

CONTRIBUTIONS TO THE THEORY AND METHODOLOGY OF QUANTILE
MODELS WITH ENDOGENEITY

A Dissertation
presented to
the Faculty of the Graduate School
at the University of Missouri

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

by
XIN LIU
Dr. David Kaplan, Dissertation Supervisor

MAY 2021

The undersigned, appointed by the Dean of the Graduate School, have examined the dissertation entitled

CONTRIBUTIONS TO THE THEORY AND METHODOLOGY OF QUANTILE
MODELS WITH ENDOGENEITY

presented by Xin Liu,

a candidate for the degree of Doctor of Philosophy and hereby certify that, in their opinion, it is worthy of acceptance.

Professor David Kaplan

Professor J. Isaac Miller

Professor Shawn Ni

Professor Shih-Kang Chao

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, David Kaplan, for his patience, encouragements and guidance. He is a great mentor and my role model.

I would like to express my appreciation to my committee. I thank Zack Miller, Shawn Ni, and Shih-Kang Chao for the invaluable input, support and encouragement on my dissertation research.

I would like to express my appreciation to the Department of Economics at the University of Missouri for the generous support on me pursuing my PhD degree.

Last, but not the least, I would like thank my family and friends for the love, support and encouragement.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	ii
LIST OF FIGURES	iv
LIST OF TABLES	vii
ABSTRACT	ix
CHAPTER	
1 Smoothed GMM for quantile models	1
2 Averaging estimation for instrumental variables quantile regression	20
3 Panel quantile regression with time-invariant rank	50
APPENDIX	
A Appendix for Chapter 1	96
B Appendix for Chapter 2	108
C Appendix for Chapter 3	145
BIBLIOGRAPHY	183
VITA	192

LIST OF FIGURES

1.1	Graph of unsmoothed $1f - 0g$ and example smoothed version $\hat{F}(\cdot)$ from (1.7).	6
2.1	Relative rRMSE in simulation model 1; $\tau = 0.5$, $n = 1000$, 200 replications, 50 bootstraps.	39
2.2	Relative rRMSE in simulation model 2 at $\tau = 0.5$ median level in 6 cases: fixed endogeneity level $c_0 = 0$ (left top), $c_0 = 0.2$ (left middle), $c_0 = 0.4$ (left bottom) and varying heterogeneity; and fixed heterogeneity level $c_h = 0$ (right top), $c_h = 0.5$ (right middle), $c_h = 1$ (right bottom) and varying endogeneity, based on 200 replications and 50 bootstraps. Sample size $n = 1000$	45
2.3	Relative rRMSE in simulation model 2 at $\tau = 0.7$ quantile level in 6 cases: fixed endogeneity level $c_0 = 0$ (left top), $c_0 = 0.2$ (left middle), $c_0 = 0.4$ (left bottom) and varying heterogeneity; and fixed heterogeneity level $c_h = 0$ (right top), $c_h = 0.5$ (right middle), $c_h = 1$ (right bottom) and varying endogeneity, based on 200 replications and 50 bootstraps. Sample size $n = 1000$	47
3.1	Left: pointwise confidence intervals (red) and joint confidence sets (blue) for the n points on the generic coefficient function in Case 1. Right: uniform confidence band for the monotone coefficient function $\beta(\cdot)$ in Case 2.	73
3.2	Comparison between Case 1 and Case 2. In Case 2, the red nondecreasing curves that are within the shaded gray area indicate the possible coefficient functions consistent with the two intervals t_k and t_{k+1} . In Case 1, we only know that the possible coefficient function (denoted as the blue curves) crosses t_k and t_{k+1} ; there is no information about the coefficient function for the value between $\beta(U_{n:k})$ and $\beta(U_{n:k+1})$	77
3.3	Confidence set for n points on the generic coefficient function in general model.	81

3.4	Joint confidence set for $\beta(\cdot)$	95
B.1	Relative rRMSE in simulation model 1 (Gaussian error) at $\tau = 0.2$ (left top), $\tau = 0.3$ (left middle), $\tau = 0.4$ (left bottom), $\tau = 0.6$ (right top), $\tau = 0.7$ (right middle), $\tau = 0.8$ (right bottom), based on 200 replications and 50 bootstraps. Sample size $n = 1000$	129
B.2	Relative rRMSE in simulation model 1 (non-Gaussian error) at $\tau = 0.5$ based on 200 replications and 50 bootstraps. Sample size $n = 1000$	130
B.3	Relative rRMSE in simulation model 1 (non-Gaussian error) at $\tau = 0.2$ (left top), $\tau = 0.3$ (left middle), $\tau = 0.4$ (left bottom), $\tau = 0.6$ (right top), $\tau = 0.7$ (right middle), $\tau = 0.8$ (right bottom), based on 200 replications and 50 bootstraps. Sample size $n = 1000$	134
B.4	Relative rRMSE in simulation model 2 at $\tau = 0.2$ quantile level in 6 cases: fixed endogeneity level $c_0 = 0$ (left top), $c_0 = 0.2$ (left middle), $c_0 = 0.4$ (left bottom) and varying heterogeneity; and fixed heterogeneity level $c_h = 0$ (right top), $c_h = 0.5$ (right middle), $c_h = 1$ (right bottom) and varying endogeneity, based on 200 replications and 50 bootstraps. Sample size $n = 1000$	138
B.5	Relative rRMSE in simulation model 2 at $\tau = 0.3$ quantile level in 6 cases: fixed endogeneity level $c_0 = 0$ (left top), $c_0 = 0.2$ (left middle), $c_0 = 0.4$ (left bottom) and varying heterogeneity; and fixed heterogeneity level $c_h = 0$ (right top), $c_h = 0.5$ (right middle), $c_h = 1$ (right bottom) and varying endogeneity, based on 200 replications and 50 bootstraps. Sample size $n = 1000$	139
B.6	Relative rRMSE in simulation model 2 at $\tau = 0.4$ quantile level in 6 cases: fixed endogeneity level $c_0 = 0$ (left top), $c_0 = 0.2$ (left middle), $c_0 = 0.4$ (left bottom) and varying heterogeneity; and fixed heterogeneity level $c_h = 0$ (right top), $c_h = 0.5$ (right middle), $c_h = 1$ (right bottom) and varying endogeneity, based on 200 replications and 50 bootstraps. Sample size $n = 1000$	140
B.7	Relative rRMSE in simulation model 2 at $\tau = 0.6$ quantile level in 6 cases: fixed endogeneity level $c_0 = 0$ (left top), $c_0 = 0.2$ (left middle), $c_0 = 0.4$ (left bottom) and varying heterogeneity; and fixed heterogeneity level $c_h = 0$ (right top), $c_h = 0.5$ (right middle), $c_h = 1$ (right bottom) and varying endogeneity, based on 200 replications and 50 bootstraps. Sample size $n = 1000$	141

B.8	Relative rRMSE in simulation model 2 at $\tau = 0.8$ quantile level in 6 cases: fixed endogeneity level $c_0 = 0$ (left top), $c_0 = 0.2$ (left middle), $c_0 = 0.4$ (left bottom) and varying heterogeneity; and fixed heterogeneity level $c_h = 0$ (right top), $c_h = 0.5$ (right middle), $c_h = 1$ (right bottom) and varying endogeneity, based on 200 replications and 50 bootstraps. Sample size $n = 1000$	142
-----	--	-----

LIST OF TABLES

1.1	Smoothed MM and GMM estimates of β_τ , log-linear model.	17
1.2	Smoothed MM and GMM estimates of γ_τ , log-linear model.	17
2.1	Relative rRMSE in simulation model 1, $\tau = 0.5$	39
2.2	Bounds of relative rRMSE in simulation model 2, $\tau = 0.5$	43
3.1	Empirical results for effect of oil wealth on military defense spending.	94
B.1	Finite-sample relative rRMSE in JTPA-based simulation model 3, $\tau = 0.5$	122
B.2	Relative rRMSE of bootstrap averaging in simulation model 3.	124
B.3	Relative rRMSE in simulation model 1 at $\tau = 0.2$	126
B.4	Relative rRMSE in simulation model 1 at $\tau = 0.3$	126
B.5	Relative rRMSE in simulation model 1 at $\tau = 0.4$	127
B.6	Relative rRMSE in simulation model 1 at $\tau = 0.6$	127
B.7	Relative rRMSE in simulation model 1 at $\tau = 0.7$	128
B.8	Relative rRMSE in simulation model 1 at $\tau = 0.8$	128
B.9	Relative rRMSE in simulation model 1 (non-Gaussian error) at $\tau = 0.5$ (median).	130
B.10	Relative rRMSE in simulation model 1 (non-Gaussian error) at $\tau = 0.2$	131
B.11	Relative rRMSE in simulation model 1 (non-Gaussian error) at $\tau = 0.3$	131
B.12	Relative rRMSE in simulation model 1 (non-Gaussian error) at $\tau = 0.4$	132

B.13	Relative rRMSE in simulation model 1 (non-Gaussian error) at $\tau = 0.6$.	132
B.14	Relative rRMSE in simulation model 1 (non-Gaussian error) at $\tau = 0.7$.	133
B.15	Relative rRMSE in simulation model 1 (non-Gaussian error) at $\tau = 0.8$.	133
B.16	The lower and upper bounds of relative rRMSE in simulation model 2 at three fixed endogeneity and varying heterogeneity.	136
B.17	The lower and upper bounds of relative rRMSE in simulation model 2 at three fixed heterogeneity and varying endogeneity.	137
B.18	Relative rRMSE of IVQR-2SLS averaging estimators in simulation model 3.	143
B.19	Relative rRMSE of IVQR-QR averaging estimators in simulation model 3.	144
C.1	The 45 countries in empirical example and their average democracy level indexed between 0 and 1.	181
C.2	The estimates of the effect oil wealth per capita of on military defense spending as ratio of GDP, its standard error, and 95% CI at ranks 0.1 to 0.9.	182

ABSTRACT

In the first chapter, I develop smoothed GMM estimation and inference for overidentified instrumental variables quantile regression (IVQR) and related quantile models. Previous approaches have been limited to exact identification, small models (due to computational burden), and/or iid sampling; we allow overidentification, large models, and non-iid sampling. I establish consistency and asymptotic normality of a GMM estimator with a general weighting matrix. The GMM framework provides the usual advantages: efficient two-step estimation, “distance metric” type inference, and overidentification testing. The new methods are illustrated in an application to consumption Euler equations under quantile utility maximization. Code is provided for all methods and applications.

In the second chapter I propose two averaging estimation methods to improve the finite-sample efficiency of the instrumental variables quantile regression (IVQR) estimator. I propose using the usual quantile regression for averaging to take advantage of cases when endogeneity is not too strong. I also propose using two-stage least squares to take advantage of cases when heterogeneity is not too strong. The first averaging method is to apply Cheng, Liao, and Shi’s (2019) averaging GMM framework to the IVQR model based on this proposed intuition. My implementation involves many computational considerations and builds on recent developments in the quantile literature. The second averaging method is a new bootstrap model averaging method that directly averages among IVQR, quantile regression, and two-stage least squares estimators. More specifically, I find the optimal weights in the bootstrap world and then apply the bootstrap-optimal weights to the original sample. The bootstrap method is simpler to compute and generally performs better in simulations, but it lacks the formal uniform dominance results of Cheng, Liao, and Shi (2019). Simulation results demonstrate that in the multiple-regressors/instruments case, both the GMM averaging and bootstrap

estimators have uniformly smaller risk than the IVQR estimator across data-generating processes with all kinds of combinations of different endogeneity levels and heterogeneity levels.

In the third chapter, I propose a quantile-based random coefficient panel data framework to study heterogeneous causal effects. The heterogeneity depends on unobservables, as opposed to heterogeneity for which we can add interaction terms. This connects it to other structural quantile models. My approach uses panel data to address “endogeneity,” meaning dependence between the explanatory variables and the random coefficients. The random coefficient vector depends on an unobserved, scalar, time-invariant “rank” variable, in which outcomes are monotonic at a particular point. I develop the theory first in a simplified model and then extend results to a more general model. First, I establish identification and uniformly consistent estimation. Second, I use a Dirichlet approach to establish small- n confidence sets or uniform confidence bands of the coefficient function. Third, I establish asymptotic normality of the coefficient estimator in the simplified model by applying the functional delta method to the empirical process. This facilitates a bootstrap confidence interval for the coefficient estimator at each specific rank. Finally, I illustrate the proposed methods by examining the causal effect of a country’s oil wealth on its political violence and military defense spending.

Chapter 1

Smoothed GMM for quantile models

1.1 Introduction

Economists and policymakers care about the causal effect of a policy. Most of the time, a policy has different effects on different groups of people. Koenker and Bassett (1978a) first introduced quantile regression, which measures a certain conditional quantile of the outcome variable. We are interested to know not only the causal effect of a policy on the mean of an outcome, but also the causal effect of a policy on certain quantile of the outcome distribution, conditional on observed characteristics, that is, the conditional quantile treatment effect.

The conditional quantile treatment effect (CQTE) is the difference between certain conditional quantile of potential outcomes under different states of treatment. However, in most cases, the treatment variables are associated with some unobserved characteristics that also affect the potential outcome.

To identify the conditional quantile treatment effect in presence of endogeneity, Chernozhukov and Hansen (2005a) introduce the instrument variable quantile regression (IVQR) model. It uses a rank similarity assumption to get a conditional moment equation. It proves that under full rank conditions, the parameters in this conditional

moment equation can be identified. Abadie, Angrist, and Imbens (2002a) propose an alternative way, the local quantile treatment effect (LQTE) model, to identify the quantile treatment effect (QTE). The LQTE estimand is interpreted as the QTE for the complier. Wüthrich (2016) shows a close connection between IVQR estimands and LQTE estimands. He shows the IVQR estimand is the same as the LQTE estimand at a transformed quantile level. This equivalence does not depend on rank similarity assumption. Based on Wüthrich (2016), even in the case that the IVQR model is misspecified, we could still estimate the IVQR model, and the estimates can be interpreted as the causal effect for the complier.

In practice, Chernozhukov and Hansen's (2006a) estimator computes very slowly when there are more than two endogenous regressors. To solve this computation difficulty, Kaplan and Sun (2017a) smooth the indicator function in the moments and get smoothed estimating equations (SEE) and SEE estimator. This estimator is simple and reliable to calculate, especially in the case of multiple endogenous regressors. This estimator has higher-order properties: a smaller asymptotic mean squared error than the unsmoothed estimator in Chernozhukov and Hansen (2006a).

Both Chernozhukov and Hansen (2006a) and Kaplan and Sun (2017a) limit to iid sampling and a linear model. De Castro, Galvao, and Kaplan (2017) generalize the Kaplan and Sun (2017a) model to a weaker assumption that allows for time-dependent data and nonlinear models, while still maintaining the idea of smoothing the moment condition. De Castro, Galvao, and Kaplan (2017) obtain a smoothed method of moments (MM) estimator and show the first-order results: the consistency and asymptotic normality of this MM estimator.

Both Kaplan and Sun (2017a) and de Castro, Galvao, and Kaplan (2017) work under the exact-identified case. This paper works in a more general GMM setting, allowing for the over-identified case. This paper keeps the main idea of smoothing the moments

as in Kaplan and Sun (2017a) and de Castro, Galvao, and Kaplan (2017), working in a general setting that allows for time-dependent data and nonlinear model. We have proved this smoothed two-step GMM estimator has consistency and asymptotic normality properties. We have also derive a distance metric test for hypothesis test and an over-identification test.

In our empirical example, we apply our smoothed two-step GMM method to estimate the quantile Euler equation, using the same data as in Yogo (2004). De Castro and Galvao (2017) develop a dynamic model to maximize the quantile utility. They derive the quantile Euler equation and claim it has the advantages of robustness to fat tail and no approximation error with log-linearization. We estimate the discount factor and elasticity of intertemporal substitution (EIS), and find economically reasonable estimates for certain quantiles even in the case when 2SLS estimates are economically unreasonable.

Section 1.2 describes one-step and two-step GMM estimators, with asymptotic theory in Section 1.3. Sections 1.4 and 1.5 describe testing of parameters and overidentifying restrictions, respectively. Section 1.6 illustrates the new methods empirically. Section 1.7 suggests directions for future research. Proofs are collected in the appendix.

Notation For scalar/vector/matrix variable formatting, \mathbf{X} is a random vector with elements X_j , \mathbf{x} is a non-random vector with elements x_j , Y and y are random and non-random scalars, respectively, and \mathbf{M} and \mathbf{m} are random and non-random matrices with row i , column j elements M_{ij} and m_{ij} . For vector/matrix multiplication, all vectors are treated as column vectors (which can be transposed to get row vectors). Also, $1 \mathcal{f} g$ is the indicator function, $E(\cdot)$ expectation, $Q_\tau(\cdot)$ the τ -quantile, $P(\cdot)$ probability, \doteq “is equal to, up to smaller-order terms,” “has exact (asymptotic) rate/order of,” and $N(\mu, \sigma^2)$ the normal distribution. For vectors, $\|\cdot\|$ is the Euclidean norm. Acronyms used include those for central limit theorem (CLT), continuous mapping theorem (CMT),

elasticity of intertemporal substitution (EIS), generalized method of moments (GMM), [smoothed] instrumental variables quantile regression ([S]IVQR), mean value theorem (MVT), probability density function (PDF), simulated annealing (SA), uniform law of large numbers (ULLN), and weak law of large numbers (WLLN).

1.2 Computation of GMM estimators

This section discusses computation of one-step and two-step GMM estimators. First, we describe the model setup and smoothing mechanism.

1.2.1 Setup

The parameter of interest $\theta_\tau \in B \subset \mathbb{R}^{d_\beta}$ satisfies the population moment condition

$$\mathbf{0} = E\{\mathbf{Z}_i [1\{f(\mathbf{Y}_i, \mathbf{X}_i, \theta_\tau) \leq 0\} - \tau]\}, \quad (1.1)$$

where τ is a given quantile index, $\mathbf{Z}_i \in Z \subset \mathbb{R}^{d_z}$ is the (full) instrument vector, $\mathbf{X}_i \in X \subset \mathbb{R}^{d_x}$ contains exogenous explanatory variables (i.e., a subset of \mathbf{Z}_i), $\mathbf{Y}_i \in Y \subset \mathbb{R}^{d_y}$ contains all endogenous variables, and $f(\cdot)$ is a known “residual function.” For example, a linear IVQR model with a single endogenous regressor Y_r instrumented for by excluded instrument Z_e would have \mathbf{Y} contain the outcome Y_o along with Y_r , $\mathbf{Z} = (\mathbf{X}, Z_e)$, and $f(\mathbf{Y}, \mathbf{X}, \mathbf{b}) = Y_o - (\mathbf{X}^r, Y_r)\mathbf{b}$.

We introduce some notation. Define the population map $\mathbf{M}: B \times \mathcal{T} \rightarrow \mathbb{R}^{d_z}$ as

$$\mathbf{M}(\theta, \tau) = E[\mathbf{g}_i^u(\theta, \tau)], \quad (1.2)$$

$$\mathbf{g}_i^u(\theta, \tau) = \mathbf{g}^u(\mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i, \theta, \tau) = \mathbf{Z}_i [1\{f(\mathbf{Y}_i, \mathbf{X}_i, \theta) \leq 0\} - \tau], \quad (1.3)$$

where superscript u denotes “unsmoothed.” The population moment condition (1.1) is

simply

$$\mathbf{M}(\beta_0, \tau) = \mathbf{0}. \quad (1.4)$$

The unsmoothed sample moment map is

$$\hat{\mathbf{M}}_n^u(\beta, \tau) = \hat{\mathbb{E}}[\mathbf{g}^u(\mathbf{Y}, \mathbf{X}, \mathbf{Z}, \beta, \tau)] = \frac{1}{n} \sum_{i=1}^n \mathbf{g}_i^u(\beta, \tau). \quad (1.5)$$

The well-known computational difficulty (e.g., Chernozhukov and Hong, 2003, Fig. 1(a) and Ex. 3, p. 297) of minimizing a GMM criterion based on $\hat{\mathbf{M}}_n^u(\beta, \tau)$ comes partly from the indicator function inside $\mathbf{g}_i^u(\beta, \tau)$. Our approach is to replace the indicator function with a smooth function, $\tilde{I}(\cdot)$.

The smoothed sample moment map is

$$\begin{aligned} \mathbf{g}_{ni}(\beta, \tau) &= \mathbf{g}_n(\mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i, \beta, \tau) - \mathbf{Z}_i [\tilde{I}(\|\mathbf{Y}_i - \mathbf{X}_i\|/h_n) - \tau], \\ \hat{\mathbf{M}}_n(\beta, \tau) &= \frac{1}{n} \sum_{i=1}^n \mathbf{g}_{ni}(\beta, \tau), \end{aligned} \quad (1.6)$$

where h_n is the bandwidth sequence and $\tilde{I}(\cdot)$ is a smoothed indicator (hence I) function.

The following example $\tilde{I}(\cdot)$ used in our code is shown in Figure 1.1:

$$\tilde{I}(u) = \begin{cases} 0 & \text{if } u \leq 0 \\ 0.5 + \frac{105}{64} \left(u - \frac{5}{3}u^3 + \frac{7}{5}u^5 - \frac{3}{7}u^7 \right) & \text{if } 0 < u < 1 \\ 1 & \text{if } 1 \leq u. \end{cases} \quad (1.7)$$

This $\tilde{I}(\cdot)$ is the integral of a fourth-order kernel function, which Kaplan and Sun (2017a) show leads to higher-order improvements in the linear iid IVQR setting.

In the exactly-identified case, the MM estimator solves the smoothed sample moment conditions,

$$\hat{\mathbf{M}}_n(\hat{\beta}_\tau, \tau) = \mathbf{0}. \quad (1.8)$$

In the over-identified case, there are d_Z moment conditions and d parameters, $d_Z > d$, so (1.8) cannot be solved. The following subsections describe our new GMM estimators

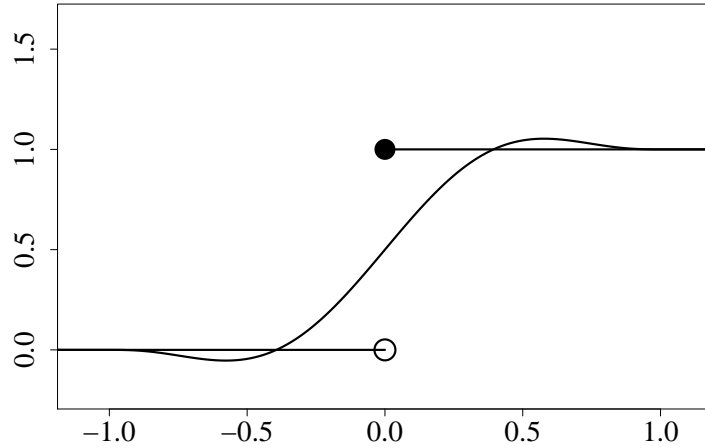


Figure 1.1: Graph of unsmoothed $1f - 0g$ and example smoothed version $\hat{f}(\cdot)$ from (1.7).

in the over-identified case.

1.2.2 One-step estimator

One way to estimate θ_τ in the over-identified case is to use the one-step estimator as in (3.11) in Newey and McFadden (1994, p. 2151). Let

$$\underline{\mathbf{G}} = \frac{1}{n} \sum_{i=1}^n \nabla_{\theta} \mathbf{g}_{ni}(\theta, \tau), \quad (1.9)$$

where ∇_{θ} denotes the partial derivative with respect to θ , and θ is an initial estimator consistent for θ_τ . With iid data, let

$$\underline{\Sigma} = \frac{1}{n} \sum_{i=1}^n \mathbf{g}_{ni}(\theta, \tau) \mathbf{g}_{ni}(\theta, \tau)' \quad (1.10)$$

be an estimator of $\underline{\Sigma} = E[\mathbf{g}_{ni}(\theta_\tau, \tau) \mathbf{g}_{ni}(\theta_\tau, \tau)']$. Then define the one-step estimator as

$$\hat{\theta}_{1s} = (\underline{\mathbf{G}}' \underline{\Sigma}^{-1} \underline{\mathbf{G}})^{-1} \underline{\mathbf{G}}' \underline{\Sigma}^{-1} \sum_{i=1}^n \mathbf{g}_{ni}(\theta, \tau) / n \quad (1.11)$$

For the initial $\tilde{\theta}$, we use the smoothed estimator from de Castro, Galvao, and Kaplan (2017), which they show is consistent even with weakly dependent data. De Castro,

Galvao, and Kaplan (2017) reduce the dimension of the instruments to the dimension of the parameter, so that the sample moment conditions may be solved exactly. For example, for a linear IVQR model, they project the regressors onto the space spanned by the instruments to get a new set of instruments for estimation.

