Fixed Point

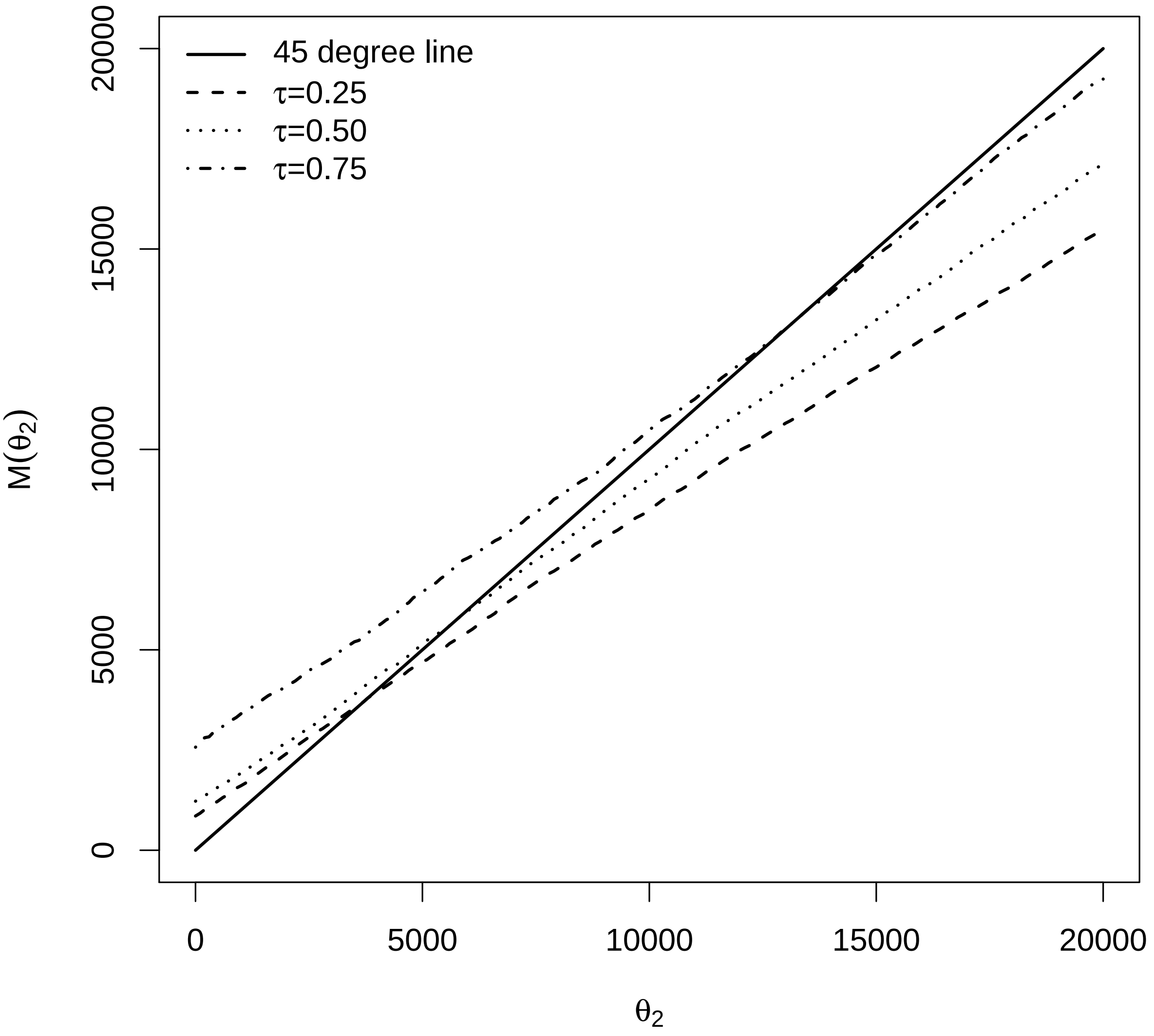


Figure 2. Comparison Point Estimates

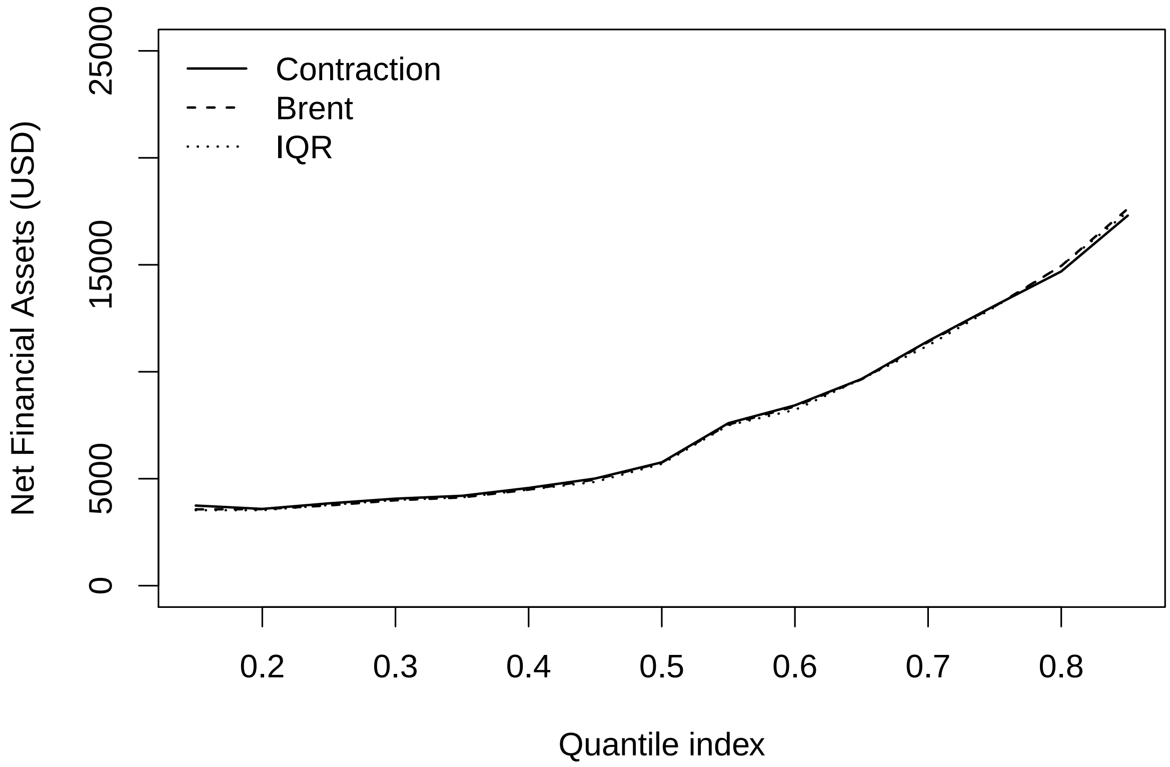