For dependent data, the estimator $\underline{\Sigma}$ in (1.10) can be replaced with a long-run variance estimator. For example, de Castro, Galvao, and Kaplan (2017) suggest the estimator from Newey and West (1987) and Andrews (1991) with a Bartlett kernel, which we use in our code. However, a rigorous consistency proof of that estimator in our setting remains an open question.

1.2.3 Two-step estimator

Another way to estimate $\beta_{0\tau}$ in (1.1) in the over-identified case is to use a two-step estimator. Let $\hat{\mathbf{W}}$ be a symmetric, positive definite weighting matrix. The smoothed GMM (minimum distance) estimator minimizes a weighted quadratic norm of the smoothed sample moment vector:

$$\hat{\beta}_{\text{GMM}} = \arg \min_{2B} \left[\sum_{i=1}^n \mathbf{g}_{ni}(\beta, \tau) \right]^{\top} \hat{\mathbf{W}} \left[\sum_{i=1}^n \mathbf{g}_{ni}(\beta, \tau) \right] = \arg \min_{2B} \hat{\mathbf{M}}_n(\beta, \tau)^{\top} \hat{\mathbf{W}} \hat{\mathbf{M}}_n(\beta, \tau). \quad (1.12)$$

The optimal weighting matrix is an estimator of the inverse long-run variance of the sample moments: $\hat{\mathbf{W}} = \underline{\Sigma}^{-1} \hat{\rho} \underline{\Sigma}^{-1}$, where $\underline{\Sigma}$ is an estimator depending on an initial estimate β_{τ} , as in (1.10) for the iid case. The efficient two-step GMM estimator is thus

$$\hat{\beta}_{2s} = \arg \min_{2B} \hat{\mathbf{M}}_n(\beta, \tau)^{\top} \underline{\Sigma}^{-1} \hat{\mathbf{M}}_n(\beta, \tau). \quad (1.13)$$

To compute the GMM estimator in (1.13), a global minimum must be found numerically. This global minimization is difficult due to the objective function's possible nonconvexity. To find the global minimum, we use the simulated annealing algorithm from the GenSA package in R (Xiang, Gubian, Suomela, and Hoeng, 2013), which is

suitable to such problems.¹ Despite its strengths, simulated annealing cannot reliably solve (1.13) without help. This is because within a reasonable amount of time, the simulated annealing method can only give us a local minimum instead of a global minimum. To get a global minimum, the simulated annealing method requires computation at a large amount of points so that the computation time takes longer. Thankfully, it is enough help to set the initial parameter value (for the numerical search) to be the estimator from de Castro, Galvao, and Kaplan (2017).

After computing (1.13), one may optionally run simulated annealing again with the *unsmoothed* objective function. Kaplan and Sun (2017a) show that the smoothed estimator may be preferred in terms of (pointwise in τ) mean squared error in the linear iid IVQR setting, but they also show that smoothing may diminish the estimated heterogeneity, so whether or not to run a final unsmoothed minimization may depend on the application.

1.3 Asymptotic theory

Under exact identification, de Castro, Galvao, and Kaplan (2017) establish consistency and asymptotic normality of the smoothed method of moments estimator. In this section, we generalize to the over-identified case.

We first introduce some assumptions, which are mostly the same as in de Castro, Galvao, and Kaplan (2017). Under these assumptions, we show the consistency and asymptotic normality of the two-step GMM estimator in (1.13).

Assumption 1. For each observation i among n in the sample, endogenous vector $\mathbf{Y}_i \in \mathcal{Y} \subset \mathbb{R}^{d_Y}$ and instrument vector $\mathbf{Z}_i \in \mathcal{Z} \subset \mathbb{R}^{d_Z}$; a subset of \mathbf{Z}_i is $\mathbf{X}_i \in \mathcal{X} \subset \mathbb{R}^{d_X}$, with $d_X \leq d_Z$. The sequence $\{f(\mathbf{Y}_i, \mathbf{Z}_i)\}$ is strictly stationary and weakly dependent.

¹We thank Ron Gallant for the suggestion of simulated annealing.

Assumption 2. The function $f : Y \times X \rightarrow B \subseteq \mathbb{R}$ is known and has (at least) one continuous derivative in its B argument for all $\mathbf{y} \in Y$ and $\mathbf{x} \in X$.

Assumption 3. The parameter space $B \subseteq \mathbb{R}^{d_\beta}$ is compact; $d_\beta \geq d_Z$. Given $\tau \in (0, 1)$, the population parameter $\theta_{0\tau}$ is in the interior of B and uniquely satisfies the moment condition

$$\mathbf{0} = E[\mathbf{Z}_i(1\{f(\mathbf{Y}_i, \mathbf{X}_i, \theta_{0\tau}) = 0\} - \tau)]. \quad (1.14)$$

Assumption 4. The matrix $E(\mathbf{Z}_i \mathbf{Z}_i^\top)$ is positive definite (and finite).

Assumption 5. The function $F(\cdot)$ satisfies $F(u) = 0$ for $u \leq -1$, $F(u) = 1$ for $u \geq 1$, and $-1 < F(u) < 1$ for $-1 < u < 1$. The derivative $F^\theta(\cdot)$ is a symmetric, bounded kernel function of order $r \geq 2$, so $\int_{-1}^1 F^\theta(u) du = 1$, $\int_{-1}^1 u^k F^\theta(u) du = 0$ for $k = 1, \dots, r-1$, and $\int_{-1}^1 |ju^r F^\theta(u)| du < 1$ but $\int_{-1}^1 u^r F^\theta(u) du \neq 0$, and $\int_{-1}^1 |ju^{r+1} F^\theta(u)| du < 1$.

Assumption 6. The bandwidth sequence h_n satisfies $h_n = o(n^{-1/(2r)})$.

Assumption 7. Given any $B \subseteq \mathbb{R}^{d_\beta}$ and almost all $\mathbf{Z}_i = \mathbf{z}$ (i.e., up to a set of zero probability), the conditional distribution of $(\mathbf{Y}_i, \mathbf{X}_i)$ given $\mathbf{Z}_i = \mathbf{z}$ is continuous in a neighborhood of zero.

Assumption 8. For a fixed $\tau \in (0, 1)$, using the definition in (1.6),

$$\sup_{B} E\|\hat{\mathbf{M}}_n(\cdot, \tau) - E[\hat{\mathbf{M}}_n(\cdot, \tau)]\| = o_p(1). \quad (1.15)$$

Assumption 9. Let $\eta_i = f(\mathbf{Y}_i, \mathbf{X}_i, \theta_{0\tau})$ and $\mathbf{D}_i = \mathbf{r}(\mathbf{Y}_i, \mathbf{X}_i, \theta_{0\tau})$, using the notation

$$\nabla = (\mathbf{y}, \mathbf{x}, \theta) \left. \frac{\partial}{\partial \theta} f(\mathbf{y}, \mathbf{x}, \theta) \right|_{\theta = \theta_{0\tau}}, \quad (1.16)$$

for the $d_\beta - 1$ partial derivative vector. Let $f_{j\mathbf{Z}}(j, \mathbf{z})$ denote the conditional PDF of η_i given $\mathbf{Z}_i = \mathbf{z}$, and let $f_{j\mathbf{Z}, \mathbf{D}}(j, \mathbf{z}, \mathbf{d})$ denote the conditional PDF of η_i given $\mathbf{Z}_i = \mathbf{z}$ and $\mathbf{D}_i = \mathbf{d}$. (i) For almost all \mathbf{z} and \mathbf{d} , $f_{j\mathbf{Z}}(j, \mathbf{z})$ and $f_{j\mathbf{Z}, \mathbf{D}}(j, \mathbf{z}, \mathbf{d})$ are at least r

times continuously differentiable in a neighborhood of zero, where the value of r is from Assumption 5. For almost all $\mathbf{z} \in Z$ and u in a neighborhood of zero, there exists a dominating function $C(\cdot)$ such that $|\frac{\partial^r}{\partial z^r}(u f(\mathbf{z}))| \leq C(\mathbf{z})$ and $E[C(\mathbf{Z})] < \infty$. (ii) The matrix

$$\mathbf{G} = \frac{\partial}{\partial \beta} E[\mathbf{Z}_i' f(\mathbf{Y}_i, \mathbf{X}_i, \tau) | \tau] = E\{\mathbf{Z}_i \mathbf{D}_i' f(\mathbf{Z}_i, \mathbf{D}_i) | \tau\} \quad (1.17)$$

has rank d_β .

Assumption 10. A pointwise CLT applies:

$$\sqrt{n}(\hat{\mathbf{M}}_n(\tau) - E[\hat{\mathbf{M}}_n(\tau)]) \xrightarrow{d} N(\mathbf{0}, \underline{\Sigma}_\tau). \quad (1.18)$$

Assumption 11. Let $Z_i^{(k)}$ denote the k th element of \mathbf{Z}_i , and similarly $\beta^{(k)}$. Let G_{kj} denote the row k , column j element of \mathbf{G} (from Assumption 9). Assume

$$\frac{1}{nh_n} \sum_{i=1}^n \mathbf{I}^j(\mathbf{Y}_i, \mathbf{X}_i, \tilde{\tau}_{\tau,k}) Z_i^{(k)} \frac{\partial}{\partial \beta^{(j)}} f(\mathbf{Y}_i, \mathbf{X}_i, \tau) \Big|_{\tau=\tilde{\tau}_{\tau,k}} \xrightarrow{p} G_{kj}. \quad (1.19)$$

for each $k = 1, \dots, d_\beta$ and $j = 1, \dots, d_\beta$, where each $\tilde{\tau}_{\tau,k}$ lies between τ and $\hat{\tau}$ (defined in Assumption 3 and (1.8), respectively).

Assumption 12. For the weighting matrix, $\hat{\mathbf{W}} \xrightarrow{p} \mathbf{W}$, and both are symmetric, positive definite matrices.

Other than Assumption 12, all the other assumptions follow from de Castro, Galvao, and Kaplan (2017). We give a brief comment of the assumptions here. Assumption 1 generalizes iid sampling in Chernozhukov and Hansen (2006a) to allow for strictly stationary and weakly dependent data. Assumption 2 generalizes Kaplan and Sun's (2017a) linear model to be nonlinear. Assumption 3 assumes global identification of parameter β . Assumption 4 assumes the instruments \mathbf{Z} have finite second moment. Assumption 5 sets up the smoothed kernel function to replace the indicator function. Assumption 6 imposes restrictions on bandwidth to make the effect of

smoothing asymptotically negligible. Assumption 7 assumes the conditional distribution of residual given instruments is continuous at zero. Assumption 8 assumes a ULLN property of smoothed moments. Assumptions 9–11 are used for asymptotic normality results. Assumption 9 enables local identification. Assumption 10 assumes a CLT property of smoothed moments. Assumption 11 ensures the consistency of covariance matrix estimator. A detailed discussion of these assumptions is in de Castro, Galvao, and Kaplan (2017).

Using Theorem 5.7 in van der Vaart (1998a), we show the GMM estimator is consistent.

Theorem 1. *Under Assumptions 1–8 and 12, the estimator from (1.12) is consistent:*

$$\hat{\beta}_{\text{GMM}}(\tau) = o_p(1). \quad (1.20)$$

The GMM estimator is also asymptotically normal.

Theorem 2. *Under Assumptions 1–12, the asymptotic distribution of the GMM estimator in (1.12) is*

$$\sqrt{n}(\hat{\beta}_{\text{GMM}}(\tau) - \beta) \xrightarrow{d} N(\mathbf{0}, (\mathbf{G}'\mathbf{W}\mathbf{G})^{-1}\mathbf{G}'\mathbf{W}\Sigma_{\tau}\mathbf{W}\mathbf{G}(\mathbf{G}'\mathbf{W}\mathbf{G})^{-1}),$$

where \mathbf{G} is defined in (1.17) and $\Sigma_{\tau} = \lim_{n \rightarrow \infty} \text{Var}(n^{-1/2} \sum_{i=1}^n \mathbf{g}_{ni}(\beta(\tau), \tau))$.

1.4 Distance metric hypothesis testing

We propose a distance metric (DM) test as discussed in Section 9 of Newey and McFadden (1994), for example. The smoothing we introduce makes possible an expansion like in Newey and McFadden (1994, p. 2223). The expansion leads to the limiting χ^2 distribution in (1.25) below.

Previously, de Castro, Galvao, and Kaplan (2017) only propose a Wald test because

they did not have a way to compute the constrained estimator necessary for the DM test. Using our approach, we can compute the constrained estimator and thus the DM statistic, and we prove its limiting distribution in Theorem 3. Compared to the Wald test, which did not perform well in the simulations of de Castro, Galvao, and Kaplan (2017), the DM test has the well-known advantages of invariance to reparameterization and not depending directly on an estimated covariance matrix.

Computationally, the simulated annealing algorithm we use does not allow constraints, but the most common null hypothesis constraints are easily accommodated by modifying the data matrices. For example, in a linear model, to impose the restriction that some coefficient is equal to zero, the corresponding column from the regressor matrix is simply deleted. To improve numerical performance, we use the unconstrained GMM estimator as the initial value.

Consider the r -dimensional null hypothesis

$$H_0: \underbrace{\mathbf{a}(\theta_\tau)}_{r \times 1} = \mathbf{0}. \quad (1.21)$$

We make the standard assumptions about $\mathbf{a}(\cdot)$ and its Jacobian.

Assumption 13. Assume $\mathbf{a}(\cdot)$ is continuously differentiable at θ_τ , and let

$$\hat{\mathbf{A}}_{r \times d_\beta} = \frac{\partial}{\partial \theta_\tau} \mathbf{a}(\cdot) \Big|_{\theta_\tau} \stackrel{p}{\rightarrow} \mathbf{A}_{r \times d_\beta} = \frac{\partial}{\partial \theta_\tau} \mathbf{a}(\cdot) \Big|_{\theta_\tau}, \quad \text{rank}(\mathbf{A}) = r = d_\beta. \quad (1.22)$$

As a special case, (1.21) includes linear hypotheses of the form $\mathbf{A} \theta_\tau = \mathbf{c}$ when $\mathbf{a}(\cdot) = \mathbf{A} \theta_\tau + \mathbf{c}$. Another special case is the full vector null hypothesis $H_0: \theta_\tau = \mathbf{b}$ when $\mathbf{a}(\cdot) = \mathbf{b}$.

Let $\hat{\theta}_n$ be the constrained two-step estimator,

$$\hat{\theta}_n = \arg \min_{\theta_\tau: \mathbf{a}(\theta_\tau) = \mathbf{0}} \hat{\mathbf{M}}_n(\hat{\theta}_n, \tau) \stackrel{1}{\geq} \hat{\mathbf{M}}_n(\hat{\theta}_n, \tau). \quad (1.23)$$

where as before $\hat{\Sigma}$ is an estimator of Σ_τ based on initial estimator $\hat{\theta}_n$.

Let the DM statistic be

$$\widehat{DM}_n = n\widehat{M}_n(\hat{\tau}) >_{\Sigma} {}^1\widehat{M}_n(\hat{\tau}) - n\widehat{M}_n(\hat{\tau}_{GMM}) >_{\Sigma} {}^1\widehat{M}_n(\hat{\tau}_{GMM}). \quad (1.24)$$

Theorem 3 gives the asymptotic distribution of \widehat{DM}_n .

Theorem 3. *Under Assumptions 1–13, under the null hypothesis in (1.21), given $\Sigma \neq \Sigma_\tau$ and the DM statistic definition in (1.24),*

$$\widehat{DM}_n \xrightarrow{d} \chi_r^2. \quad (1.25)$$

1.5 Overidentification test

Given our two-step GMM estimator, we can formulate a J -test for overidentifying restrictions as in Hansen (1982, Lem. 4.2).² Our proof follows the same structure as in Section 9.5 of Newey and McFadden (1994).

The null hypothesis is that the unconditional moment conditions are satisfied by the (pseudo)true population parameter,

$$H_0: E[\mathbf{g}_i^u(\tau)] = \mathbf{0}, \quad \mathbf{g}_i^u(\tau) = \mathbf{Z}_i[1\{(\mathbf{Y}_i, \mathbf{X}_i) \in \mathcal{C}\} - \tau]. \quad (1.26)$$

The usual overidentification testing caveats apply. For example, rejection of H_0 may be due to misspecification of \mathcal{C} (e.g., misspecification of the structural functional form) rather than an invalid instrument. Also, non-rejection should not be interpreted as strong evidence of instrument validity when power may be low (e.g., in small samples).

Define the J -statistic to be

$$\hat{J}_n = n\widehat{M}_n(\hat{\tau}_{2S}) >_{\Sigma} {}^1\widehat{M}_n(\hat{\tau}_{2S}). \quad (1.27)$$

²Chernozhukov, Hansen, and Wüthrich (2017) note this possibility, writing, “We note that if the GMM problem is overidentified, overidentification-type tests can be used to assess the joint validity of the underlying assumptions.”

Theorem 4. *Under Assumptions 1–12, under the null hypothesis in (1.26), with $\underline{\Sigma}_\tau$ defined in Theorem 2, consistent estimator $\underline{\Sigma}^P$ $\underline{\Sigma}_\tau$, and $\hat{\alpha}_{2s}$ defined in (1.13), the J -statistic defined in (1.27) has a limiting chi-square distribution with $d_Z - d_\beta$ degrees of freedom: $\hat{J}_n \xrightarrow{d} \chi_{d_Z - d_\beta}^2$.*

1.6 Empirical example: quantile Euler equations

In this section, we apply our smoothed two-step GMM method to an empirical example: to estimate the quantile Euler equation. The quantile Euler equation is derived from the theoretical foundation of quantile utility maximization in de Castro and Galvao (2017). The setup is the same as de Castro and Galvao (2017), but the method and results are new.

De Castro and Galvao (2017) develop a dynamic model in which an agent has a quantile utility preference rather than a conventional expected utility preference. That is, the agent is assumed to maximize its τ -quantile of the stream of future utility every period,

$$\max U(C_0) + Q_\tau[\beta_\tau U(C_1)] + Q_\tau[\beta_\tau^2 U(C_2)] + Q_\tau[\beta_\tau^3 U(C_3)] + \dots + Q_\tau[\beta_\tau^{j-2} U(C_{j-1})] + Q_\tau[\beta_\tau^{j-1} U(C_j)],$$

while subject to a linear budget constraint,

$$C_t + x_{t+1}p_t = x_t(p_t + d_t)$$

where C_t is the consumption at time t ; x_t is the amount of asset he invests in time t ; p_t is the price of the asset at time t ; d_t is the dividend value of the asset at time t . The agent makes an inter-temporal allocation between consumption and saving in every period, based on quantile utility maximization.

De Castro and Galvao (2017) solve this dynamic model and derive the corresponding quantile Euler equation in equilibrium.

$$\mathbb{Q}_\tau \left[\beta_\tau (1 + r_{t+1}) \frac{U'(C_{t+1})}{U'(C_t)} \mid \mathcal{I}_t \right] = 1. \quad (1.28)$$

where \mathbb{Q}_τ means the τ -quantile operator, β is the discount factor, r_{t+1} is the interest rate at time $t + 1$, C_t is the consumption level in period t , and \mathcal{I}_t is the information set at time t . The quantile Euler equation looks similar to standard expected utility-based Euler equation, but with a quantile operator instead of expectation operator.

Our interest is to investigate the intertemporal allocation of consumption and estimate the elasticity of intertemporal substitution (EIS) for a quantile utility maximizer. We follow the literature to use an isoelastic utility function:

$$U(C_t) = \frac{C_t^{1-\gamma}}{1-\gamma}, \quad \gamma > 0, \quad (1.29)$$

where γ is our parameter of interest since $1/\gamma$ is interpreted as the EIS.

With an isoelastic utility function, the quantile Euler equation becomes

$$\mathbb{Q}_\tau [\beta_\tau (1 + r_{t+1}) (C_{t+1}/C_t)^{-\gamma} \mid \mathcal{I}_t] = 0. \quad (1.30)$$

Because of the equivariance property of the quantile operator, we can log-linearize the quantile Euler equation with no approximation error:

$$\mathbb{Q}_\tau [\ln \beta_\tau + \ln(1 + r_{t+1}) - \gamma \ln(C_{t+1}/C_t) \mid \mathcal{I}_t] = 0. \quad (1.31)$$

We could rewrite it as

$$\mathbb{Q}_{1-\tau} [\ln(C_{t+1}/C_t) - \gamma^{-1} \ln \beta_\tau - \gamma^{-1} \ln(1 + r_{t+1}) \mid \mathcal{I}_t] = 0. \quad (1.32)$$

The constant and slope of the $1 - \tau$ IV quantile regression of $\ln(C_{t+1}/C_t)$ onto a constant and $\ln(1 + r_{t+1})$ are $\gamma^{-1} \ln \beta_\tau$ and γ^{-1} . By definition, EIS captures the responsiveness of the growth rate of consumption to the real interest rate. In the special case of isoelastic utility function, the slope ($1/\gamma$) of this IV quantile regression can be interpreted as the

EIS.

One benefit of this quantile utility maximization framework is that the quantile level τ also measures an agent's risk attitude. Both γ and τ measure the risk attitude. Thus we are able to separate two notions: EIS ($1/\gamma$) and risk attitude (τ). In contrast, in an expected utility framework with isoelastic utility, there is no separation between EIS ($1/\gamma$) and risk attitude (γ). Other advantages of the quantile Euler equation include its robustness to fat tails of consumption levels and no approximation error with log-linearization.

We use the same dataset as in Yogo (2004), which is from Campbell (2003) originally. It consists of the aggregate quarterly data of consumption and real interest rate in eleven developed countries. Our empirical example uses data for four countries: the United States (USA), the United Kingdom (UK), Australia (AUS), and Sweden (SWE). The real interest rate is calculated as the short-term interest rate deflated by consumer price index. The aggregate consumption level is calculated as the total consumption for all countries, except for US that the aggregate consumption is calculated as the sum of the nondurable goods and services due to data availability. Following conventions, the aggregate consumption is measured at the beginning of each period. The sample period for the four countries are: USA 1947Q1-1995Q4, UK 1970Q1-1999Q1, AUS 1970Q1-1998Q4, and SWE 1970Q1-1997Q4.

Our empirical analysis follows Table 2 of Yogo (2004), but with quantile estimates. Yogo (2004) uses a 2SLS method to the log-linearized model $\ln(C_{t+1}/C_t) = \delta_0 + \delta_1 \ln(1 + r_{t+1}) + u_{t+1}$ and obtains estimates of the EIS. The instruments for the endogenous real interest rate consist of the twice lagged nominal interest rate, twice lagged dividend-price ratio, twice lagged inflation rate, and twice lagged log consumption growth.

Table 1 shows the estimated β for four countries for different τ in the quantile utility maximization model as well as the 2SLS estimates for the expected utility maximization

Table 1.1: Smoothed MM and GMM estimates of β_τ , log-linear model.

τ	US		UK		AUS		SWE	
	$\hat{\beta}_{MM}$	$\hat{\beta}_{GMM}$	$\hat{\beta}_{MM}$	$\hat{\beta}_{GMM}$	$\hat{\beta}_{MM}$	$\hat{\beta}_{GMM}$	$\hat{\beta}_{MM}$	$\hat{\beta}_{GMM}$
0.1	0.92	0.92	0.00	0.12	0.88	0.91	0.95	0.94
0.2	0.90	0.92	1.86	0.75	0.88	0.87	0.95	0.95
0.3	0.85	1.13	1.16	0.82	0.82	0.75	0.95	0.95
0.4	0.11	1.04	1.07	0.87	2.38	1.35	0.95	0.95
0.5	1.14	1.02	1.04	1.03	1.13	1.09	0.90	0.80
0.6	1.04	1.01	1.02	1.01	1.02	1.03	0.94	1.07
0.7	1:02	1:01	1:00	1:00	1:00	1:00	0:97	0:97
0.8	1:00	1:00	0:98	0:99	0:98	0:99	0:97	0:97
0.9	0.99	1.00	0.96	0.97	0.96	0.98	0.97	0.96
2SLS	1.08		1.03		1.11		0.27	

Table 1.2: Smoothed MM and GMM estimates of γ_τ , log-linear model.