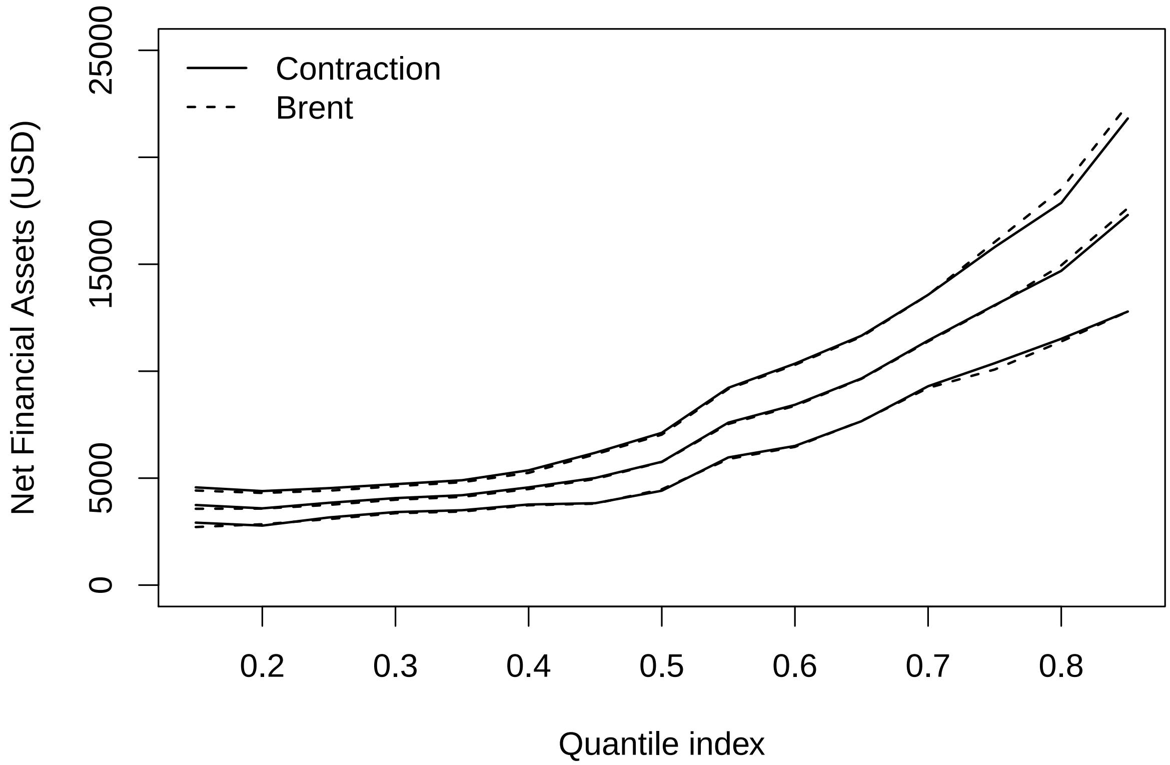


Figure 3. Pointwise 95% Bootstrap Confidence Intervals


Contraction
Brent
Net Financial Assets (USD)
Quantile index