τ	US		UK		AUS		SWE	
	$\hat{\gamma}_{MM}$	$\hat{\gamma}_{GMM}$	$\hat{\gamma}_{MM}$	$\hat{\gamma}_{GMM}$	$\hat{\gamma}_{MM}$	$\hat{\gamma}_{GMM}$	$\hat{\gamma}_{MM}$	$\hat{\gamma}_{GMM}$
0.1	7.2	7.1	744.0	114.1	6.9	5.5	3.4	3.7
0.2	11.5	8.0	43.1	19.0	9.0	9.1	4.0	4.4
0.3	20.0	16.5	13.5	15.7	17.0	24.6	5.5	6.1
0.4	342.5	6.9	8.3	14.1	98.9	33.9	8.7	10.2
0.5	26.5	4.5	7.3	5.3	20.7	14.8	38.7	89.5
0.6	10.9	3.7	6.6	4.1	7.3	8.0	151.0	301.5
0.7	6:6	3:4	5:9	3:7	5:9	5:1	13:9	13:6
0.8	4:9	3:2	5:3	3:7	5:0	3:7	5:3	6:4
0.9	4.9	3.1	5.0	4.0	4.4	0.0	3.3	4.8
2SLS	16.7		6.0		22.1		544.4	

model. Similar to Table 1, Table 2 shows the corresponding γ estimates. As we know, the EIS should be positive to be economically meaningful. Thus, our γ should also be positive to be economically reasonable.

From Table 2, we can see for upper level quantile indices, i.e., τ greater than or equal to 0.7, we obtain reasonable estimates of γ and EIS. But for lower level quantile indices, especially when τ is less than or equal to 0.4, the estimates are economically implausible.

To compare the results of 2SLS estimates and two-step GMM estimates, we could also see that the $\tau = 0.5$ model has similar estimates with 2SLS estimates. The model with an agent maximizing the $\tau = 0.7$ or 0.8 quantile of utility may better fit the data. We can see that for the case where Sweden has economically implausible 2SLS estimates ($\gamma = 544.4, \beta = 0.27$), our quantile model can still obtain economically reasonable estimates for upper τ levels ($\tau = 0.7$).

The MM estimates are the estimates in the exact-identified case. When we have more instruments than parameters to be estimated, we could project all the regressors onto the space spanned by the full set of instruments and use the projected values as the instruments in the exact-identified case. For most cases, the MM estimates and GMM estimates are similar, especially in the case $\tau = 0.7, 0.8$ that we are more interested in. However, for US $\tau = 0.7$ case, there is a significant difference between MM and two-step GMM estimates.

1.7 Conclusion

In this paper, we introduce the smoothed two-step GMM method to improve efficiency. We derive the asymptotic properties, distance metric hypothesis test and overidentification test for this smoothed two-step GMM estimator. It works in a general setting, allowing nonlinear models, time-dependent data, and over-identification. Computation is relatively faster and reliable, even with multiple endogenous regressors.

We use this method to estimate a quantile Euler equation, obtaining economically reasonable estimates at certain quantile levels even when the 2SLS estimates are implausible.

Future work would extend this paper to averaging IVQR estimators like in Hansen (2017a) and Cheng, Liao, and Shi (2016) to improve efficiency. This is in progress. Further work would use J -statistics to empirically determine τ , complete a rigorous

LRV consistency proof, and extend to a nonparametric model.

Chapter 2

Averaging estimation for instrumental variables quantile regression

2.1 Introduction

Since Koenker and Bassett's (1978b) seminal work, quantile regression (QR) has become a useful tool to capture unobserved heterogeneous effects in policy analysis or program evaluation. However, in practice endogeneity commonly results in inconsistent estimates of the conventional quantile regression. To address this endogeneity issue in quantile regression, Chernozhukov and Hansen (2005b) propose an instrumental variables method to identify the structural quantile function or (conditional) quantile treatment effects. Subsequently, Chernozhukov and Hansen (2006b) and then many others have provided estimation methods for the instrumental variables quantile regression (IVQR) model. Although there are important distinctions, for now I simply refer to "the" IVQR estimator.

Although the IVQR estimator has desirable large-sample properties like consistency and asymptotic normality, it can have imprecise estimates due to its large finite-sample

variance,¹ just as the two-stage least squares (2SLS) estimator may have a substantial finite-sample dispersion.

I propose two new methods to improve the finite-sample estimation efficiency of IVQR. The first applies the averaging generalized method of moments (GMM) framework of Cheng, Liao, and Shi (2019) (hereafter CLS) to IVQR. The second is a bootstrap averaging method, which uses the bootstrap world's optimal averaging weights on IVQR, QR, and 2SLS estimators.

I make four main contributions. First, I consider QR and 2SLS in averaging to improve on IVQR, for reasons detailed below. Second, the implementation of the CLS averaging GMM framework is not trivial for IVQR. In particular, it needs two-step GMM estimation and nonparametric Jacobian matrix estimation, which are discussed or extended in my paper (and my code). Third, the bootstrap averaging method is new and outperforms the CLS averaging GMM method in simulations, although it lacks theoretical results like in CLS. Fourth, my simulations compare various methods in a wide range of data-generating process (DGP) types (varying endogeneity, heterogeneity, distributional shapes, etc.), although they are still limited.

My use of QR and 2SLS is motivated by the bias–variance tradeoff. When there is not much endogeneity, the conventional QR estimator is not very biased, and it usually has smaller variance than the IVQR estimator. Therefore, averaging IVQR with QR can reduce variance and may achieve an overall reduction in mean squared error (MSE), although it might increase the bias. When there is not much heterogeneity across quantiles, 2SLS and IVQR (at any quantile) will have similar slope estimands; at the same time, 2SLS usually has smaller variance than IVQR, except in cases like a fat-tailed error term. Therefore, using 2SLS to average IVQR can also improve efficiency by reducing the overall MSE.

¹It is possible that the IVQR variance is infinite, similar to IV (Kinal, 1980). More technically, the word “variance” in this paper means a different measure of dispersion/spread that is never infinite, like trimmed/truncated variance or interquartile range.

The bootstrap averaging method is new. It averages the IVQR, QR, and 2SLS estimators in the bootstrap world with a grid of fixed weights, and picks the weight that minimizes the robust root mean squared error (robust RMSE, or rRMSE) as the bootstrap optimal weight. This optimal weight is then used to average the IVQR, QR, and 2SLS estimators in the original sample to obtain the bootstrap averaging estimator.

The bootstrap averaging method, although lacking theoretical results, has potential advantages over the CLS averaging GMM method when applied to IVQR. Bootstrap averaging is easier for computation since it avoids highly over-identified quantile GMM, which has a difficult criterion function to minimize. Moreover, the bootstrap averaging estimator has better performance in simulation results. This is partly because the bootstrap can average among three different estimators, whereas CLS only averages between two estimators. It might also be because often in finite samples QR outperforms the aggressive IVQR-QR GMM estimator and 2SLS outperforms the aggressive IVQR-2SLS GMM estimator.

Literature

Averaging estimation Averaging estimation originates from Stein-like shrinkage estimation and has recently been reinvestigated and extended by many authors to improve estimation efficiency. Stein and James (1961) propose an estimator that shrinks the least squares (LS) estimator toward zero. The James–Stein estimator can be viewed as in the class of averaging estimation, in that it averages the LS estimator and zero. Under certain conditions, the James–Stein estimator dominates the LS estimator in terms of a uniform reduction of MSE. Maasoumi (1978) applies the idea of Stein-like shrinkage estimation to simultaneous equations by averaging the three-stage least squares (3SLS) estimator with the LS estimator. He shows while the 3SLS and 2SLS estimators have no finite moments (therefore, unbounded risk) in some cases, the averaging estimator has finite moments (therefore, bounded risk). Recently, Hansen (2017b) applies averaging

estimation to the usual IV model. He averages 2SLS and OLS estimators with weight depending on the statistic for testing exogeneity in the model. He shows that the averaging estimator uniformly dominates 2SLS in asymptotic risk when the number of endogenous regressors is greater than two. This result allows non-normal distributions but requires homoskedasticity. The averaging GMM method of CLS works in a general framework with no normality or homoskedasticity restriction. CLS provide supporting simulations to show the uniform dominance results in their averaging GMM framework can hold with both Gaussian and non-Gaussian errors, when there are multiple endogenous regressors.

IVQR estimation and computation Since Chernozhukov and Hansen's (2005b) seminal work on IVQR identification, many studies have focused on IVQR estimation methods and improving computational efficiency. To the best of my knowledge, besides the one exception noted below, there are no studies on improving IVQR estimation efficiency in finite samples. My paper helps fill this gap.

The challenging computation of IVQR estimation comes from the non-differentiable IVQR moments and the non-convex GMM objective function. Chernozhukov and Hansen (2006b) first propose a two-step inverse quantile regression method with a grid search on the endogenous regressors' coefficients to compute the IVQR estimator. This IVQR estimator is asymptotically equivalent to a GMM estimator. However, its computation time scales poorly with the number of endogenous regressors. Chen and Lee (2018) propose an exact GMM estimator using mixed-integer quadratic programming, but it also has long computation time. Zhu (2019) proposes a k -step correction approach using mixed integer linear programming. This estimator is asymptotically equivalent to the GMM estimator and has computational efficiency in models with multiple endogenous regressors. Kaido and Wüthrich (2021) decompose IVQR estimation into conventional QR sub-problems. Kaplan and Sun (2017b) propose smoothing the IVQR moments, which helps both estimation and computational efficiency. However, their results only apply to

an exactly-identified linear model and iid sampling. De Castro, Galvao, Kaplan, and Liu (2019) propose a GMM estimator using the same smoothed IVQR moments, extending results to over-identified nonlinear models and dependent data. I use de Castro et al.’s (2019) method to interpret IVQR as a GMM estimator. This facilitates using CLS’s averaging GMM framework.

Outline Section 2.2 describes the model setup. Section 2.3 proposes the components for averaging to improve IVQR. Section 2.4 implements the GMM averaging estimation method. Section 2.5 introduces the bootstrap averaging estimation method. Section 2.6 presents simulation results. Section 2.7 concludes. Proofs and additional computational details are collected in the appendix. Code is provided for all methods and simulations.

Notation For scalar/vector/matrix variable formatting, \mathbf{X} is a random vector with elements X_j , \mathbf{x} is a non-random vector with elements x_j , Y and y are random and non-random scalars, respectively, and \mathbf{M} and \mathbf{m} are random and non-random matrices with row i , column j elements M_{ij} and m_{ij} . For vector/matrix multiplication, all vectors are treated as column vectors. Also, $1_f g$ is the indicator function, $E(\cdot)$ expectation, $Q_\tau(\cdot)$ the τ -quantile, $P(\cdot)$ probability, and $N(\mu, \sigma^2)$ the normal distribution. Acronyms used include those for instrumental variables (IV), two-stage least squares (2SLS), generalized method of moment (GMM), [smoothed] instrumental variables quantile regression ([S]IVQR), probability density function (PDF), cumulative distribution function (CDF), quantile regression (QR), conditional quantile function (CQF), data-generating process (DGP), [root] mean squared error ([R]MSE), asymptotic mean squared error (AMSE), and interquantile range (IQR).

2.2 Model setup

We are interested in estimating the parameter $\beta_{0\tau} \in \mathbb{R}^{d_\beta}$ in a linear quantile model that uniquely satisfies the conditional probability

$$\tau = P(Y_i \leq \beta_{0\tau} + \mathbf{X}_i' \boldsymbol{\beta}_{0\tau} | \mathbf{Z}_i), \quad (2.1)$$

where $\tau \in (0, 1)$ is a given quantile level; Y_i is the outcome variable; $\mathbf{X}_i = (\mathbf{X}_{exog,i}', \mathbf{D}_i')' \in \mathbb{R}^{d_x}$ is the vector of regressors; \mathbf{D}_i is the vector of potentially endogenous explanatory variables; $\mathbf{X}_{exog,i}$ is the vector of exogenous explanatory variables; and $\mathbf{Z}_i = (\mathbf{X}_{exog,i}', \mathbf{Z}_{excl,i}')' \in \mathbb{R}^{d_z}$ is the full vector of instruments, which contains both the exogenous explanatory variables $\mathbf{X}_{exog,i}$ and excluded instruments $\mathbf{Z}_{excl,i}$. The conditional probability in (2.1) comes from Chernozhukov and Hansen's (2005b) identification result in their Theorem 1, which states conditions under which the $\beta_{0\tau}$ satisfying (2.1) is a structural parameter or includes a (conditional) quantile treatment effect parameter.

The conditional probability in (2.1) can be written as a conditional expectation,

$$0 = E[1 - F(Y_i - \beta_{0\tau} - \mathbf{X}_i' \boldsymbol{\beta}_{0\tau}) | \mathbf{Z}_i]. \quad (2.2)$$

By the law of iterated expectations, (2.2) implies the unconditional moments

$$\mathbf{0}_{d_z} = E[\mathbf{Z}_i (1 - F(Y_i - \beta_{0\tau} - \mathbf{X}_i' \boldsymbol{\beta}_{0\tau}))]. \quad (2.3)$$

Like most IVQR estimators, I consider (2.3) to be the IVQR moments. In principle, the \mathbf{Z}_i in (2.3) could be replaced by the (nonparametrically estimated) optimal instruments based on (2.2), but this is beyond my scope. This paper estimates structural parameter $\beta_{0\tau}$ based on the unconditional moments in (2.3).

2.3 Averaging components for IVQR

Generally, the idea of averaging estimation is to average the IVQR estimator with some other estimator that has a smaller variance (and possibly larger bias) than the IVQR estimator. It is essentially a bias–variance tradeoff. The goal is to make the averaging estimator have an overall smaller MSE than the IVQR estimator.

Specifically, I propose using two estimators (or moments) for averaging to improve the IVQR estimator: QR and 2SLS.

2.3.1 QR and endogeneity

The motivation for considering QR for averaging is from the model endogeneity perspective. The averaging between QR and IVQR to improve on IVQR is analogous to averaging between OLS and 2SLS to improve on 2SLS as studied by Hansen (2017b). If the model has little endogeneity, then QR is only a little biased. Meanwhile, QR usually has lower variance than IVQR. Therefore, using QR may result in an averaging estimator that has only a little bias but significantly lower variance than the IVQR estimator. If instead the DGP has severe endogeneity, then QR has much larger bias than IVQR. Ideally, almost all the weight is put on IVQR in this case, whereas less weight is put on IVQR when there is little endogeneity.

If the structural quantile function is also a conditional τ -quantile function (CQF), then

$$P(Y_i \leq \mathbf{X}_i' \beta_{0\tau} \mid \mathbf{X}_i) = \tau. \quad (2.4)$$

Parallel to (2.2), this conditional quantile restriction can be rewritten as a conditional expectation,

$$0 = E[\mathbb{1}_{Y_i \leq \mathbf{X}_i' \beta_{0\tau}} - \tau \mid \mathbf{X}_i]. \quad (2.5)$$

Using the law of iterated expectations like in (2.3), (2.5) implies the unconditional

QR moments in (2.6), which are also the first-order conditions of the population QR minimization problem:

$$0 = E f \mathbf{X}_i [1 f Y_i \quad \mathbf{X}_i^0 \quad 0 \tau] g. \quad (2.6)$$

Computational issues aside, the QR estimator can be considered as the solution to the sample analog of (2.6).

2.3.2 2SLS and heterogeneity

The motivation for using 2SLS is from the heterogeneity perspective. When there is not much heterogeneity across quantiles, 2SLS and IVQR (at any quantile) have similar slope estimands (but not intercepts). Thus, when considering 2SLS as an estimator of the IVQR slopes, it is only slightly biased. Meanwhile, 2SLS usually has smaller variance than IVQR, especially at quantile levels that are away from the median and closer to the tails.² For example, even in an empirically-based simulation with substantial heterogeneity (that causes 2SLS to be biased), Table 3 of Kaplan and Sun (2017b) shows 2SLS to be more efficient than IVQR at four out of five quantile levels.

Incorporating the 2SLS estimator in averaging, the averaging estimator can improve efficiency over the IVQR estimator by putting more weight on 2SLS when it has large enough reduction in variance to outweigh the bias, and putting more weight on IVQR when the bias increase overwhelms the variance reduction.

The averaging weights are crucial for good performance of the averaging estimator. In some cases, too little weight on QR or 2SLS may fail to take advantage of their lower variance, and the averaging estimator's MSE may be the same as that of IVQR. In other cases, too much weight on QR or 2SLS may result in a large squared bias and large MSE, even worse than IVQR. The optimal averaging weight ideally balances the variance and

²One exception to 2SLS having smaller variance is with fat-tailed error terms, which I investigate in the simulations in Section 2.6.

the squared bias, minimizing MSE. This is like the bandwidth choice for kernel regression and other nonparametric estimators. In Section 2.4, I use CLS empirical optimal weight under their averaging GMM framework. In Section 2.5, I propose to obtain the averaging weight through bootstrap in the new bootstrap averaging method.

2.4 Averaging GMM method

This section describes the first averaging estimation method. I apply the CLS averaging GMM framework to IVQR based on the ideas in Section 2.3.

First, I define the estimator. Then, I discuss when this averaging estimator can improve the IVQR estimator, as well as special considerations on smoothing bandwidths used in the estimation of parameters, covariance matrix, and Jacobian matrix in the IVQR model.

2.4.1 Estimator

Define the averaging estimator as

$$\hat{\theta}_{AVG} = (1 - \hat{w}) \hat{\theta}_{IVQR} + \hat{w} \hat{\theta}_{AGG}, \quad (2.7)$$

where I use the empirical averaging weight \hat{w} from CLS (reproduced in my (B.20)); I use de Castro et al.'s (2019) smoothed two-step GMM method to obtain the IVQR estimator $\hat{\theta}_{IVQR}$, which is interpreted as the “conservative estimator” in CLS; and $\hat{\theta}_{AGG}$ is the “aggressive estimator” for which I use de Castro et al.'s (2019) method with the IVQR moments and additional QR or 2SLS moments below, based on Section 2.3.³

I propose two types of additional moments. The additional moment conditions are written $E[\mathbf{g}_i(\theta, \tau)] = \mathbf{0}$. The corresponding QR and 2SLS moment functions are, re-

³Details of the estimator are in Appendix B.2.

spectively,

$$\mathbf{g}_i(\beta, \tau) = \mathbf{D}_i[1/Y_i \quad \mathbf{X}_i^\ell \quad 0/g \quad \tau], \quad (2.8)$$

$$\mathbf{g}_i(\beta, \tau) = (\mathbf{Z}_{-1,i} \quad \mathbf{Z}_{-1})'(Y_i \quad \mathbf{X}_i^\ell), \quad (2.9)$$

where again \mathbf{D}_i are the endogenous regressors (because exogenous regressors are already included in the IVQR moments), $\mathbf{Z}_{-1,i}$ is the full vector of moments excluding the intercept, and \mathbf{Z}_{-1} is their average over $i = 1, \dots, n$. The demeaning in (2.9) is needed because the 2SLS intercept moment condition is not used (i.e., $Y_i \quad \mathbf{X}_i^\ell$ is not mean-zero); (2.9) captures only the zero correlation between instruments and $Y_i \quad \mathbf{X}_i^\ell$.

Combining the IVQR moments together with the QR moments, I obtain the IVQR-QR aggressive estimator $\hat{\beta}_{AGG}$, which combines with the weight \hat{w} in (2.7) to yield the IVQR-QR averaging estimator $\hat{\beta}_{AVG}$. I call this the IVQR-QR type of averaging. Similarly, combining the IVQR moments together with the 2SLS slope moments, I obtain the corresponding IVQR-2SLS aggressive estimator and IVQR-2SLS averaging estimator. I call this the IVQR-2SLS type of averaging.

The ‘‘additional moments’’ might or might not be valid. If the additional moments are valid (i.e., $\mathbf{0} = E[\mathbf{g}(\beta_{0\tau}, \tau)]$), then adding additional valid information to estimation will reduce variance and improve efficiency. If the additional moments are misspecified (i.e., $\mathbf{0} \notin E[\mathbf{g}(\beta_{0\tau}, \tau)]$), then combining these invalid additional moments with the original valid IVQR moments will result in a biased aggressive GMM estimator. However, the misspecified moments could still be helpful, if as a result the aggressive GMM estimator has a large reduction in variance, and an overall reduction in MSE.

Besides the two types of additional moments proposed in this paper, the same idea suggests using IVQR slope moments with other quantile levels (τ) as the additional moments. The intuition is that when there is no or little heterogeneity, there will not be much difference across quantiles. This is related to the L -estimation method in

conventional quantile regression by Koenker and Portnoy (1987).

2.4.2 Improvement on IVQR

In principle, the averaging estimator can always be at least as good as the conservative IVQR estimator, if it puts zero weight on the aggressive estimator when the additional moments are misspecified severely enough. In practice, this desirable result may not exactly hold due to the estimation error of the empirical weight in finite samples.

CLS Theorem 5.2 provides a sufficient condition for the averaging GMM estimator to uniformly dominate the conservative estimator asymptotically. I call it the “uniform dominance condition” (UDC) in this paper. Appendix B.3 provides a detailed discussion of it. “*Uniform dominance*” means the averaging estimator has a (weakly) smaller asymptotic risk than the IVQR estimator, regardless of the true DGP. Therefore, the averaging estimator is unambiguously preferred to the IVQR estimator.

In general, averaging estimation studies like CLS suggest that as long as the additional moments (here QR and 2SLS) can reasonably lower the variance and the model has enough endogenous regressors, then the averaging estimator in this section should uniformly dominate the IVQR estimator asymptotically.

2.4.3 Bandwidth choice

When implementing the CLS averaging GMM framework for IVQR using de Castro et al.’s (2019) smoothed GMM framework, it is important to consider the smoothing bandwidth choice. The CLS empirical weight formula in (B.20) requires estimation of the conservative and aggressive *parameters*, *covariance matrix*, and (population) *Jacobian matrix*, which involves a conditional density that must be nonparametrically estimated. The performance of the averaging GMM method heavily depends on whether the empirical weight is estimated accurately or not.

To estimate the *parameters* and the *covariance matrix* with the smoothed GMM approach of de Castro et al. (2019), I use the smallest possible smoothing bandwidth, for two reasons. First, although Kaplan and Sun (2017b) note that smoothing can reduce MSE,⁴ the goal of my paper is to demonstrate that it is the averaging method (not smoothing) that can improve estimation efficiency. Second, the CLS averaging GMM framework assumes the conservative moments are valid and that the conservative estimator is not biased, but smoothing introduces some bias. Using the smallest possible smoothing bandwidth makes the bias of the conservative IVQR estimator due to smoothing as small as possible.

For IVQR and even QR, the population Jacobian matrix involves a conditional PDF, which is commonly estimated by a nonparametric kernel estimator. The usual kernel estimator is actually the same as the standard sample Jacobian when $\hat{\cdot}$ is based on smoothed moments.

To precisely estimate the *Jacobian matrix* of the IVQR moments and of the aggressive moments, I modify Kato's (2012) bandwidth. He provides the optimal bandwidth for Jacobian estimation with conventional QR based on asymptotic mean squared error (AMSE). He also provides a simplified version assuming independent, standard normal regression errors. I extend his Gaussian plug-in bandwidth to allow for any error variance, and I adapt his formulas for QR to IVQR. With QR, the vector \mathbf{X} acts as both the regressors and the instruments; for IVQR, the parts of the bandwidth formulas where \mathbf{X} acts as instruments are replaced by \mathbf{Z} . For the QR Gaussian plug-in, I build on Kato's (2012) results to prove in Proposition 21 that in the model $Y = \mathbf{X}'\beta_0 + U$ with $U | \mathbf{X} \sim N(\mu, \sigma^2)$ and $Q_\tau(U | \mathbf{X}) = 0$, the sample analog of the AMSE-optimal

⁴Kaplan and Sun (2017b) derive an MSE-optimal bandwidth for estimating the smoothed estimating equation IVQR estimator. This bandwidth is typically much larger than the smallest possible smoothing that makes computation feasible.

bandwidth is

$$\hat{h}_{\text{opt}} = n^{-1/5} \left(\frac{4.5 \sum_{j,k=1}^{d_X} (n^{-1} \sum_{i=1}^n X_{ij}^2 X_{ik}^2)}{\hat{\Delta}(\tau) \sum_{j,k=1}^{d_X} (n^{-1} \sum_{i=1}^n X_{ij} X_{ik})^2} \right)^{1/5},$$

$$\hat{\Delta}(\tau) = \frac{1}{\hat{\sigma}^5} \left[1 - \left(\Phi^{-1}(\tau) \right)^2 \right]^2 \phi \left(\Phi^{-1}(\tau) \right),$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the standard normal CDF and PDF. Details are in Appendix B.1.

2.5 Bootstrap model averaging method

The second averaging method I propose is a new bootstrap model averaging method.

This method averages the three estimators IVQR, 2SLS, and QR, using the optimal weight from the bootstrap world. That is, first, in the bootstrap world, I compute many possible weighted averages of IVQR, QR, and 2SLS using a grid of fixed weights. Then, I pick the optimal weight choice that minimizes the RMSE of the averaging estimator in the bootstrap world; I call this the bootstrap optimal weight. Finally, I average the real-world IVQR, QR, and 2SLS estimators using the bootstrap optimal weight, and I call the resulting averaging estimator the bootstrap averaging estimator.

The bootstrap averaging estimator comes from the same motivation as in Section 2.3. That is, compared to the IVQR estimator, the QR and 2SLS estimators usually have smaller variance, and as long as their bias is not too large, averaging with them can decrease MSE overall. This is especially true with only mild endogeneity or heterogeneity.

The bootstrap method algorithm is as follows.

1. Compute the IVQR, 2SLS, and QR estimators $(\hat{\beta}_{\text{IVQR}}, \hat{\beta}_{\text{2SLS}}, \hat{\beta}_{\text{QR}})$ using the original sample $(Y_i, \mathbf{X}_{\text{endo},i}, \mathbf{X}_{\text{exog},i}, \mathbf{Z}_{\text{excl},i})$ for $i = 1, \dots, n$.
2. Draw bootstrap sample b : $(Y_i^{(b)}, \mathbf{X}_{\text{endo},i}^{(b)}, \mathbf{X}_{\text{exog},i}^{(b)}, \mathbf{Z}_{\text{excl},i}^{(b)})$ for $i = 1, \dots, n$.
3. Use the bootstrap sample to compute the bootstrapped IVQR, 2SLS, and QR estimators $(\hat{\beta}_{\text{IVQR}}^{(b)}, \hat{\beta}_{\text{2SLS}}^{(b)}, \hat{\beta}_{\text{QR}}^{(b)})$.

4. Compute averaging estimators in the bootstrap sample for a fixed grid of weights (w_{1s}, w_{2s}, w_{3s}) for $s = 1, \dots, S = 13,071$, with $0 \leq w_{1s}, w_{2s}, w_{3s} \leq 1$ and $w_{1s} + w_{2s} + w_{3s} = 1$; specifically, $\hat{\text{AVG}}_{s}^{(b)} = w_{1s} \hat{\text{IVQR}}^{(b)} + w_{2s} \hat{\text{2SLS}}^{(b)} + w_{3s} \hat{\text{QR}}^{(b)}$ ⁵
5. Repeat steps 2.-4. for $b = 1, \dots, B$.
6. Treating $\hat{\text{IVQR}}$ as the true population parameter in the bootstrap world, compute the RMSE of the 13,071 different averaging estimators:

$$\text{RMSE}(\hat{\text{AVG}}_{s}) = \sqrt{\frac{1}{B} \sum_{b=1}^B (\hat{\text{AVG}}_{s}^{(b)} - \hat{\text{IVQR}})^2}.$$

7. Define the “bootstrapped optimal weight” as minimizing the RMSE in step 6.:

$$(w_1, w_2, w_3) = (w_{1s^*}, w_{2s^*}, w_{3s^*}), \quad s^* = \underset{s \in \{1, \dots, 13071\}}{\text{arg min}} \text{RMSE}(\hat{\text{AVG}}_{s}).$$

8. Using the bootstrapped optimal weight in step 7., define the “bootstrapped averaging estimator” as $\hat{\text{AVG}}_{\text{BS}} = w_1 \hat{\text{IVQR}} + w_2 \hat{\text{2SLS}} + w_3 \hat{\text{QR}}$.

The computation time for the bootstrap averaging estimator is reasonably fast even with a large weight grid size like 13,071. The actual IVQR, 2SLS, and QR estimators only need to be computed once per bootstrap draw; the 13,071 is just arithmetic. In

⁵The 13,071 sets of fixed weights come from the following procedure.

To start, let w_1 increase from 0 to 0.99 by a grid of 0.01. For each fixed w_1 , let w_2 and w_3 share the rest $(1 - w_1)$ with 101 combinations, i.e., $w_2 = 0(1 - w_1), w_2 = 0.01(1 - w_1), \dots, w_2 = 0.99(1 - w_1), w_2 = 1(1 - w_1)$ and $w_3 = 1 - w_1 - w_2$. This produces $100 \cdot 101 = 10,100$ combinations.

The end point case $w_1 = 1, w_2 = 0, w_3 = 0$ is one more combination.

The finer the weights, the closer the weights picked to the optimal weights in the bootstrap world, and the better performance of the bootstrap averaging estimator. But a finer grid of weights also results in more time for computation. I use additional finer weight combinations, but with the number of weight sets as small as possible, trying to improve the bootstrap performance but without sacrificing the computation time too much. I consider the special case of two estimators averaging with finer weights. That is, either $w_2 = 0$ the IVQR-QR averaging or $w_3 = 0$ the IVQR-2SLS averaging. Meanwhile, if the optimal weights are chosen as the $w_2 = 0$ combinations under the availability of the previous 10,000 combinations with nonzero w_2 , that means the w_1 is also relatively large. So I consider $w_1 = 0.85$ to 1 increasing by a grid of 0.0001, i.e., $w_1 = 0.8501, \dots, 0.9999, w_2 = 0, w_3 = 1 - w_1$, which has $99 - 15 = 1485$ combinations, and $w_1 = 0.8501, \dots, 0.9999, w_3 = 0, w_2 = 1 - w_1$, which has another $99 - 15 = 1485$ combinations.

In total, there are $10,100 + 1 + 1485 + 1485 = 13,071$ combinations of fixed weights.

my implementation, $\hat{\beta}_{IVQR}$ is computed using de Castro et al.'s (2019) smoothed two-step GMM method and $\hat{\beta}_{IVQR}^{(b)}$ are obtained by first projecting the regressors onto the instruments, and then solving smoothed versions of the exactly-identified equations, using standard numerical methods. So the additional computation time of bootstrap averaging over GMM averaging is not too big.

There are some differences with Section 2.4 that may enable the bootstrap's better performance in simulations. Section 2.4 considers averaging between two estimators (conservative and aggressive), whereas the bootstrap method averages among three estimators: IVQR, 2SLS, and QR. Computationally, bootstrap averaging is simpler and easier, not requiring two-step GMM with a large degree of overidentification.

Investigating a theoretical-based optimal weight formula is left for future research, since deriving an optimal vector weight for three-estimators-averaging is not easy. To the best of my knowledge, the current averaging studies are two-estimators-averaging, which enables a scalar optimal weight, through substantial work. In contrast, the bootstrap method in my paper selects the optimal weights from a grid of fixed weights, which enables the fast computation. It performs very well in simulations in extensive DGPs, even without a theory-based optimal weight formula.

2.6 Simulations

This section reports simulation results showing the finite-sample performance of two types (IVQR-QR and IVQR-2SLS) of GMM averaging estimator, and the bootstrap averaging estimator, relative to the IVQR estimator.

I consider three different simulation models, with many DGPs within each model. Simulation models 1 and 2 present cases where all the averaging estimators uniformly dominate the IVQR estimator. Simulation model 1 closely follows the CLS simulation model 1, but with modification to IVQR. Simulation model 2 includes slope hetero-

geneity, since one important feature of quantile regression is to capture unobserved heterogeneity. Simulation model 3 presents a case where the averaging estimators do not uniformly dominate the IVQR estimator because there is only a single endogenous regressor.⁶

The averaging estimators' performance is measured by their "robust" root mean squared error (rRMSE) relative to that of the IVQR estimator. The rRMSE is computed similar to RMSE but replacing bias with median bias and replacing standard deviation with interquartile range (IQR) divided by 1.349. The rRMSE equals RMSE exactly for normal sampling distributions but is more robust to outliers.⁷ More specifically, with d_θ parameters, rRMSE is computed as

$$\text{rRMSE} = \sqrt{\sum_{j=1}^{d_\theta} [(\text{median bias}_j)^2 + (\text{IQR}_j/1.349)^2]}, \quad (2.10)$$

where the median bias is the median of estimators among the M simulation replications minus the true parameter value,

$$\text{median bias}_j = \text{median}_{m2f1, \dots, Mg}(\hat{\theta}_{jm}) - \theta_j, \text{ for } j = 1, \dots, d_\theta, \quad (2.11)$$

the IQR is the difference between the 0.75-quantile and 0.25-quantile of the estimators among the M replications,

$$\text{IQR}_j = 0.75\text{-quantile}_{m2f1, \dots, Mg}(\hat{\theta}_{jm}) - 0.25\text{-quantile}_{m2f1, \dots, Mg}(\hat{\theta}_{jm}), \text{ for } j = 1, \dots, d_\theta. \quad (2.12)$$

For each simulation DGP, unless otherwise noted, I run 200 simulation replications, with sample size $n = 1000$ and 50 bootstrap replications. Averaging would work even better with more bootstrap replications (or larger n), but the time to run simulations for

⁶Similarly, CLS simulation model 2 shows the averaging GMM estimator is not uniformly better than the conservative GMM estimator in a linear IV model with a single endogenous regressor.

⁷Using either rRMSE or RMSE does not affect the results qualitatively. The rRMSE and RMSE only have minor quantitative difference, and are even closer when considering the relative ratio to that of the IVQR estimator.

my many DGPs would become prohibitive.⁸ In practice, with only one dataset instead of thousands, using more bootstrap replications would not take so much additional time. Thus, the bootstrap averaging simulation results are “conservative” because performance should be even better in practice.

I normalize the rRMSE of the IVQR estimator to 1 and use the relative rRMSE to see how other estimators perform relative to the IVQR estimator. That is, I divide all rRMSEs by the IVQR rRMSE to get the relative rRMSE. If the relative rRMSE is above (below) 1, that means the estimator has larger (smaller) rRMSE than IVQR, i.e., it performs worse (better) than IVQR. Ideally, an estimator has relative rRMSE (weakly) below 1 in all different DGPs. In this case, we say the proposed estimator uniformly dominates the IVQR estimator: regardless of the true DGP, it performs as good or better than IVQR. Such an estimator is unambiguously preferred to IVQR. An estimator may still be preferred even without uniform dominance, but it would depend on the user’s preferences (e.g., the estimator’s Bayes risk may be larger or smaller than IVQR’s depending on the user’s prior over DGPs).

2.6.1 Simulation model 1

This simulation model is close to simulation model 1 in CLS. The CLS model considers endogenous regressors at a fixed endogeneity level with a set of valid instruments, while introducing potentially invalid instruments in the additional moments with varying degrees of endogeneity. Instead of varying the endogeneity of additional instruments, I let

⁸If using 200 bootstrap replications, it will take a few months to run all the simulations, since I include extensive DGPs. For example, simulation model 3 has 242 DGPs per quantile level τ . To run simulations at quantile $\tau = 0.2$ to $\tau = 0.8$, it will run $242 \times 7 = 1694$ times. The bootstrap estimator already performs very well with only 50 bootstrap replications.

As n increases, the averaging estimators will tend to perform better, in part because the large-sample theory in CLS will apply with less approximation error. There is no exact finite-sample theory about the averaging estimator. For each specific n , we could run through the simulations to see the performance of the averaging estimator. For the sake of computation time and space, results at multiple n are not included in this paper.

the regressors' endogeneity level vary across DGPs. Otherwise, the DGPs are the same as in CLS. This simulation model 1 has enough endogenous regressors and instruments that uniform dominance holds.

Simulation DGP

Consider the linear model

$$Y = \theta_0 + \sum_{j=1}^6 \theta_j X_j + u, \quad (2.13)$$

where $\theta_1 = \theta_2 = \dots = \theta_6 = 2.5$ and $\theta_0 = 1$; Y is the dependent variable; and there are six endogenous regressors (X_1, \dots, X_6) and twelve excluded valid instruments (Z_1, \dots, Z_{12}). The data $(Y_i, X_{1,i}, \dots, X_{6,i}, Z_{1,i}, \dots, Z_{12,i})$ are sampled iid for $i = 1, \dots, n$. The regressors are generated by

$$X_j = (Z_j + Z_{j+6})/2 + \epsilon_j, \text{ for } j = 1, \dots, 6. \quad (2.14)$$

The $(Z_1, \dots, Z_{12}, \epsilon_1, \dots, \epsilon_6, u)$ are generated from a multivariate normal distribution with mean zero and covariance matrix $\text{diag}(\mathbf{I}_{12 \times 12}, \underline{\Sigma}_{7 \times 7})$, where

$$\underline{\Sigma}_{7 \times 7} = \begin{bmatrix} \mathbf{I}_{6 \times 6} & c_0 \mathbf{1}_{6 \times 1} \\ c_0 \mathbf{1}_{1 \times 6} & 1 \end{bmatrix}. \quad (2.15)$$

The $\mathbf{I}_{6 \times 6}$ and $\mathbf{I}_{12 \times 12}$ are the identity matrix. The $\mathbf{1}_{6 \times 1}$ and $\mathbf{1}_{1 \times 6}$ are the column and row vector of ones, respectively.

In each DGP, c_0 is a fixed constant that measures the amount of endogeneity, specifically the covariance between regressors and the structural error term. For $\underline{\Sigma}_{7 \times 7}$ to be positive definite it requires $c_0 < 0.4$. Thus, let c_0 take value in $\{0, 0.05, 0.1, \dots, 0.4\}$. The higher the c_0 value, the more endogenous the regressors. In the special case $c_0 = 0$, the regressors are generated independently from the structural error. The simulation results demonstrate that in the case $c_0 = 0.4$, the regressors are endogenous enough that the QR estimator performs much worse than IVQR estimator.

In this data-generating process, the error term u follows a Gaussian distribution. Then, I make a location shift to make its τ -quantile to be zero. This shift is to make the model to satisfy the conditional IVQR restriction that conditional on instruments, the τ -quantile of the error term is zero.

The conservative IVQR moment functions use the twelve instruments and a constant,

$$g_{1,j}(\theta) = Z_j(1 \text{ } fY \text{ } \theta_0 \text{ } X_1\theta_1 \text{ } X_6\theta_6 \text{ } 0g \text{ } \tau) \text{ for } j = 0, \dots, 12, \quad (2.16)$$

where $Z_0 = 1$ is the constant term.

The additional moment functions for IVQR-QR averaging are

$$g_j(\theta) = X_j(1 \text{ } fY \text{ } \theta_0 \text{ } X_1\theta_1 \text{ } X_6\theta_6 \text{ } 0g \text{ } \tau) \text{ for } j = 1, \dots, 6. \quad (2.17)$$

The additional moment functions for IVQR-2SLS averaging are

$$g_j(\theta) = (Z_j \text{ } Z_j)(Y \text{ } \theta_0 \text{ } X_1\theta_1 \text{ } X_6\theta_6) \text{ for } j = 1, \dots, 12. \quad (2.18)$$

The IVQR estimator is the GMM estimator from the 13 moments using (2.16). The IVQR-QR aggressive estimator is the GMM estimator from the 19 moments using (2.16) and (2.17). The IVQR-2SLS aggressive estimator is the GMM estimator from the 25 moments using (2.16) and (2.18).

Simulation results

In simulation model 1, all three averaging estimators uniformly dominate the IVQR estimator. Figure 2.1 shows these three averaging estimators' relative rRMSE is bounded below 1 in a class of DGPs with different endogeneity level. Table 2.1 reports the rRMSE of various estimators relative to that of the IVQR estimator. The second column reports the value of c_0 , which indicates how much regressor endogeneity there is in the DGP. Columns 3, 6, and 9 are the three averaging estimators proposed in this paper. Columns 4 and 5 are the IVQR-2SLS aggressive estimator and the 2SLS estimator for reference.

Table 2.1: Relative rRMSE in simulation model 1, $\tau = 0.5$.

DGP	Endog	IVQR.2SLS			IVQR.QR			BS	SEE
		AVG	AGG	2SLS	AVG	AGG	QR		
1	0	0.92	0.91	0.80	0.65	0.47	0.44	0.58	0.84
2	0.05	0.92	0.93	0.73	0.67	0.55	0.53	0.60	0.82
3	0.10	0.90	0.90	0.83	0.79	0.87	0.84	0.69	0.83
4	0.15	0.87	0.90	0.82	0.87	1.14	1.13	0.79	0.84
5	0.20	0.92	0.86	0.78	0.93	1.47	1.46	0.84	0.84
6	0.25	0.91	0.90	0.78	0.95	1.72	1.71	0.83	0.83
7	0.30	0.92	0.96	0.79	0.96	2.18	2.16	0.88	0.85
8	0.35	0.88	0.85	0.73	0.96	2.31	2.27	0.83	0.79
9	0.40	0.88	0.91	0.80	0.98	2.90	2.84	0.83	0.85

200 replications. 50 bootstraps. Sample size is 1000.

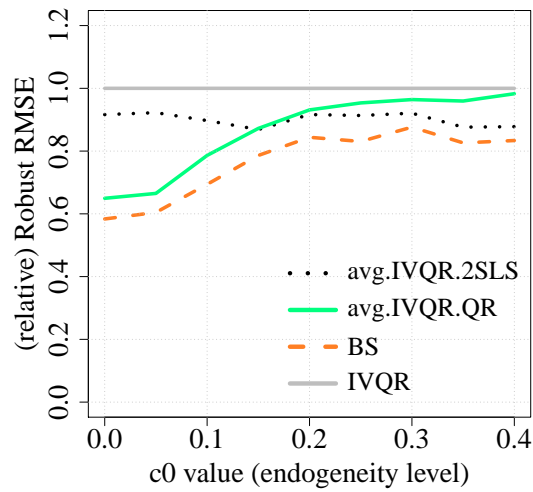


Figure 2.1: Relative rRMSE in simulation model 1; $\tau = 0.5$, $n = 1000$, 200 replications, 50 bootstraps.

Columns 7 and 8 are the IVQR-QR aggressive estimator and the QR estimator for reference. Column 10 is the smoothed estimating equations (SEE) estimator proposed by Kaplan and Sun (2017b), using their code’s plug-in bandwidth.

In simulation model 1, with fixed coefficients (no slope heterogeneity) and Gaussian errors, naturally IVQR-2SLS averaging performs well across all DGPs. The IVQR-2SLS aggressive estimator always has rRMSE below 1, and the IVQR-2SLS averaging takes advantage of this, performing similarly well.

The IVQR-QR averaging also performs as desired with rRMSE always at least as good as IVQR, taking advantage of the QR moments to reduce rRMSE when endogeneity is low, but shifting weight to IVQR when the endogeneity and corresponding QR bias become large. In DGPs 1–3 with the least endogeneity, the QR and aggressive IVQR-QR estimators perform better than the IVQR estimator. In DGPs 4–9 with more endogeneity, they are increasingly worse than IVQR. However, the IVQR-QR averaging estimator has relative rRMSE less than 1 in all DGPs, even in the DGPs where the QR moments are very misspecified and the QR and IVQR-QR aggressive estimators are much worse than IVQR. This shows the empirical optimal weight formula works as desired, putting more weight on the aggressive estimator when the additional moments reduce variance more relative to the increased bias, and putting more weight on the conservative estimator (IVQR) when the additional moments are severely misspecified.

The bootstrap averaging estimator performs best of all. It not only uniformly dominates the IVQR estimator, with up to 42% efficiency gain, but also is uniformly better than either type of GMM averaging estimator, as seen readily in Figure 2.1. Moreover, the bootstrap averaging estimator usually performs better than the SEE estimator, although not uniformly.

The same patterns are seen across $\tau = 0.2, \dots, 0.8$ in the results in Appendix B.5.1: all three averaging estimators uniformly dominate IVQR, and bootstrap averaging uni-

formly dominates both GMM averaging estimators.

Results with non-Gaussian error term

The uniform dominance in this fixed-coefficient model with multiple endogenous regressors also holds with a non-Gaussian error term. To illustrate this, I set the error term to follow a chi-square distribution with 4 degrees of freedom. More specifically, I first generate the error term in the same way as in (2.15). Second, I transform the error term to $u = F_{\chi_4^2}^{-1}(F(u))$ where $F(\cdot)$ is the CDF of the original Gaussian error term u and $F_{\chi_4^2}^{-1}$ is the inverse CDF of a χ_4^2 distribution. Finally, I shift u to have its τ -quantile equal zero.

The simulation results have similar patterns as in the Gaussian error case; see the results tables and figures in Appendix B.5.1. All three averaging estimators have relative rRMSE less than 1 in all the DGPs at all quantiles, except three cases of IVQR-2SLS averaging at $\tau = 0.2$ having relative rRMSE between 1.01 and 1.03. However, these three rare cases seem due to simulation error since their relative rRMSEs are less than or equal to 1 when running more simulation replications. Bootstrap averaging again uniformly dominates both IVQR-2SLS and IVQR-QR averaging across $\tau = 0.4$ to $\tau = 0.8$. At $\tau = 0.2$ and $\tau = 0.3$, for which 2SLS is not very helpful, bootstrap averaging and IVQR-QR are similarly good. At $\tau = 0.8$, for which 2SLS is particularly helpful, the bootstrap averaging relative rRMSE is always below 0.75 and as low as 0.52.

2.6.2 Simulation model 2

Seeing slope heterogeneity across quantiles is part of the value of quantile regression. Simulation model 2 extends simulation model 1 to allow for slope heterogeneity.

Simulation DGP

Consider a linear random-coefficient model

$$Y = \theta_0 + \sum_{j=1}^6 \theta_j X_j + u, \quad (2.19)$$

with $\theta_0 = 1$ and random coefficients for individual i equal to

$$\theta_{1i} = \theta_{2i} = \dots = \theta_{6i} = c_h [F(u_i)]^4, \text{ for } i = 1, \dots, n, \quad (2.20)$$

where $F(\cdot)$ is the CDF of u . The slope term (2.20) is set to be a function of the rank of the error term in its distribution. It represents the slope heterogeneity feature. The term c_h is a fixed constant in a DGP, the same as c_2 in simulation model 3. When $c_h = 0$, the slope is a constant zero, which means the model has no slope heterogeneity. The larger the c_h value, the more slope heterogeneity in the DGP. In simulation model 2, c_h takes value in $\{0, 0.1, \dots, 1\}$. I consider both Gaussian and non-Gaussian errors. The non-Gaussian error follows the same transformation as in Section 2.6.1. The error term has the same location shift as in Section 2.6.1 to make its τ -quantile equal zero.

As in simulation model 1, there are six endogenous regressors (X_1, \dots, X_6) and twelve valid instruments (Z_1, \dots, Z_{12}), and sampling is iid. The regressors are generated by

$$X_j = Z_j + Z_{j+6} + \epsilon_j, \text{ for } j = 1, \dots, 6. \quad (2.21)$$

The monotonicity condition in IVQR requires that the structural quantile function is increasing in u given any $\mathbf{X} = \mathbf{x} \geq \mathbf{X}$ (Chernozhukov and Hansen, 2005b, Condition A1). For the monotonicity condition to hold in this simulation model, I restrict the regressors to have non-negative support. More specifically, I shift each regressor to the right by 3.1 times its standard deviation. There is less than 0.001 probability that a regressor remains negative, in which case I set it to zero.

The $(Z_1, \dots, Z_{12}, \epsilon_1, \dots, \epsilon_6, u)$ are generated the same way as in simulation model

1, following a multivariate normal distribution with mean zero and covariance matrix $\text{diag}(\mathbf{I}_{12}, \Sigma_{7,7})$, $\Sigma_{7,7}$ defined in (2.15). Since $c_0 \in \{0, 0.05, 0.1, \dots, 0.4\}$, there are 11 \times 9 DGP combinations of (c_h, c_0) .

I simulate the rRMSE as in (2.10), where the true population parameters are $\theta_j(\tau) = c_h \tau^4$ for all $j = 1, \dots, 6$.

Simulation results

Table 2.2: Bounds of relative rRMSE in simulation model 2, $\tau = 0.5$.

	Endogeneity Level					
	No ($c_0 = 0$)		Some ($c_0 = 0.2$)		Much ($c_0 = 0.4$)	
	Lower	Upper	Lower	Upper	Lower	Upper
$\hat{\theta}_{\text{AVG.2SLS}}$	0.922	1.006	0.960	1.024	0.959	0.997
$\hat{\theta}_{\text{AVG.QR}}$	0.831	0.919	0.997	1.002	0.998	1.001
$\hat{\theta}_{\text{BS}}$	0.770	0.929	0.850	1.014	0.860	1.006
	Heterogeneity Level					
	No ($c_h = 0$)		Some ($c_h = 0.5$)		Much ($c_h = 1$)	
	Lower	Upper	Lower	Upper	Lower	Upper
$\hat{\theta}_{\text{AVG.2SLS}}$	0.906	0.991	0.973	1.022	0.981	1.016
$\hat{\theta}_{\text{AVG.QR}}$	0.857	1.001	0.876	1.001	0.888	1.001
$\hat{\theta}_{\text{BS}}$	0.803	0.883	0.877	1.004	0.886	1.018

200 replications. 50 bootstraps. Sample size is 1000.

In Table 2.2, I fix the endogeneity level indicator c_0 at zero, 0.2, or 0.4 (highest value), and vary the slope heterogeneity across all values $c_h = 0, 0.1, \dots, 1$. Similarly, I fix c_h at zero, 0.5, or 1 (highest value), and vary the endogeneity level across all values $c_0 = 0, 0.05, \dots, 0.4$. It covers around half of the 99 possible combinations of (c_h, c_0) , including all the boundary/extreme values. Table 2.2 reports the upper and lower bounds of the relative rRMSE of the three averaging estimators in these six cases (i.e., 3 fixed endogeneity levels plus 3 fixed heterogeneity levels).

Table 2.2 summarizes results at the median in these six cases. More detailed results tables are included in the supplemental appendix. We can see that all three averaging estimators have relative rRMSE lower bounds strictly less than 1. The bootstrap averaging estimator has the smallest lower bound in five cases and is essentially tied for smallest (up to simulation error) in the sixth. With no endogeneity and varying heterogeneity, IVQR-QR and bootstrap averaging both have upper bound strictly below 1, while the IVQR-2SLS averaging estimator has upper bound roughly 1 (up to simulation error). With no heterogeneity and varying endogeneity, IVQR-2SLS and bootstrap averaging both have upper bound strictly below 1, while IVQR-QR averaging has upper bound roughly 1. In the other four cases, the three averaging estimators all have upper bound roughly 1. Although some exceed 1 (largest value 1.024), this is believed to be entirely due to simulation error: all the upper bounds reduce to 1.000 ± 0.001 with a larger number of simulation replications.

Figure 2.2 is similar to Table 2.2, but it visualizes results for all 60 DGPs (not just bounds). It shows that in simulation model 2, all three averaging estimators uniformly dominate the IVQR estimator. Using any of these three averaging estimators would provide more efficient estimation than the IVQR estimator. Further, bootstrap averaging is always at least as good as the others (up to simulation error) and is significantly better in some cases.

Results at other quantiles

Appendix B.5.2 includes results like Figure 2.2 and Table 2.2 at quantiles ranging from $\tau = 0.2$ to $\tau = 0.8$. The supplemental appendix provides yet more detailed tables for these cases.

For some quantile levels, for some DGPs, the bootstrap averaging estimator's relative rRMSE is above 1, but always close enough that it may be explained by the low number of simulation replications (200) and bootstrap replications (50). Among 420 values

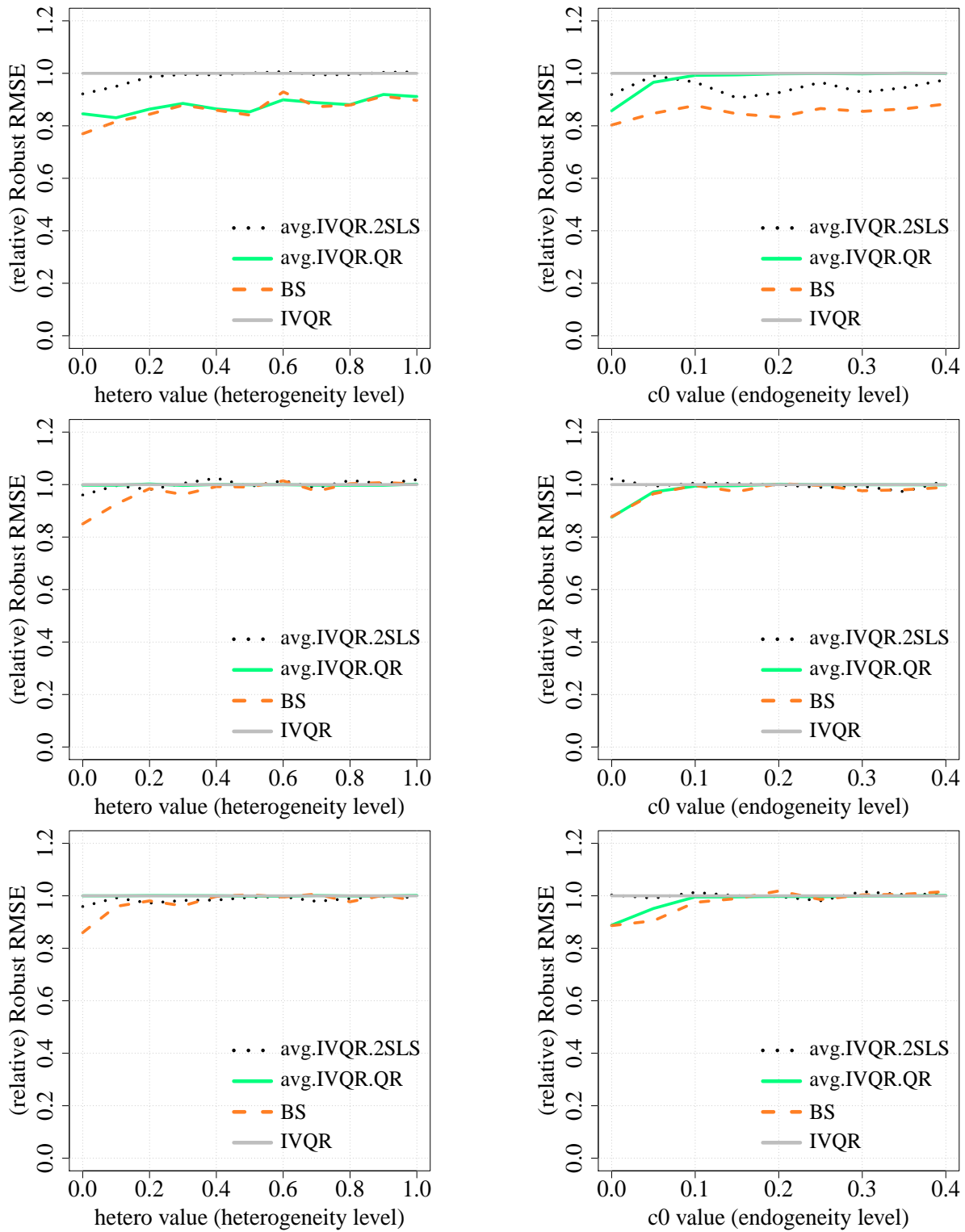


Figure 2.2: Relative rRMSE in simulation model 2 at $\tau = 0.5$ median level in 6 cases: fixed endogeneity level $c_0 = 0$ (left top), $c_0 = 0.2$ (left middle), $c_0 = 0.4$ (left bottom) and varying heterogeneity; and fixed heterogeneity level $c_h = 0$ (right top), $c_h = 0.5$ (right middle), $c_h = 1$ (right bottom) and varying endogeneity, based on 200 replications and 50 bootstraps. Sample size $n = 1000$.

(i.e., 60 DGPs with 7 values of τ each), the largest relative rRMSE is 1.075. However, increasing the bootstrap replications to 1000 reduces this to merely 1.014. This 1.014 may even reduce further with more simulation replications, but even if not, a worst-case rRMSE only 1.4% worse than IVQR is a small price to pay for the large efficiency gains seen across many other DGPs. With the usual 50 bootstrap replications, the relative rRMSE exceeds 1.05 only 19 times out of 420, and it is between 1 and 1.04 in 96 out of 420 DGPs. In the other 305 out of 420 cases, the bootstrap averaging estimator has relative rRMSE less than 1. At every quantile, the bootstrap averaging estimator's relative rRMSE can reach as low as 0.79, and at some quantiles even 0.574. These values should also improve with a higher number of bootstrap replications.

The QR estimator and the IVQR-QR averaging estimator each have similar patterns across quantiles in the 60 DGPs shown in the table in the supplemental appendix. One finding is that in the much-endogeneity case ($c_0 = 0.35$), the IVQR-QR aggressive estimator has a computation problem and performs very poorly. The IVQR-QR averaging estimator, however, still has relative rRMSE of 1. This indicates the empirical weight puts almost all weight on the conservative IVQR estimator and zero weight on the IVQR-QR aggressive estimator. The averaging method works in a desirable way. The IVQR-QR averaging estimators all show uniform dominance over the IVQR estimator.

For IVQR-2SLS averaging, there is a “lucky quantile” at which 2SLS performs well across all DGPs, regardless of heterogeneity. That is because there is a certain quantile level τ at which the IVQR slope coincidentally equals the 2SLS slope. As τ gets farther from the “lucky quantile,” the IVQR slope increasingly differs from the 2SLS slope, with the rate of increase depending on c_h .

The simulation results show that at $\tau = 0.6, 0.7, 0.8$, the 2SLS estimator has relative rRMSE less than 1 uniformly in all 60 DGPs, even in the DGPs with high c_h values. Correspondingly, the bootstrap averaging estimator is much lower than 1 at these three

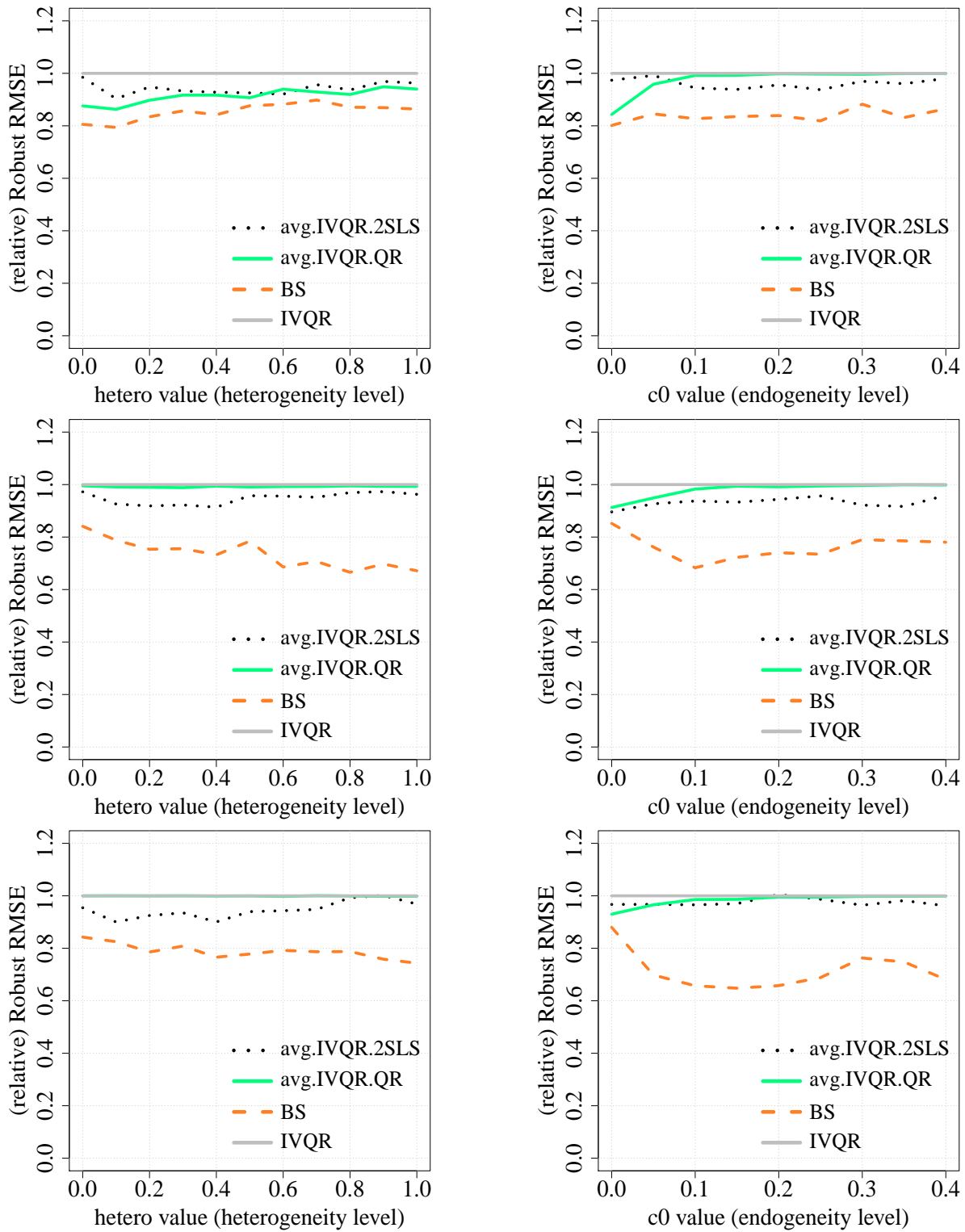


Figure 2.3: Relative rRMSE in simulation model 2 at $\tau = 0.7$ quantile level in 6 cases: fixed endogeneity level $c_0 = 0$ (left top), $c_0 = 0.2$ (left middle), $c_0 = 0.4$ (left bottom) and varying heterogeneity; and fixed heterogeneity level $c_h = 0$ (right top), $c_h = 0.5$ (right middle), $c_h = 1$ (right bottom) and varying endogeneity, based on 200 replications and 50 bootstraps. Sample size $n = 1000$.

quantiles. Figure 2.3 shows the performance of the three averaging estimators at $\tau = 0.7$. The bootstrap averaging estimator's relative rRMSE is significantly below 1, and much below that of the IVQR-2SLS averaging estimator. This is probably because the bootstrap method averages the 2SLS estimator directly, rather than through aggressive GMM that may not weight the 2SLS slope moments as heavily as is optimal.

2.6.3 Simulation model 3

Simulation model 3 acknowledges the limitation of the averaging method, especially in small models with a single endogenous regressor, which is the same finding in other averaging studies. I consider a model based on the Job Training Partnership Act that includes 242 DGPs with combinations of different endogeneity, heterogeneity, and fat-tail levels. The simulation results suggest the three averaging estimators are not uniformly better than the IVQR estimator in this model. The averaging estimators have smaller rRMSE than IVQR estimator in some cases, but not all. More details on this simulation model and results are discussed in Appendix B.4.

2.7 Conclusion

This paper contributes two averaging estimation methods to improve finite-sample efficiency in IVQR estimation. First, I implement the averaging GMM of Cheng, Liao, and Shi (2019), proposing two types of additional moments based respectively on conventional QR and 2SLS, and considering other important implementation details such as bandwidths, extending a result of Kato (2012) along the way. Second, I propose a new method that uses the bootstrap to estimate optimal weights for averaging IVQR, QR, and 2SLS estimators.

This paper provides simulation evidence that these three averaging estimators outperform the IVQR estimator across all kinds of DGPs in large models with multiple

endogenous regressors, as well as across quantile levels from $\tau = 0.2$ to $\tau = 0.8$. Bootstrap averaging offers especially substantial efficiency gains, even with only 50 bootstrap draws.

Future work could involve developing theory for the practically successful bootstrap method, or investigating averaging across quantiles motivated by Koenker and Portnoy (1987), or adding non-trivial smoothing into this averaging framework. Additionally, an IVQR application of Armstrong and Kolesár (2021) may better use additional-but-possibly-misspecified moments when the object of interest is a scalar instead of the full parameter vector.

Chapter 3

Panel quantile regression with time-invariant rank

3.1 Introduction

Panel data models are popular in applied research because they can address endogeneity in structural models without using instruments. They achieve identification of causal effects by controlling for unobserved time-invariant heterogeneity, since we observe multiple time periods for each individual. The term “individual” may broadly refer to a person, country, firm, school, etc.

Much research is interested in not only the causal effect of variables on the mean outcome, but also the effect on distributional features, such as (conditional) quantiles. Quantile regression is appealing since it can account for unobserved heterogeneity and capture heterogeneous effects.

Recently, there have been a growing number of studies on panel quantile regression. Such studies are challenging because the conditional quantile operator is nonlinear. In the standard fixed effects model, the unobserved individual effect U usually correlates with the regressor vector \mathbf{X} , so pooled OLS results in biased estimates of the structural

parameter vector β . This β is identified by first-differencing or demeaning to get rid of the unobserved individual effect. However, we cannot first-difference or demean in quantile models due to the nonlinearity (i.e., the difference of quantiles is not the quantile of the difference).

I propose a structural random coefficient model that addresses endogeneity with panel data and captures heterogeneous causal effects as in quantile models. Quantile models can be interpreted as random coefficient models in a cross-sectional setting (Koenker, 2005, Section 2.6). The linear quantile regression model

$$Q_\tau(Y | \mathbf{X}) = \mathbf{X}^\top \beta(\tau) \quad (3.1)$$

is generated by the structural random coefficient model

$$Y = \mathbf{X}^\top \beta(U) \quad (3.2)$$

with $U \sim \text{Unif}(0, 1)$ given the exogeneity assumption $U \perp \mathbf{X}$ and the monotonicity condition that $\beta(u)$ is increasing in u . With panel data, I can relax the independence restriction $U \perp \mathbf{X}$ to allow unrestricted dependence between U and \mathbf{X} .

To represent the unobserved heterogeneity, I use the “rank variable” idea from Chernozhukov and Hansen’s (2005b) instrumental variable quantile regression model. The rank variable determines the individual’s ranking (quantile) in the outcome distribution if hypothetically every individual were made to have the same explanatory variable values, like U in (3.2). The standard fixed effects panel models use an additively separable individual characteristic to represent the unobserved time-invariant heterogeneity. Instead, I allow the unobserved time-invariant individual characteristic to enter nonseparably as the rank variable, similar to (3.2):

$$Y_{it} = \mathbf{X}_{it}^\top \beta(U_i) + V_{it}, \quad (3.3)$$

where the additional idiosyncratic error V_{it} is assumed exogenous. Different from many studies that assume the individual rank variable is iid across time, the time-invariant rank variable in this paper makes more economic intuition in many cases.

My structural model includes the standard fixed effects (FE) model as a special case. If only the intercept (not slope coefficients) depends on the rank variable U_i , then $(U_i)^\theta = (\beta_0(U_i), \beta_1, \beta_2, \dots)$ (α_i, θ_1) , and writing $\mathbf{X}^\theta = (1, \mathbf{X}_1)$, then (3.3) becomes the traditional linear structural FE model

$$Y_{it} = \mathbf{X}_{1,it}^\theta + \alpha_i + V_{it}. \quad (3.4)$$

In (3.4), the structural effect of \mathbf{X} on Y is homogeneous across individuals. In contrast, my model in (3.3) allows a heterogeneous (individual-specific) structural effect of \mathbf{X} on Y , through the dependence of (U_i) on U_i .

My model can be applied to a wide range of economic settings. For example, consider estimating a Cobb–Douglas production function given a panel dataset of farm i over time periods t (seasons or years). The productivity is heterogeneous across farms, due to soil quality, labor skill, machine productivity, etc. If we consider the general productivity of each farm as its latent rank variable, we could write the generalized Cobb–Douglas production function as in (3.3), where Y_{it} is the log output of farm i at time t ; \mathbf{X}_{it} includes an intercept and the log of the inputs, such as labor and capital; U_i denotes farm i 's general productivity; and the idiosyncratic disturbance V_{it} denotes some external shock outside the farm's control, such as rainfall. This model is more general than the traditional Cobb–Douglas production function analysis, which assumes the productivity of labor and capital are identical across farms.

I make four primary contributions. First, I contribute a new identification strategy for heterogeneous causal effects in both a simplified model and a long panel version of the model in (3.3). The structural coefficients measure the causal effects of explana-

tory variables vector \mathbf{X} on the outcome Y for individuals in the population with rank variable value $U = \tau$. Identification is achieved by assuming endogeneity is due only to time-invariant (not idiosyncratic) unobservables, along with a weaker version of the monotonicity assumption of Chernozhukov and Hansen (2005b).

Second, I establish uniform consistency of the coefficient estimators over all individual rank values under a certain rate relation between n and T (the cross-sectional and time series dimensions of the data). I first estimate the individual-specific coefficients, relying on established time series OLS consistency results. Then I implement a sorting approach on a hypothetical outcome at certain \mathbf{x} with the estimated individual coefficients. Under a certain rate condition between n and T , the order of the rank variables can be recovered through the order of the hypothetical outcomes. Therefore, the estimator of coefficient at any specific rank value is determined and consistently estimated.

Third, I contribute a novel inference approach to build a uniform confidence band for the coefficient function using a Dirichlet method, based on the fact that the finite-sample uniform order statistics jointly follow an ordered Dirichlet distribution. If each individual's coefficient were known, then the unobserved individual rank variable would be the only source of uncertainty. More generally, there is also estimation error. In the general model, the individual coefficient has estimation error, which essentially comes from the idiosyncratic error. The uncertainty comes from both the rank variable and the individual coefficient estimation. For the coefficient as a function of rank, I construct confidence sets both pointwise and joint over n rank values, which I extend to a uniform confidence band if the coefficient is an increasing function of rank.

Finally, I propose a bootstrap inference method for the coefficient estimator at any rank value. I establish the asymptotic normality of the coefficient function estimator uniformly over all rank values, by applying the functional delta-method to the empirical quantile process. I establish bootstrap consistency results in a simplified model to pro-

vide the intuition and theoretic grounds for using the bootstrap inference method in the general model. Based on the uniform asymptotic normality and the bootstrap consistency results in the simplified model, I use a cluster-robust resampling method to build a bootstrap standard error and bootstrap confidence interval for the causal parameters at any specific rank value.

I also apply my new methodology in an empirical application examining the causal effect of oil wealth on a country's military spending. Many researchers (e.g., Besley and Persson, 2011; Bohn and Deacon, 2000; Collier and Hoeffler, 1998, 2004; Fearon and Laitin, 2003) have investigated the macro-level "oil-fuels-war" hypothesis. Various methods are proposed, and they arrive at different conclusions. I use my new methods to estimate the causal effect of oil wealth for countries at any specific rank, interpreting rank as a general military spending tendency. I build confidence sets using the Dirichlet method and construct a bootstrap confidence interval for the causal effect of oil wealth. Although there is some heterogeneity, I arrive at the same conclusion that oil wealth has no significant effect on military spending, as found by Cotet and Tsui (2013) assuming the effect is homogeneous.

These contributions raise additional research questions that I am pursuing in separate papers. One question relates to partial identification. When the data set has missing values or the number of time series observations is very small, point identification fails. I plan to characterize upper and lower bounds for the coefficient function in these situations. Another research question concerns a bias-variance tradeoff, motivated by empirical applications. If covariates lack variation for each individual, then the variance from the time series estimation can be very large. Grouping individuals with similar covariate values may be able to reduce the standard error and improve efficiency, if the increase in bias is small enough.

Paper structure and notation Section 3.2 reviews the literature. Section 3.3 presents identification results. Section 3.4 defines the estimator and shows uniform consistency. Section 3.5 derives confidence sets and uniform confidence bands for the coefficient function based on the Dirichlet approach. Section 3.6 derives the uniform asymptotic normality of the coefficient function estimator. It discusses bootstrap consistency and presents bootstrap confidence intervals. Section 3.7 provides empirical illustrations. Section 3.8 concludes and discusses future research.

Acronyms used include those for fixed effects (FE) and quantile regression (QR). Notationally, random and non-random vectors are respectively typeset as, e.g., \mathbf{X} and \mathbf{x} , while random and non-random scalars are typeset as X and x (with the exception of certain constants like the time series dimension T), and random and non-random matrices as $\underline{\mathbf{X}}$ and $\underline{\mathbf{x}}$; $1_{\mathcal{F}}g$ is the indicator function. The Dirichlet distribution with parameters a_1, \dots, a_K is written $\text{Dir}(a_1, \dots, a_K)$, the beta distribution $\text{Beta}(a, b)$, and the uniform distribution $\text{Unif}(a, b)$; in some cases these stand for random variables following such distributions. Let \rightharpoonup denote weak convergence. Let δ_x denote the Dirac measure.

3.2 Literature

3.2.1 Quantile regression

Quantile regression is first introduced by Koenker and Bassett (1978b). A variety of approaches have been developed to address endogeneity in quantile models. This includes the instrumental variable quantile regression methods of Chernozhukov and Hansen (2005b, 2006b, 2008), the local quantile treatment effect of Abadie, Angrist, and Imbens (2002b) (and see Melly and Wüthrich, 2017, for a survey), and the control function methods for triangular structural models in Chesher (2003), Lee (2007), and Imbens and Newey (2009), among others. Chernozhukov and Hansen (2005b) introduce

the “rank variable” idea to represent unobserved heterogeneity in quantile models. They achieve identification of structural heterogeneous effects by restricting the rank variable and imposing a monotonicity assumption on the conditional outcome as a function of the rank variable.

3.2.2 Panel quantile regression

A variety of approaches have been proposed to study quantile regression with panel data, mainly in two categories.

First, a series of papers study the panel fixed effects quantile regression (FE-QR), including Koenker (2004), Kato, Galvao, and Montes-Rojas (2012), Canay (2011), and Galvao and Wang (2015), among others. They assume the conditional quantile function is additive in an individual effect and a linear function of the explanatory variables. The individual effect cannot be consistently estimated in a short panel because the number of parameters is proportional to n , and the individual effect cannot be differenced out due to the nonlinearity of the conditional quantile operator. Consequently, these FE-QR studies assume a large- T (long panel) framework, i.e., $T \rightarrow \infty$. Koenker (2004) first proposes a penalized estimator in panel quantile regression. It assumes the individual effect is a location shift that is the same at all quantiles. Kato, Galvao, and Montes-Rojas (2012) generalize the individual effect to be a function of the quantile. They establish large-sample theory for the FE-QR estimator when both n and T go to infinity. Canay (2011) proposes a two-step estimator that first estimates the individual effect and then runs quantile regression after subtracting the individual effect from the outcome variable. It treats the idiosyncratic error as the rank variable that follows a standard uniform distribution, and it is assumed to be iid across both individual and time dimension. Frumento, Bottai, and Fernández-Val (2020) extend Canay’s (2011) model to a two-level quantile regression with the level-one quantile regression the same as Canay’s (2011)

model and the level-two quantile regression modeling the scalar-valued individual fixed effect as a quantile regression of the time-invariant characteristics.

There are several extensions of these FE-QR models. Galvao (2011) studies a dynamic FE-QR model. Galvao and Kato (2016) provide a smoothed FE-QR estimator. Harding and Lamarche (2009) use an instrumental variables approach to study the panel quantile regression with endogenous regressors. Rosen (2012) derives set identification in the FE-QR model with a short panel.

Second, the other approach to study panel quantile regression is based on the correlated random coefficient model (CRC-QR). As shown by Koenker (2005, Section 2.6), a quantile regression model can be equivalently written as a random coefficient model with a monotonicity condition and an independence assumption between the explanatory variables and the unobserved heterogeneity (rank variable) in the cross-sectional setting. With the availability of panel data, the independence assumption can be relaxed. The correlated random coefficient model starts from Chamberlain (1982). Abrevaya and Dahl (2008) study the effect of birth inputs on birthweight using quantile estimation in a random coefficient short panel model. They allow the regressors to be correlated with the unobserved individual characteristics. Arellano and Bonhomme (2016) assume a random coefficient panel data model and achieve a linear interpretation of the conditional quantile function under certain conditions. They assume the coefficient is a function of the idiosyncratic error, which follows a standard uniform distribution and is iid in both individual and time dimensions. Graham, Hahn, Poirier, and Powell (2018) propose a partially varying coefficient quantile regression model. They show this model can be derived from a random coefficient model with coefficients depending on the regressors and the unobserved individual effect. They assume the unobserved individual effect conditionally follows a standard uniform distribution and is iid in both individual and time dimensions.

There are some other studies on quantile effects in nonlinear panel data models. Chernozhukov, Fernández-Val, Hahn, and Newey (2013) study the average treatment effect and quantile treatment effects in nonlinear panel models. They treat time as randomly assigned. Under this time homogeneity condition, they establish identification, partial identification, and estimation of average and quantile treatment effects with discrete-valued regressors. Powell (2020) studies panel quantile regression with nonadditive fixed effects in an instrumental variable framework. It extends Chernozhukov and Hansen's (2005b) cross-sectional IVQR model to panel model, but with the same strong monotonicity assumption. It can estimate consistently in small- T case, but the estimation and computation of standard error can be numerically challenging, as pointed out in Baker, Smith, and Powell (2016).

3.2.3 Random coefficient panel model

The random coefficient panel model is an alternative to the fixed effect panel model. The fixed effect model considers an additive unobserved individual heterogeneity, whereas the random coefficient model uses the individual random coefficient to capture the unobserved individual heterogeneity.

The marginal effect of a variable might be varying over time and be different for different individuals depending on many factors. If all the factors are captured in the model, then the constant slope model might be reasonable. However, many variables are unobserved. Therefore, a varying coefficient model is more reasonable to capture the unobserved heterogeneity.

A variety of random coefficient panel models have been discussed, depending on different assumptions imposed on the random coefficient. The classical random-coefficient panel model assumes the random coefficient is a constant plus an idiosyncratic error. The average causal effect in this model will be the same as the causal effect in the

fixed-coefficient model. Arellano and Bonhomme (2012) assume the individual-specific coefficient is a random draw from an unrestricted conditional distribution given the regressors and identify the distributional characteristics (variance and conditional density) of the random coefficient. The correlated random coefficient (CRC) model assumes the random coefficient as a function of the regressors and/or the unobservables. Fernández-Val and Lee (2013) first use each individual’s time series observations to estimate each individual-specific coefficient, then incorporate the individual-specific unobserved heterogeneity in a GMM framework. It includes the correlated random coefficient model with random coefficient depending on the individual-specific unobserved heterogeneity as a special case. When the random coefficient is assumed to be a function of the rank variable, it becomes the panel quantile models. The existing panel quantile studies assume the rank variable is iid over time and over individual. My paper considers all the unobserved variables summarized into one scalar variable normalized on $(0, 1)$, which is the rank variable. I assume the rank variable is iid only over individuals and is time-invariant.

3.3 Identification

This section presents the panel quantile models and identification results. First, I introduce a simplified model to develop intuition. Second, I give identification results for the main model.

3.3.1 Simplified model point identification

Consider a simple random coefficient panel data structural model

$$Y_{it} = \mathbf{X}_{it}^\theta (U_i), \quad i = 1, \dots, n, \quad t = 1, \dots, T, \quad (3.5)$$

where Y_{it} is the dependent variable for individual i at time t , \mathbf{X}_{it} is the corresponding K -vector of explanatory variables, and U_i is the unobserved individual-specific rank variable normalized to $U_i \sim \text{Unif}(0, 1)$. This \mathbf{X}_{it} usually includes a constant, so the first component of $\beta(U_i)$ is the individual-specific intercept, whereas the standard FE model writes the individual-specific intercept separately. The dependence between U_i and \mathbf{X}_{it} is unrestricted. The function $\beta(\cdot): \mathbb{R} \rightarrow \mathbb{R}^K$ is deterministic. Different individuals have different coefficient vectors $\beta(U_i)$ depending on their U_i . For example, $\beta(0.5)$ is the structural coefficient vector for an individual with median level $U_i = 0.5$. The $\beta(U_i)$ is a random coefficient that can vary across individuals, but the source of the randomness is restricted to be a scalar U_i .

The K -vector coefficient $\beta(U_i)$ as a function of the unobserved individual component captures the heterogeneity in response.

The number of time periods T is assumed to be fixed and greater than or equal to the number of random coefficients K . When $T = K$, we are able to identify all individual coefficients $\beta(U_i)$ perfectly under a rank condition on \mathbf{X}_{it} . When $T > K$, having more time periods observations will add no additional information to identification because all the extra time periods' observations fit perfectly on the same line as identified with K time periods' observations.

Assumption 14 (iid sampling). Let $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{iT})^\theta$ and $\underline{\mathbf{X}}_i = (\mathbf{X}_{i1}, \dots, \mathbf{X}_{iT})^\theta$. The observables $(\mathbf{Y}_i, \underline{\mathbf{X}}_i)$ are assumed to be independent and identical distributed (iid) across $i = 1, \dots, n$.

Assumption 15 (Time-invariant rank variable). (i) The unobserved time-invariant individual rank variable U_i is iid across i . It follows a standard uniform distribution, $U_i \stackrel{iid}{\sim} \text{Unif}(0, 1)$. (ii) The relationship between $\underline{\mathbf{X}}_i$ and U_i is unrestricted.

Assumption 16 (Monotonicity). There exists known K -vector \mathbf{x}^0 in the support of \mathbf{X}_{it} such that the map $u \mapsto \beta(u) \cdot \mathbf{x}^0$ is strictly increasing on $[0, 1]$.

Assumption 17 (Continuity). At $\mathbf{X} = \mathbf{x}$, the function $u \mapsto \mathbf{x}^\theta(u)$ is continuous in u on $(0,1)$.

Assumption 18 (Rank condition). For each individual i , the observed $T \times K$ explanatory variable matrix $\underline{\mathbf{X}}_i = (\mathbf{X}_{i1}, \dots, \mathbf{X}_{iT})^\theta$ has (full column) rank K with probability one.

Assumption 15 assumes the rank variable is time-invariant and is iid only across i . Assumption 15 is different from most existing studies that assume the individual rank is iid both across i and across t . Assumption 16 is similar to the monotonicity assumption in Chernozhukov and Hansen (2005b). Assumption 16 is a weaker version that assumes monotonicity only at a single \mathbf{x} value. Assumption 17 guarantees a unique solution for identification. Assumption 18 ensures that for each individual, we are able to identify their individual-specific coefficient $\beta_i = \beta(U_i)$ using their T observations. Assumption 18 requires the K variables are linearly independent. If \mathbf{X} includes an intercept term, then it cannot include any other time-invariant variable. Moreover, Assumption 18 implies there is no “stayer” for more than $(T - K)$ periods in the data set.

Under Assumptions 14–18, we can identify all the individual coefficients from their time series dimension variation. We can also identify the coefficient at any rank value.

Theorem 5 (Identification in simplified model). *Suppose Assumptions 14–18 hold.*

(i) *For each representative individual i , its individual coefficient $\beta(U_i)$ is uniquely determined by its observations $(\mathbf{Y}_i, \underline{\mathbf{X}}_i)$, with probability one. That is, the random matrix $(\mathbf{Y}_{T-1}, \underline{\mathbf{X}}_{T-K})$ uniquely determines the random vector $\beta(U)$.*

(ii) *For any $\tau \in (0,1)$, there exists a $\beta(\tau)$ in the support of $\beta(U)$ that uniquely satisfies*

$$P(\mathbf{x}^\theta(U) = \mathbf{x}^\theta(\tau)) = \tau.$$

Generally, $\beta(\tau)$ represents the marginal effect of \mathbf{X} on Y for the individual whose rank variable value is $U = \tau$. It captures the heterogeneous effect across individuals.

Additionally, given $\mathbf{X} = \mathbf{x}$, τ can be interpreted as the marginal effect of \mathbf{X} on the τ -quantile structural function introduced by Imbens and Newey (2009). If Assumption 16 is strengthened to hold at all \mathbf{x} in the support of \mathbf{X} , as in Chernozhukov and Hansen (2005b), then $\mathbf{x}^\theta(U)$ is the τ -quantile structural function, the same as identified by Chernozhukov and Hansen (2005b).

3.3.2 Large- T point identification

I extend the simplified model to a general large- T model by adding idiosyncratic disturbances.

Consider the structural model

$$Y_{it} = \mathbf{X}_{it}^\theta(U_i) + V_{it}, \quad i = 1, \dots, n, \quad t = 1, \dots, T, \quad (3.6)$$

where Y_{it} , \mathbf{X}_{it} , and U_i are the same as before, and V_{it} is the idiosyncratic disturbance. Again, this model is a random coefficient model that can capture heterogeneous effects across individuals with endogeneity and unobserved heterogeneity.

Assumption 19 (Idiosyncratic disturbance). The idiosyncratic disturbance V_{it} satisfies

$$\mathbb{E}(\mathbf{X}_{it}V_{it}) = \mathbf{0}, \quad \text{for all } i = 1, \dots, n \text{ and } t = 1, \dots, T, \quad (3.7)$$

and each individual-specific coefficient $\tau(U_i)$ is identified.

For example, Assumption 19 is satisfied in the special case of covariance stationarity with invertible matrix $\mathbb{E}[\mathbf{X}_{it}\mathbf{X}_{it}^\theta]$.

The general model includes the same assumptions about the rank variable, monotonicity, and continuity as in the simplified model. Assumption 19 replaces Assumption 18 for identification of individual coefficients, after which the argument is the same for identification of the coefficients at different ranks τ .

Theorem 6 (Identification in general model). *In the general model, under Assumptions 14–17 and 19, the coefficients are identified at any rank value. That is, for any $\tau \in (0, 1)$, there exists $\beta(\tau)$ in the support of U uniquely satisfying*

$$P(\mathbf{x}^\theta(U) = \mathbf{x}^\theta(\tau)) = \tau.$$

Again, the parameter $\beta(\tau)$ is interpreted as the marginal effect of \mathbf{X} on outcome variable Y for an individual in the population with rank variable value $U = \tau$.

3.3.3 Relation to other models

Nonseparable idiosyncratic error

The model in (3.6) has an additive idiosyncratic error, outside the coefficient function. Alternatively, we can assume an even more general model with nonseparable idiosyncratic error,

$$Y_{it} = \mathbf{X}_{it}^\theta(U_i, V_{it}), \quad i = 1, \dots, n, \quad t = 1, \dots, T. \quad (3.8)$$

There are two ways to connect the nonseparable and additive idiosyncratic error models, (3.8) and (3.6) respectively. One is to consider model (3.6) as a special case of model (3.8), in which the idiosyncratic error only enters the intercept, not the slope coefficients, and it does so additively. That is, writing β_j for the j th component of vector β , let $\beta_1(U_i, V_{it}) = \gamma_1(U_i) + V_{it}$, $\beta_k(U_i, V_{it}) = \gamma_k(U_i)$ for $k = 2, \dots, K$, and $\mathbf{X}_{it} = (1, \mathbf{X}_{1,it}^\theta)^\theta$, so

$$Y_{it} = \mathbf{X}_{it}^\theta(U_i, V_{it}) = \gamma_1(U_i) + V_{it} + \sum_{k=2}^K X_{kit} \gamma_k(U_i) = \mathbf{X}_{it}^\theta(U_i) + V_{it}.$$

The other way to connect the two models is to assume an additive idiosyncratic error inside the coefficient function, $\beta(U_i, V_{it}) = \beta(U_i + V_{it})$, and make a Taylor expansion of $\beta(U_i + V_{it})$ at $\beta(U_i)$. If $\beta(\cdot)$ is differentiable and V_{it} is relatively small, then $\beta(U_i + V_{it})$

$(U_i) + \beta_1(U_i)V_{it}$. This is an exact equality in the special case of linear (3.6). Thus,

$$Y_{it} = \mathbf{X}_{it}^\beta (U_i + V_{it}) = \mathbf{X}_{it}^\beta (U_i) + \mathbf{X}_{it}^\beta [\beta_1(U_i)V_{it}].$$

The second term on the right-hand side satisfies the linear projection property when V_{it} is mean-zero and mean independent of \mathbf{X}_{it} and U_i ,

$$E[\mathbf{X}_{it}^\beta \mathbf{X}_{it}^\beta [\beta_1(U_i)V_{it}]g] = E[\mathbf{X}_{it}^\beta \mathbf{X}_{it}^\beta [\beta_1(U_i)] \overbrace{E[V_{it} | \mathbf{X}_{it}, U_i]}^{=0} g] = \mathbf{0}.$$

Thus, model (3.8) is approximately the same as model (3.6) with redefined idiosyncratic error.

Standard FE model

Moreover, the standard FE panel structural model can be considered as a special case of the general model studied in this paper. If we restrict the slope to be constant across individuals, let $\mathbf{X}^\beta = (1, \mathbf{X}_{-1})$ include an intercept term, and let $\alpha_i = \beta_1(U_i)$ and $\beta_k(U_i) = \beta_k$ for $k = 2, \dots, K$ with $\beta = (\beta_2, \dots, \beta_K)^\beta$, then the general model (3.6) becomes

$$Y_{it} = \mathbf{X}_{1,it}^\beta + \alpha_i + V_{it}.$$

3.4 Estimation

3.4.1 Simplified model point estimation

First, I describe the coefficient estimator at any rank value in the simplified model. Then, I show uniform consistency of this estimator. These results develop intuition and intermediate theoretical results that are useful later for the full model.

Order statistics of rank variable

Consider order statistics of the individual rank variable U_i . Let $U_{n:k}$ denote the k th order statistic in a sample of size n : $U_{n:k}$ is the k th-smallest value among $\{U_1, \dots, U_n\}$. Assumption 15 assumes $U_i \stackrel{iid}{\sim} \text{Unif}(0, 1)$, which results in the following properties from Wilks (1962, pp. 236–238):

$$U_{n:k} \sim \text{Beta}(k, n + 1 - k), \quad (3.9)$$

$$(U_{n:1}, U_{n:2}, \dots, U_{n:n}) \sim \text{Dir}(\overbrace{1, \dots, 1}^{n+1}), \quad (3.10)$$

a Dirichlet distribution with all $n + 1$ parameters equal to 1.

Although the individual rank variable U_i is unobserved, we know the joint distribution of its order statistics from (3.10), which proves useful for inference. Assumption 16 enables us to uncover the order of the unobserved individual ranks through the order of observable values. Specifically, at $\mathbf{X} = \mathbf{x}$, the ordering of the hypothetical outcome variables $Y_i = \mathbf{x}^\theta(U_i)$ represents the ordering of the unobserved individual rank variables U_i :

$$Y_{n:1} = \mathbf{x}^\theta(U_{n:1}) < Y_{n:2} = \mathbf{x}^\theta(U_{n:2}) < \dots < Y_{n:n} = \mathbf{x}^\theta(U_{n:n}). \quad (3.11)$$

Estimation

In the simplified model, we can perfectly estimate the coefficient parameter $\theta(U_i)$ from data, although the value U_i remains unknown. Without loss of generality, assume $T = K$. Under the full rank condition Assumption 18,

$$\theta(U_i) = \mathbf{X}_i^{-1} \mathbf{Y}_i, \quad i = 1, \dots, n. \quad (3.12)$$

Given the n individual coefficients $\theta(U_i)$ for $i = 1, \dots, n$, for any $\tau \in (0, 1)$, my coefficient estimator at rank $U = \tau$ is

$$\hat{\theta}(\tau) = \theta(U_{n:\lceil n\tau \rceil}), \quad (3.13)$$

where $U_{n:\lceil \tau n \rceil}$ is the $\lceil \tau n \rceil$ th order statistic, and $\lceil \tau n \rceil$ denotes the ceiling function of τn , i.e., the least integer greater than or equal to τn . Because $0 < \tau < 1$, $1 \leq \lceil \tau n \rceil \leq n$. Although the U_i values are unobserved, their order can be discovered through the order of hypothetical outcome values Y_i at $\mathbf{X} = \mathbf{x}$ under Assumption 16, which determines $(U_{n:k})$.

Define the permutation $\sigma(\cdot)$ to associate the individual “ i ” with its order “ k ”:

$$\sigma(k) = \hat{f}_i : U_{n:k} = U_i g. \quad (3.14)$$

So we can write the coefficient $(U_{n:k}) = (U_{\sigma(k)})$, and the estimator is $\hat{\tau} = (U_{\sigma(\lceil \tau n \rceil)})$. The permutation $\sigma(\cdot)$ can be learned from the Y_i ordering because $Y_{n:k} = Y_{\sigma(k)}$.

Specifically, the estimation of $\hat{\tau}$ consists of four steps.

Step 1: For each individual, use its T time period observations to (perfectly) estimate its individual coefficient: $(U_i) = \mathbf{X}_i^{-1} \mathbf{Y}_i$, $i = 1, \dots, n$.

Step 2: Given all perfectly estimated individual coefficients $(U_1), (U_2), \dots, (U_n)$ from Step 1, compute the n corresponding outcome values Y at $\mathbf{X} = \mathbf{x}$,

$$Y_i = \mathbf{x}^\theta(U_i), \quad i = 1, \dots, n. \quad (3.15)$$

Step 3: Sort the outcome values $\{Y_i\}_{i=1}^n$ in increasing order. The order of the Y represents the order of the U : the permutation $\sigma(\cdot)$ from (3.14) satisfies $Y_{n:k} = Y_{\sigma(k)}$ for $k = 1, \dots, n$.

Step 4: Using the (U_i) from Step 1 and the permutation $\sigma(\cdot)$ from Step 3, the coefficient estimator at rank τ is

$$\hat{\tau} = (U_{\sigma(\lceil \tau n \rceil)}). \quad (3.16)$$

Large- n uniform consistency

This subsection shows the estimator in the simplified model is uniformly consistent over $0 < \tau < 1$.

Assumption 20 (Uniform continuity). Each element of the coefficient $\beta_k(\cdot)$ is a uniformly continuous function on $(0, 1)$. That is, writing $\beta_k(\cdot) = (\beta_1(\cdot), \dots, \beta_K(\cdot))^\theta$, for any $k = 1, \dots, K$,

$$\exists \epsilon > 0, \exists \delta > 0 \text{ such that } \forall x, y \in (0, 1), |x - y| < \delta \Rightarrow |\beta_k(x) - \beta_k(y)| < \epsilon.$$

Assumption 20 is a stronger condition than Assumption 17. With the uniform continuous coefficient function assumed in Assumption 20 and the uniform consistency of the uniform sample quantile function from Shorack and Wellner (1986, Thm. 3, p. 95), I establish the uniform consistency of the coefficient function estimator $\hat{\beta}_k(\cdot)$ on $(0, 1)$.

Theorem 7 (Uniform consistency in simplified model). *Under Assumptions 14–16, 18, and 20, the coefficient function estimator uniformly converges to the true coefficient function on $(0, 1)$:*

$$\sup_{\tau \in (0, 1)} \left\| \hat{\beta}_k(\tau) - \beta_k(\tau) \right\| \xrightarrow{a.s.} 0 \text{ as } n \rightarrow \infty.$$

3.4.2 Large- T point estimation

I now describe the estimator in the general large- T model and show consistency. I assume that each individual's coefficient vector estimator is consistent as $T \rightarrow \infty$, based on well-established time series theories. I show the uniform consistency of the estimator as both T and n go to infinity under a certain rate condition.

Estimation consists of two steps. First, the individual coefficients $\hat{\beta}_k(U_i)$ are estimated for each i . Second, the order of the n unobserved individual ranks U_i is estimated based on the estimated coefficients, and the estimator $\hat{\beta}_k(\tau)$ at any rank $\tau \in (0, 1)$ is the estimated coefficient at the estimated $n\tau$ th order statistic.

One of the major differences in estimation between the general model and the simplified model is that instead of perfectly estimating the true individual coefficients in the simplified model, the estimated individual coefficients now have estimation error in general model. This might affect the ordering of the fitted outcome values \hat{Y}_i , implying an imprecise ordering of the unobserved rank values U_i .

Time series OLS consistency

Let \hat{U}_i denote the time series OLS (TS-OLS) estimator for individual i 's coefficient (U_i) ,

$$\hat{U}_i = \left(\frac{1}{T} \sum_{t=1}^T \mathbf{X}_{it} \mathbf{X}_{it}' \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^T \mathbf{X}_{it} Y_{it} \right). \quad (3.17)$$

There are many large-sample results for the TS-OLS estimator under various assumptions. For example, Hamilton (1994, Ch. 8.2) lists five different scenarios under which the TS-OLS estimator is consistent. Relying on these well-established results, I assume that the TS-OLS estimator of each individual-specific coefficient is consistent as $T \rightarrow \infty$.

Assumption 21 (TS-OLS consistency). (i) The TS-OLS coefficient estimator is consistent for all individuals.

$$\hat{U}_i \xrightarrow{p} U_i \text{ as } T \rightarrow \infty, \quad \forall i = 1, \dots, n. \quad (3.18)$$

(ii) The convergence rate is $T^{-\kappa}$ for some positive $\kappa > 0$:

$$\hat{U}_i - U_i = O_p(T^{-\kappa}) \text{ as } T \rightarrow \infty, \quad \forall i = 1, \dots, n. \quad (3.19)$$

(iii) As $n \rightarrow \infty$, the TS-OLS coefficient estimates are uniformly consistent in probability, i.e., for any positive $\epsilon > 0$,

$$\mathbb{P} \left(\sup_{i=1, \dots, n} \left\| \hat{U}_i - U_i \right\| > \epsilon \right) \rightarrow 0 \text{ as } n, T \rightarrow \infty. \quad (3.20)$$

The norm in Assumption 21(iii) is taken as the supremum norm for simplicity, but it

can be interpreted as any L_p norm due to the equivalence of norms in finite-dimensional space. Assumption 21 implicitly restricts how fast T must grow with respect to n . A sufficient n/T rate relation is given in the uniform consistency result of Theorem 9.

Large- T estimation

Define the estimated permutation $\hat{\sigma}(\cdot)$ to associate the individual i with its fitted outcome value order k :

$$\hat{\sigma}(k) = \{i : \hat{Y}_{n:k} = \hat{Y}_i\}. \quad (3.21)$$

The estimated permutation $\hat{\sigma}(k)$ depends on the data and the \mathbf{x} value. Later Proposition 8 shows the estimated order equals the true order with probability approaching one under a certain rate relation between n and T , so $P(\hat{\sigma}(\cdot) = \sigma(\cdot)) \rightarrow 1$.

Specifically, the estimation of $\hat{\tau}$ consists of four steps.

Step 1: For each individual, use its T time period observations to estimate its individual coefficient: $\hat{U}_i = (\mathbf{X}_i^\top \mathbf{X}_i)^{-1} (\mathbf{X}_i^\top \mathbf{Y}_i)$, $i = 1, \dots, n$.

Step 2: Given the estimated individual coefficients $\hat{U}_1, \hat{U}_2, \dots, \hat{U}_n$ from Step 1, compute the n fitted outcome values \hat{Y}_i at $\mathbf{X} = \mathbf{x}$,

$$\hat{Y}_i = \mathbf{x}^\top \hat{U}_i, \quad i = 1, \dots, n. \quad (3.22)$$

Step 3: Sort the fitted outcome values \hat{Y}_i in increasing order: $\hat{Y}_{n:1} \leq \hat{Y}_{n:2} \leq \dots \leq \hat{Y}_{n:n}$. The order represents the estimated order of the U_i : the estimated permutation $\hat{\sigma}(\cdot)$ satisfies $\hat{Y}_{n:k} = \hat{Y}_{\hat{\sigma}(k)}$ for $k = 1, \dots, n$.

Step 4: Using the \hat{U}_i from Step 1 and the estimated permutation $\hat{\sigma}(\cdot)$ from Step 3, the estimator at rank τ is

$$\hat{\tau} = \hat{U}_{\hat{\sigma}(d_{n\tau})}. \quad (3.23)$$

Large- T , large- n uniform consistency

The estimator in the general model is uniformly consistent over $0 < \tau < 1$, under the following additional assumption.

Assumption 22. (Extended monotonicity condition) Let $g(u) = \mathbf{x}^\theta(u)$. The derivative function $g^\theta(u)$ has a positive lower bound L on $(0, 1)$: $\inf_{0 < u < 1} g^\theta(u) = L > 0$.

The monotonicity condition Assumption 16 assumes the function $g(u) = \mathbf{x}^\theta(u)$ is strictly increasing on $(0, 1)$. Assumption 22 strengthens this slightly, so $g(u)$ has a strictly positive lower bound of its derivative. It prevents the situation that the function $g(u)$ is almost flat on some intervals; for example, $g(u) = \sin(u\pi/2)$ satisfies Assumption 16 because $g^\theta(u) > 0$ for all $0 < u < 1$ but violates Assumption 22 because $\lim_{u \rightarrow 1} g^\theta(u) = 0$. This restriction is not a strong assumption in the sense that it still allows an individual element function of the vector-valued coefficient $\beta(u)$ to be non-monotone or almost flat.

Let $Z_i = k_{T,i}^\wedge$ denote a random variable based on the TS-OLS estimator. The Z_i are iid over i from Assumption 14.

Assumption 23. For all $T > T_0$, $E(T^\kappa Z_i) < 1$, where $T_0 \geq \mathbb{N}$ is some constant.

Assumption 23 helps establish the stochastic order of $\max_{1 \leq i \leq n} k_{T,i}^\wedge$ in terms of n , which then translates to T using a specified n/T rate relation.

Proposition 8 shows a sufficient n/T rate relation that guarantees the true order of the U_i can be discovered from the order of \hat{Y}_i . That is, Proposition 8 implies the estimated permutation is asymptotically the same as the true permutation.

Proposition 8 (Estimated permutation consistency). *Suppose Assumptions 14–17 and 19–23 hold. Under the rate relation $n = o(T^{\kappa/(3+\delta)})$, for some small positive $\delta > 0$,*

$$\mathbb{P}\left(\frac{\min_{1 \leq j \leq n} Y_{n:2} \cdots Y_{n:n-1} g}{2} > \max_{1 \leq i \leq n} \hat{Y}_{T,i} - Y_i\right) \rightarrow 1 \text{ as } n, T \rightarrow \infty. \quad (3.24)$$

The coefficient estimator defined based on the estimated permutation is uniformly consistent over $(0, 1)$.

Theorem 9 (Uniform consistency in general model). *Under Assumptions 14–16 and 19–23, as n and T go to infinity with the rate relation $n = o(T^{\kappa/(3+\delta)})$, the estimated coefficient function uniformly converges to the true coefficient function,*

$$\sup_{\tau \in (0,1)} \left\| \hat{\beta}(\tau) - \beta(\tau) \right\| = o_p(1).$$

3.5 Inference

In this section, I provide a novel method to obtain confidence sets or uniform confidence bands for the coefficient function $\beta(\cdot)$ on $(0, 1)$. My approach uses the ordered Dirichlet distribution of uniform order statistics to capture uncertainty in the unobserved rank variable values. It sheds light on the shape of the coefficient function, showing patterns in the heterogeneous effects across individuals.

First, I derive results in the simplified model to develop intuition and intermediate theoretical results. Specifically, I derive confidence sets for points on the coefficient function as well as a uniform confidence band if the coefficient function is assumed monotonic. Then, I generalize these to the general model.

Notationally, the scalar $\beta(\cdot)$ refers to any single component of the vector-valued $\beta(\cdot)$.

3.5.1 Dirichlet method in simplified model

In the simplified model, we perfectly estimate all the individual coefficients. There is no uncertainty in the vertical direction in the coefficient function graph. The uncertainty comes from the random rank variable U in the horizontal direction. Therefore, the confidence interval for individual coefficient $\beta(U_{n:k})$ is horizontal, instead of a conventional vertical one.

Pointwise confidence interval

This subsection presents individual pointwise confidence intervals. These are depicted as the red intervals in the left graph in Figure 3.1.

Define the k th horizontal interval as

$$I_k = \{x \mid f(x, y) \in [a_k, b_k], y = \beta(U_{n:k})\mathcal{G}\}, \quad (3.25)$$

where a_k and b_k are the $\alpha/2$ -quantile and $(1 - \alpha/2)$ -quantile of the $\text{Beta}(k, n + 1 - k)$ distribution. Using the uniform order statistic property in (3.9), $\mathbb{P}(a_k \leq U_{n:k} \leq b_k) = 1 - \alpha$, so the interval I_k is an equal-tailed $100(1 - \alpha)\%$ confidence interval in the sense that it contains $\beta(U_{n:k})$ with probability $1 - \alpha$, with equal probability of $\beta(U_{n:k})$ being too far to the left or right. Having fixed endpoints (a_k, b_k) and random $\beta(U_{n:k})$ is different than the usual case of a random interval containing a fixed point, but the interpretation of frequentist coverage probability is otherwise the same. It can also be interpreted that the coefficient function $\beta(\cdot)$ crosses the interval I_k with $1 - \alpha$ probability.

Each single interval I_k ($k = 1, \dots, n$) includes the corresponding individual-specific coefficient $\beta(U_{n:k})$ with exact probability $1 - \alpha$, even with small n . However, as usual, these n separate intervals $\prod_{k=1}^n I_k$ have less than $1 - \alpha$ joint coverage probability. In the next subsection, I construct confidence sets that can cover the coefficient function with the desired joint coverage probability.

Joint confidence sets and uniform confidence bands

First, I construct confidence sets for the n individual coefficients $\beta(U_{n:1}), \dots, \beta(U_{n:n})$ jointly, based on results from Goldman and Kaplan (2018). Second, I restrict the coefficient function to be monotone¹ and establish uniform confidence bands for the whole

¹The monotone coefficient function in this section is different from the monotonicity condition in Assumption 16. The monotonicity condition Assumption 16 is monotonicity of the inner product of the vector $\mathbf{X} = \mathbf{x}$ and the vector-valued coefficient $\beta(\cdot)$ as a function of the rank variable, whereas here the monotone coefficient function is referring to each scalar-valued coefficient function $\beta(\cdot)$ itself being

coefficient function.

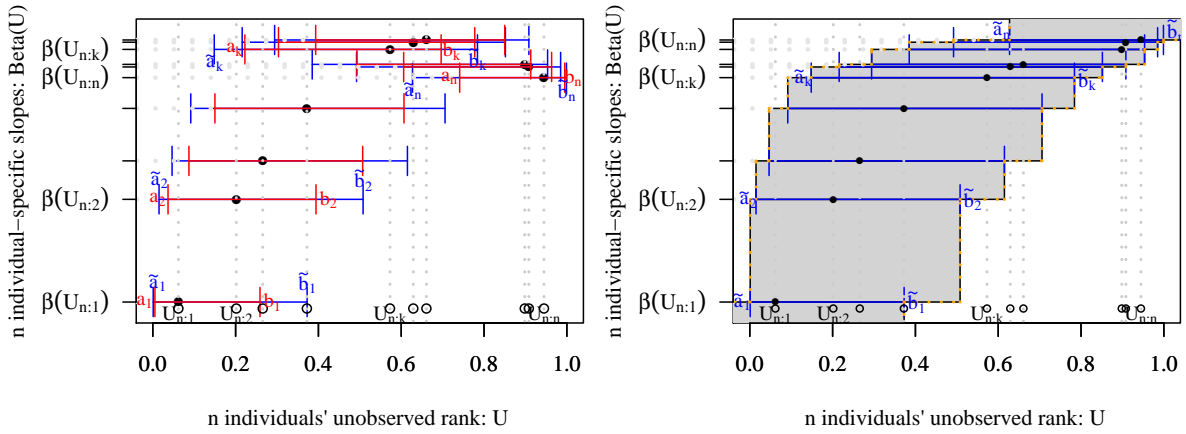


Figure 3.1: Left: pointwise confidence intervals (red) and joint confidence sets (blue) for the n points on the generic coefficient function in Case 1. Right: uniform confidence band for the monotone coefficient function $\beta(\cdot)$ in Case 2.

Case 1: Generic coefficient function and confidence sets

Goldman and Kaplan (2018) provide joint confidence intervals for the n jointly ordered Dirichlet-distributed uniform order statistics $(U_{n:1}, \dots, U_{n:n})$. That is, given any $\alpha \in (0, 1)$,

$$P(\mathbf{a}_1 < U_{n:1} < \mathbf{b}_1 \text{ and } \dots \text{ and } \mathbf{a}_n < U_{n:n} < \mathbf{b}_n) = 1 - \alpha, \quad (3.26)$$

where \mathbf{a}_k is the $\alpha/2$ -quantile of the $\text{Beta}(k, n + 1 - k)$ distribution; \mathbf{b}_k is the $(1 - \alpha/2)$ -quantile of the $\text{Beta}(k, n + 1 - k)$ distribution; and α is theoretically from Theorem 5 in Goldman and Kaplan (2018). This α can be analytically approximated from Fact 6 in Goldman and Kaplan (2018),

$$\alpha = \exp \left\{ -c_1(\alpha) \frac{c_2(\alpha)}{\ln \ln n} - c_3(\alpha) [\ln n]^{c_4(\alpha)} g \right\}, \quad (3.27)$$

where $c_1(\alpha) = 2.75 - 1.04 \ln(\alpha)$, $c_2(\alpha) = 4.76 - 1.20\alpha$, $c_3(\alpha) = 1.15 - 2.39\alpha$ and $c_4(\alpha) = 3.96 + 1.72\alpha^{0.171}$.

monotone.

Define the n intervals in the coefficient function graph

$$t_k = f(x, y) \mid_{\alpha_k} \quad x \quad \tilde{b}_k; y = \beta(U_{n:k})g, \quad k = 1, \dots, n, \quad (3.28)$$

where the α_k and \tilde{b}_k are the same as in (3.26). These t_k are depicted as the blue intervals in the left graph in Figure 3.1. They jointly cover the n individual coefficients with at least $1 - \alpha$ probability.

Theorem 10 (Confidence set for generic coefficient function in simplified model). *Under Assumptions 14–18, the joint intervals $[_{k=1}^n t_k$ form a confidence set for the n individual coefficient points on the generic coefficient function at a $100(1 - \alpha)\%$ confidence level:*

$$1 - \alpha \leq \mathbb{P}\left(f[_{k=1}^n(U_{n:k}, \beta(U_{n:k}))g \in f[_{k=1}^n t_k g\right).$$

It can be interpreted that the coefficient function $\beta(\cdot)$ crosses all n intervals $[_{k=1}^n t_k$ jointly with probability at least $1 - \alpha$. Equality is achieved when the n individual coefficients are distinct, in which case the coefficient function crosses the joint intervals $[_{k=1}^n t_k$ with exact probability $1 - \alpha$.

Generalization to vector-valued coefficient function

We can generalize the confidence sets for the n coefficient estimates on the scalar-valued coefficient function $\beta(\cdot)$ to that on the vector-valued coefficient function $\beta(\cdot)$. The former can be considered as one element function of the coefficient in the latter case.

The uncertainty of the vector-valued coefficient function in the simplified model only comes from the uncertainty of the individual rank variable. First, in the simplified model, there is no vertical uncertainty from estimation error of the individual coefficients. The only uncertainty comes from the horizontal direction about the individual rank order statistics $U_{n:k}$. Second, the model assumes that each element of the random coefficient vector $\beta(U)$ depends on the same scalar U . There is no more uncertainty about the full

vector than any individual component. Both reduce to the uncertainty about the same scalar order statistics $U_{n:k}$.

Consider a K -vector coefficient function $\boldsymbol{\beta}(\cdot) = (\beta_1(\cdot), \dots, \beta_K(\cdot))^\theta$. We can construct the confidence sets on each of the element coefficient functions $\beta_1(\cdot), \dots, \beta_K(\cdot)$. That is, we draw the n intervals as the confidence sets in each of the K graphs with axes $u \nabla \beta_1(u), u \nabla \beta_2(u), \dots, u \nabla \beta_K(u)$. The horizontal endpoints $f(\mathbf{a}_j, \mathbf{b}_j) \mathcal{G}_{j=1}^n$ of the n intervals in the K graphs are the same. The only difference is the vertical axis values in the K graphs.

Consider the l th element function of the vector-valued coefficient. Define the n intervals

$$t_k^l = f(x, y) \int \mathbf{a}_k \quad x \quad \mathbf{b}_k; y = \beta_l(U_{n:k})g, \quad k = 1, \dots, n,$$

where the \mathbf{a}_k and \mathbf{b}_k are the same as in (3.26). From Theorem 10, the confidence set $[\prod_{k=1}^n t_k^l$ covers the n individual coefficients on the l th element coefficient function with probability $1 - \alpha$ at least.

Let the ‘‘vertical’’ dimension denote the K -vector coefficient. Define the K -dimensional intervals

$$\mathcal{I}_k = \left\{ (x, \mathbf{y}) \int \mathbf{a}_k \quad x \quad \mathbf{b}_k; \mathbf{y} = (\beta_1(U_{n:k}), \dots, \beta_K(U_{n:k}))^\theta \right\}. \quad (3.29)$$

The confidence set $[\prod_{k=1}^n \mathcal{I}_k$, which consists of n K -dimensional intervals, jointly covers the n vector-valued points $(U_{n:1}), \dots, (U_{n:n})$ on the vector-valued coefficient function with probability $1 - \alpha$ at least.

Corollary 11 (Confidence set for vector-valued generic coefficient function in simplified model). *Under Assumptions 14–18, with a vector-valued coefficient function $\boldsymbol{\beta}(\cdot)$, using the definitions in (3.26), (3.27), and (3.29), $[\prod_{k=1}^n \mathcal{I}_k$ is a $100(1 - \alpha)\%$ confidence set for the n vector-valued individual coefficients $(U_{n:1}), \dots, (U_{n:n})$:*

$$1 - \alpha = \mathbb{P}\left(f\left[\prod_{k=1}^n (U_{n:k}, (U_{n:k}))g \quad \left\{ \prod_{k=1}^n \mathcal{I}_k \right\} \right) \right).$$

The intuition for Corollary 11 is that each element of the vector coefficient function depends on the same scalar U ; $U_{n:k}$ is the only source of uncertainty of each element coefficient function. For a vector-valued coefficient function, we can construct the confidence set on each element coefficient function at the same time. It has the property that the first element function $\beta_1(\cdot)$ crosses its confidence set $[\tau_{k=1}^n, \tau_k^1]$ if and only if the second element function $\beta_2(\cdot)$ crosses its confidence set $[\tau_{k=1}^n, \tau_k^2]$ if and only if the third element function $\beta_3(\cdot)$ crosses its confidence set, etc. Each element coefficient function going through its corresponding n -interval confidence set is the same as the vector-valued coefficient function going through the n K -dimensional intervals.

Case 2: Monotonic coefficient function and uniform confidence bands

Assume the coefficient function is strictly monotone. Monotonicity enables us to build a uniform confidence band for the whole coefficient function based on the n individual coefficients. For simplicity, I focus on a scalar-valued coefficient function. The results can be generalized to vector-valued coefficient functions as well.

With monotonicity, we are able to bound the coefficient function value between two observed coefficients $\beta(U_i)$ and $\beta(U_j)$. That is, a monotonic coefficient function guarantees that for any value u in between the U_i and U_j values, the coefficient function value $\beta(u)$ is between the two observed coefficient values $\beta(U_i)$ and $\beta(U_j)$.

Figure 3.2 illustrates the difference between the confidence sets in the generic coefficient function case and the uniform confidence bands in the monotone coefficient function case. Without loss of generality, assume the coefficient function $\beta(\cdot)$ is increasing. Consider two adjacent rank variable order statistics and their coefficients $\beta(U_{n:k})$ and $\beta(U_{n:k+1})$. When the coefficient function is unrestricted, we only know that (with high probability) the coefficient function goes through the two horizontal intervals τ_k and τ_{k+1} . The coefficient function can cross the two horizontal intervals from any direction, either from top-left to bottom-right or from bottom-left to top-right. In contrast, when

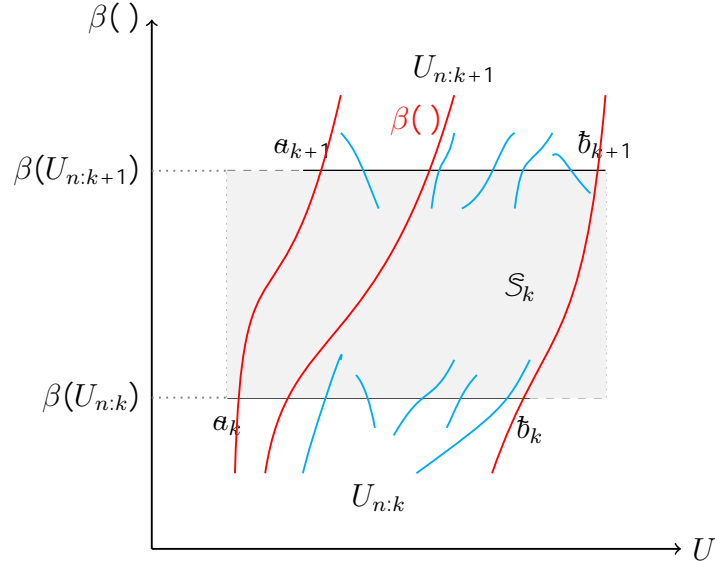


Figure 3.2: Comparison between Case 1 and Case 2. In Case 2, the red nondecreasing curves that are within the shaded gray area indicate the possible coefficient functions consistent with the two intervals t_k and t_{k+1} . In Case 1, we only know that the possible coefficient function (denoted as the blue curves) crosses t_k and t_{k+1} ; there is no information about the coefficient function for the value between $\beta(U_{n:k})$ and $\beta(U_{n:k+1})$.

the coefficient function is increasing, the two intervals t_k and t_{k+1} can be strengthened to a confidence band for the coefficient function. That is, for all rank values between $U_{n:k}$ and $U_{n:k+1}$, the coefficient value $\beta(\cdot)$ has to be between $\beta(U_{n:k})$ and $\beta(U_{n:k+1})$. That gives the rectangle S_k defined later. Additionally, it restricts a monotonic coefficient function to cross all horizontal intervals from the same direction: either all from bottom-left to top-right or all from top-left to bottom-right.

To construct the band, first construct the $n - 1$ rectangles S_1, S_2, \dots, S_{n-1} ,

$$\begin{aligned} S_k &= [a_k, b_{k+1}] \times [\beta(U_{n:k}), \beta(U_{n:k+1})] \\ &= \{f(x, y) \mid a_k \leq x \leq b_{k+1}; \beta(U_{n:k}) \leq y \leq \beta(U_{n:k+1})\}, \quad k = 1, \dots, n - 1, \end{aligned} \quad (3.30)$$

and then construct two open-ended areas S_0 and S_n ,

$$S_0 = [1, b_1] \times [1, \beta(U_{n:1})] = \{f(x, y) \mid 1 \leq x \leq b_1; y \leq \beta(U_{n:1})\},$$

$$S_n = [\mathbf{a}_n, + 1] [\beta(U_{n:n}), + 1] = f(x, y) j x \quad \mathbf{a}_n; y \quad \beta(U_{n:n})g,$$

where \mathbf{a}_k and \mathbf{b}_k are the same as in (3.26). Define the total area S as

$$S = \int_{k=0}^n S_k. \quad (3.31)$$

The area S , which is composed of the $n - 1$ rectangles plus two open-ended areas, provides a uniform confidence band for the monotone coefficient function $\beta(\cdot)$. It is depicted as the gray area in the right graph in Figure 3.1.

Theorem 12 (Uniform confidence band for monotone coefficient function in simplified model). *Under Assumptions 14–18, the area S is a uniform confidence band for the strictly monotone coefficient function at a $100(1 - \alpha)\%$ confidence level:*

$$1 - \alpha = P(f(u, y) : y = \beta(u), 0 \leq u \leq 1 | g \in S).$$

The vertical intervals of the uniform confidence band collapse as n goes to infinity because $(U_{n:k+1} - U_{n:k}) \sim \text{Beta}(1, n) \xrightarrow{P} 0$ for any $k = 1, \dots, n - 1$ and $\beta(\cdot)$ is continuous. Although not as immediately obvious, the horizontal intervals collapse as well as $n \rightarrow \infty$, since each interval for the random variable $U_{n:k} \sim \text{Beta}(k, n + 1 - k)$ shrinks to a point $U_{n:n} \xrightarrow{P} \tau$. Therefore, the uniform confidence band collapses in both horizontal and vertical directions to an infinitesimal small point, so it traces out the true coefficient function $\beta(\cdot)$.

3.5.2 Dirichlet method in general model

This subsection extends the Dirichlet inference method to the large- T general model. First, I establish the pointwise confidence set for each individual coefficient. Second, I build joint confidence sets for the n individual coefficients.

In the general model, we no longer observe the exact individual coefficient parameters.

Instead, we observe n individual TS-OLS coefficient estimates $\hat{\beta}(U_{n:k})$ for $k = 1, \dots, n$. Now, there are two dimensions of uncertainty in the coefficient function graph. One is from the uncertainty in the rank variable $U_{n:k}$, since it is an unobserved random variable following a beta distribution. The other source of uncertainty is from the TS-OLS estimation of coefficient, which essentially comes from the idiosyncratic error V_{it} . Therefore, the confidence set is made up of rectangles in the general model, in contrast to the horizontal intervals in the simplified model.

Assumption 24 (Inference for TS-OLS individual coefficients). Given n individuals, for any $\alpha \in (0, 1)$, we can obtain confidence interval T_i^α for each individual i 's coefficient β_i with asymptotic coverage probability $1 - \alpha$ as $T \rightarrow \infty$:

$$\lim_{T \rightarrow \infty} \mathbb{P}(\beta_i \in T_i^\alpha) = 1 - \alpha.$$

The confidence interval T_i^α depends implicitly on the TS-OLS estimator $\hat{\beta}_i$. For example, when the time dimension observations are iid and regular OLS results apply ($\kappa = 1/2$ in Assumption 21), we can write the asymptotic 95% confidence interval as $T_i^{0.05} = \hat{\beta}(U_i) \pm 1.96\hat{\sigma}/\sqrt{T}$, for some standard deviation estimator $\hat{\sigma}$ that is consistent for the asymptotic variance of $\sqrt{T}(\hat{\beta}_i - \beta_i)$. The parameter α denotes nominal coverage probability $1 - \alpha$ of the confidence interval T_i^α .

Another complication is that TS-OLS estimation error can make the estimated permutation $\hat{\sigma}(\cdot)$ from (3.21) not equal the true permutation $\sigma(\cdot)$ in finite samples. However, Proposition 8 shows that they are equal asymptotically with probability approaching one. Thus, at least asymptotically, this source of error has a negligible effect on coverage probability.

Pointwise confidence set

Define the rectangle

$$W_k^{(\alpha_1, \alpha_2)} = [a_k^{\alpha_1}, b_k^{\alpha_1}] \times T_{\delta(k)}^{\alpha_2} = \left\{ (x, y) \mid a_k^{\alpha_1} \leq x \leq b_k^{\alpha_1}; y \geq T_{\delta(k)}^{\alpha_2} \right\}, \quad k = 1, \dots, n, \quad (3.32)$$

where $a_k^{\alpha_1}$ and $b_k^{\alpha_1}$ are from (3.25) with coverage probability $1 - \alpha_1$, and $T_{\delta(k)}^{\alpha_2}$ is defined through the estimated permutation with asymptotic coverage probability $1 - \alpha_2$. The rectangle provides a pointwise confidence set as described in Theorem 13.

Theorem 13 (Pointwise confidence set for individual coefficient in the fixed- n , large- T general model). *Let Assumptions 14–17, 19, and 21–24 hold. For any $\alpha \in (0, 1)$ and any constant $w \in (0, 1)$, let $\alpha_1 = w\alpha$ and $\alpha_2 = (1 - w)\alpha$. Then, the rectangle $W_k^{(\alpha_1, \alpha_2)}$ contains the point $(U_{n:k}, \beta(U_{n:k}))$ with asymptotic coverage probability bounded below by $1 - \alpha$ as $T \rightarrow \infty$: $\lim_{T \rightarrow \infty} \mathbb{P}(U_{n:k}, \beta(U_{n:k})) \in W_k^{(\alpha_1, \alpha_2)} \geq 1 - \alpha$ for any $k = 1, \dots, n$.*

The pointwise confidence sets each help quantify uncertainty about a single point on the function $\beta(\cdot)$. They are smaller than the rectangles in the joint confidence sets and uniform confidence band introduced below, but they may not properly capture the overall shape of the coefficient function.

Joint confidence set

Parallel to Section 3.5.1 for the simplified model, this subsection provides joint confidence sets for n points on the coefficient function in the general model.

Given some constant $\alpha_1, \alpha_2 \in (0, 1)$, define the rectangles in the coefficient function graph

$$R_k^{(\alpha_1, \alpha_2)} = [a_k^{\alpha_1}, b_k^{\alpha_1}] \times T_{\delta(k)}^{\alpha_2} = \left\{ (x, y) \mid a_k^{\alpha_1} \leq x \leq b_k^{\alpha_1}; y \geq T_{\delta(k)}^{\alpha_2} \right\}, \quad k = 1, \dots, n, \quad (3.33)$$

where $a_k^{\alpha_1}$ and $b_k^{\alpha_1}$ are from (3.26) with (joint) coverage probability $1 - \alpha_1$, and $T_{\delta(k)}^{\alpha_2}$ is

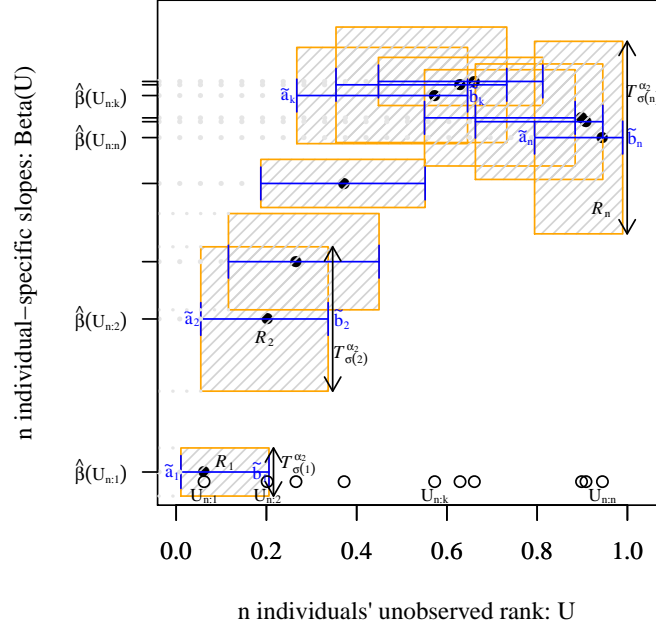


Figure 3.3: Confidence set for n points on the generic coefficient function in general model.

defined through the estimated permutation $\hat{\delta}(\cdot)$. Define the total area $R^{(\alpha_1, \alpha_2)}$ as

$$R^{(\alpha_1, \alpha_2)} = \prod_{k=1}^n R_k^{(\alpha_1, \alpha_2)}. \quad (3.34)$$

It is depicted as the n gray boxes in Figure 3.3. It provides a joint confidence set as described in Theorem 14.

Theorem 14 (Joint confidence set for generic coefficient function in the fixed- n , large- T general model). *Under Assumptions 14–17, 19, and 21–24, in the general model, given n individuals, the area $R^{(\alpha_1, \alpha_2)}$ is a confidence set for the n individual coefficients on the generic coefficient function. Specifically, for any $\alpha \in (0, 1)$, with $\alpha_1 = \alpha/2$ and $\alpha_2 = 1 - (\alpha/2)^{1/n}$, the confidence set $R^{(\alpha_1, \alpha_2)}$ has asymptotic coverage probability bounded below by $1 - \alpha$ as $T \rightarrow \infty$:*

$$\lim_{T \rightarrow \infty} \mathbb{P}(\bigcap_{k=1}^n (U_{n:k}, \beta(U_{n:k})) \in R^{(\alpha_1, \alpha_2)}) \geq 1 - \alpha.$$

Generally, the α_1 and α_2 in Theorem 14 can be generalized to any values that satisfy

$\alpha_1 + 1 - (1 - \alpha_2)^n = \alpha$, with $0 < \alpha_1, \alpha_2 < 1$. For example, let $\alpha_1 = w\alpha$ and $\alpha_2 = 1 - (1 - (1 - w)\alpha)^{1/n}$, where w is a constant $0 < w < 1$. It satisfies $\alpha_1 + 1 - (1 - \alpha_2)^n = \alpha$, and the confidence set $\mathcal{R}^{(\alpha_1, \alpha_2)}$ still has at least $1 - \alpha$ asymptotic coverage probability.

The α_1 and α_2 relationship can be used to adjust the width and height of the confidence set rectangles. As α_1 increases, the confidence level $1 - \alpha_1$ decreases, so each horizontal interval gets narrower. As α_2 decreases, the confidence level $1 - \alpha_2$ increases, so each vertical interval gets taller. Therefore, if the n rectangles of the confidence set are initially wide and short, we can increase α_1 and decrease α_2 to narrow the width and increase the height. Conversely, if the n rectangles are initially narrow and tall, we can decrease α_1 and increase α_2 to widen and shorten the rectangles.

The confidence set consisting of n boxes in Figure 3.3 will shrink to n horizontal intervals as in the left figure in Figure 3.1 as $T \rightarrow \infty$. This is because the horizontal intervals are the joint confidence set for the n uniform order statistics. The horizontal interval length depends on n , but not on the time dimension T . The vertical intervals depend on both n and T , since they have to satisfy the joint coverage of all n individual coefficients. In the fixed- n , large- T general model, the horizontal interval is fixed as T^{-1} , whereas the vertical interval height shrinks to zero.

If the coefficient function is monotone, then the confidence set in Figure 3.3 extends to a uniform confidence band, parallel to the extension from the left graph to the right graph in Figure 3.1 for the simplified model. For example, if $\beta(\cdot)$ is increasing, instead of two separate rectangles $\mathcal{R}_k^{(\alpha_1, \alpha_2)}$ and $\mathcal{R}_{k+1}^{(\alpha_1, \alpha_2)}$, the uniform confidence band expands to the box starting from the bottom-left corner of $\mathcal{R}_k^{(\alpha_1, \alpha_2)}$ to the top-right corner of $\mathcal{R}_{k+1}^{(\alpha_1, \alpha_2)}$.

3.6 Bootstrap inference

I show the uniform asymptotic normality of the coefficient function estimator in both the simplified model and the general model. Then, I provide bootstrap confidence intervals and show bootstrap consistency.

3.6.1 Asymptotic normality

I establish the uniform asymptotic normality of the coefficient function estimator by applying the functional delta-method to the well-known Donsker's theorem, assuming differentiability of the coefficient function.

Donsker's theorem (e.g., van der Vaart, 1998b, Thm. 19.3) states the asymptotic Gaussianity of the empirical distribution function. Letting F_n be the sequence of the empirical distribution function indexed by sample size n , assuming iid sampling from distribution function F , then

$$\sqrt{n}(F_n - F) \xrightarrow{D} G \text{ in } D[-1, +1], \text{ as } n \rightarrow \infty, \quad (3.35)$$

where G is the standard Brownian bridge.² The limit process is mean zero and has covariance function $\text{Cov}[G_F(u), G_F(v)] = F(u) \wedge F(v) - F(u)F(v)$. The $D[s, t]$ denotes the Skorohod space of all cadlag functions on the interval $[s, t] \subset \mathbb{R}$.

Assumption 25. Consider the K -vector coefficient function $\beta: (0, 1) \rightarrow \mathbb{R}^K$. Assume each of its element function $\beta_k: (0, 1) \rightarrow \mathbb{R}$, $k = 1, \dots, K$, is differentiable with uniformly continuous and bounded derivative.

Assumption 25 is a regularity condition that assumes the differentiability of the coefficient function. Applying the fundamental property in Lemma 23 to a standard

²The standard Brownian bridge process G on the unit interval $[0, 1]$ is mean zero with covariance function $\text{Cov}[G(u), G(v)] = u \wedge v - uv$, where $u \wedge v$ denotes $\min\{u, v\}$.

uniform distribution function, we get the uniform asymptotic normality of the coefficient function estimator in the simplified model.

Theorem 15 (Uniform asymptotic normality of coefficient function estimator in simplified model). *Under Assumptions 14–18 and 25 in the simplified model, for any element coefficient function $k = 1, \dots, K$, we have the uniform asymptotic normality of the coefficient function estimator*

$$\rho_{\bar{n}}(\hat{\beta}_k(\cdot) - \beta_k(\cdot)) \xrightarrow{d} \beta_k^0 G, \text{ in } \ell^1(0, 1), \text{ as } n \rightarrow \infty.$$

The limit process is Gaussian with mean zero and covariance function $\beta_k^0(u)\beta_k^0(v)(u \wedge v - uv)$, $u, v \in (0, 1)$.

The symbol $\ell^1(T)$ denotes the collection of all bounded functions $f: T \rightarrow \mathbb{R}$ with the norm $\|f\|_1 = \sup_{t \in T} |f(t)|$.

Corollary 16 applies Theorem 15 at a fixed τ to get the pointwise asymptotic normality of the coefficient estimator at any specific rank.

Corollary 16 (Pointwise asymptotic normality of coefficient estimator in simplified model). *Under Assumptions 14–18 and 25 in the simplified model, for any $\tau \in (0, 1)$,*

$$\rho_{\bar{n}}(\hat{\beta}_k(\tau) - \beta_k(\tau)) \xrightarrow{d} \beta_k^0(\tau)G(\tau), \text{ as } n \rightarrow \infty,$$

where $\beta_k^0(\tau)G(\tau)$ is a Gaussian random variable with mean zero and variance $\tau(1 - \tau)(\beta_k^0(\tau))^2$.

The uniform asymptotic normality of the coefficient function estimator also holds in the general model under the same n/T rate relation for its uniform consistency. In the general model, the TS-OLS individual coefficient estimates do not exactly equal to the true coefficient parameters, but the order of the rank variables can still be discovered under the rate condition. Under the same rate condition, the TS-OLS estimation error goes away even after scaling by $\rho_{\bar{n}}$. Thus, the coefficient function estimator has the same limit as in the simplified model.

Theorem 17 (Uniform asymptotic normality in general model). *Under Assumptions 14–17, 19–23, and 25 in the general model, as $n, T \rightarrow \infty$ with rate relation $n = o(T^{\kappa/(3+\delta)})$, the coefficient function estimator converges to a Gaussian limit,*

$$\sqrt{n}(\hat{\beta}_k(\cdot) - \beta_k(\cdot)) \rightarrow \beta_k^0 G, \text{ in } \ell^1(0, 1),$$

where the limit is mean zero and has covariance function $\beta_k^0(u)\beta_k^0(v)(u \wedge v - uv)$, $u, v \in (0, 1)$.

3.6.2 Bootstrap in the simplified model

In this subsection, I first show bootstrap consistency, which provides the theoretic grounds for the bootstrap method suggested later. Then, I construct a bootstrap confidence interval for the coefficient at any specific rank value. Finally, I present a cluster-robust bootstrap algorithm.

Generally, let P_n be the empirical measure for a size n iid sample X_1, X_2, \dots, X_n from probability measure P :

$$P_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i},$$

where δ_x is the Dirac measure, whose mass is concentrated at a point x on the domain of Borel sets in \mathbb{R} . This P_n is the probability measure for the empirical distribution function F_n ,

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{X_i \leq x\}}.$$

Let F be the set of indicator functions $F = \{\mathbf{1}_{(-\infty, t]} : t \in \mathbb{R}\}$. It is known that each such indicator function is measurable and that F is a Donsker class for any underlying measure P of X_1, \dots, X_n . Donsker's theorem can alternatively be written as

$$\sqrt{n}(P_n - P) \rightarrow H, \text{ in } \ell^1(F), \text{ as } n \rightarrow \infty, \quad (3.36)$$

where the limit process H is the same as $G \circ F$ previously, with G being the standard

Brownian bridge.

Consider the standard bootstrap method. Given a sample X_1, X_2, \dots, X_n , the bootstrap sample $\hat{X}_1, \hat{X}_2, \dots, \hat{X}_n$ is an iid sample drawn from the original sample X_1, X_2, \dots, X_n with replacement. It treats the empirical measure P_n as the true measure for bootstrap resampling. Let \hat{P}_n denote the bootstrap empirical measure,

$$\hat{P}_n = \frac{1}{n} \sum_{i=1}^n M_{ni} \delta_{X_i}, \quad (3.37)$$

where M_{ni} denotes the number of times X_i is redrawn in the size n bootstrap sample. Conditional on the data, the random vector (M_{n1}, \dots, M_{nn}) follows a multinomial distribution with parameters n and $(1/n, \dots, 1/n)$.

The bootstrap empirical process is

$$\hat{G}_n = \sqrt{n}(\hat{P}_n - P_n).$$

It is well known that the bootstrap empirical process converges to the same limit process as in Donsker's theorem.

Lemma 18 (The bootstrap empirical process convergence, as a special case of van der Vaart and Wellner (1996) Theorem 3.6.13). *Conditional on almost all data sequences X_1, X_2, \dots , the bootstrap empirical process converges to a limit process,*

$$\hat{G}_n \Rightarrow G, \text{ in } \ell^1(F), \text{ as } n \rightarrow \infty,$$

where the limit is the same as in (3.36).

Bootstrap consistency

Definition 1 (Validity of bootstrap). A bootstrap method is “valid” if the following two sufficient conditions are met. First, the centered and scaled estimator asymptotically

converges to some distribution,

$$\rho_{\bar{n}}(\hat{\theta} - \theta_0) \stackrel{d}{\rightarrow} N(0, V), \text{ as } n \rightarrow \infty. \quad (\text{I})$$

Second, conditional on almost all sequences X_1, X_2, \dots drawn iid from the true CDF F , the centered and scaled bootstrap estimator asymptotically converges to the same limiting distribution,

$$\rho_{\bar{n}}(\hat{\theta} - \hat{\theta}_j(X_1, \dots, X_n)) \stackrel{d}{\rightarrow} N(0, V), \text{ as } n \rightarrow \infty, \quad (\text{II})$$

where θ_0 is the true parameter, $\hat{\theta}$ is the estimator for θ_0 from the sample, and $\hat{\theta}_j$ is the bootstrap estimator.

To show bootstrap validity for the coefficient estimator in this paper, I show the first condition is met by Theorem 15. Applying the delta-method to Lemma 18, I show the second condition is met.

Lemma 19 (Bootstrap consistency). *Let F be a distribution function with compact support $[s, t]$ and be continuously differentiable on its support with strictly positive derivative f . Let F_n be the empirical distribution of an iid sample from F with size n . Let \hat{F}_n be the bootstrap empirical distribution of an iid sample from F_n with size n . Let $\beta_k(\cdot)$ denote the k th element of the vector-valued coefficient function $\beta(\cdot) = (\beta_1(\cdot), \dots, \beta_k(\cdot), \dots, \beta_K(\cdot))$. Under Assumption 25, for all $k = 1, \dots, K$, conditional on almost all sequences X_1, X_2, \dots , the bootstrap empirical process converges to a Gaussian process,*

$$\rho_{\bar{n}}(\beta_k(\hat{F}_n^{-1}) - \beta_k(F_n^{-1})) \rightarrow \beta_k^0(F^{-1}) \frac{G}{f(F^{-1})}, \text{ in } \ell^1(0, 1), \text{ as } n \rightarrow \infty, \quad (3.38)$$

where the limit process is the same as in Lemma 23.

Let $\hat{\beta}_k(\cdot)$ denote the coefficient estimator as a function of the rank value τ uniformly over $(0, 1)$ from the bootstrapped sample. Let $\hat{\beta}_k(\cdot)$ denote the coefficient estimator as a function of rank value τ uniformly over $(0, 1)$ from the original sample. Applying

Lemma 19 to the standard uniform distribution function, we have the conditional convergence of the bootstrap coefficient function estimator uniformly over all τ on $(0, 1)$ in the simplified model.

Theorem 20 (Bootstrap consistency in simplified model). *Under Assumptions 14–18 and 25 in the simplified model,*

$$\rho_{\bar{n}}(\hat{\beta}_k(\cdot) - \beta_k(\cdot)) \overset{D}{\rightarrow} j(X_1, \dots, X_n) \beta_k^0 G, \text{ in } \ell^1(0, 1), \text{ as } n \rightarrow \infty,$$

where the limit process is the same as in Theorem 15.

Theorem 20 establishes bootstrap consistency in the simplified model. This provides a suggestive ground for using bootstrap methods in the general model in practice.

Bootstrap confidence interval

In this section, I compute a bootstrap standard error and construct a bootstrap confidence interval for the coefficient estimator at any rank value.

For any fixed rank value τ , we can obtain the coefficient estimates $\hat{\beta}_k(\tau)$, denoted as $\hat{\beta}_\tau^k$, as in Section 3.4.2. For simplicity of notation, I suppress the τ and k , using $\hat{\beta}$ to denote the coefficient estimates at rank value τ for the k th element of the coefficient function.

I use Algorithm 3 of Chernozhukov, Fernández-Val, and Melly (2013) to estimate the standard error based on the bootstrap interquartile range,

$$\widehat{\text{SE}}_{bs} = \frac{\hat{\beta}_{0.75} - \hat{\beta}_{0.25}}{z_{0.75} - z_{0.25}}, \quad (3.39)$$

where z_τ denote the τ -quantile of a standard normal distribution, and $\hat{\beta}_{0.75}$ and $\hat{\beta}_{0.25}$ are the 0.75- and 0.25-quantile of the bootstrapped estimators f^b , $b = 1, \dots, Bg$ among B bootstrap replications. The bootstrap standard deviation is $\hat{\sigma}_{bs} = \rho_{\bar{n}} \widehat{\text{SE}}_{bs}$. As $n \rightarrow \infty$, the bootstrapped standard deviation converges in probability to the real-world standard

deviation of the ρ_n -scaled estimator, i.e., $\hat{\sigma}_{bs} \stackrel{p}{\rightarrow} \sigma$.

Construct the bootstrap confidence interval as

$$[\hat{\beta} - CV_{1-\alpha}^{bs} \widehat{SE}_{bs}, \hat{\beta} + CV_{1-\alpha}^{bs} \widehat{SE}_{bs}], \quad (3.40)$$

where $\hat{\beta}$ is the estimator from the given dataset, the standard error estimator \widehat{SE}_{bs} is derived from (3.39), and the bootstrap critical value $CV_{1-\alpha}^{bs}$ is defined as the $1 - \alpha$ quantile of the $j\hat{\beta}^b - \hat{\beta}j/\widehat{SE}_{bs}$ among the B bootstrap replications,

$$CV_{1-\alpha}^{bs} = (1 - \alpha)\text{-quantile of } \{j\hat{\beta}^b - \hat{\beta}j/\widehat{SE}_{bs}, b = 1, \dots, B\}. \quad (3.41)$$

The bootstrap confidence interval has asymptotic coverage probability $1 - \alpha$. From bootstrap consistency, $\rho_n(\hat{\beta}^b - \hat{\beta})$ converges in distribution to the same limit as $\rho_n(\hat{\beta} - \beta)$, conditional on almost all sequences of data. Further, $\hat{\sigma}_{bs} \stackrel{p}{\rightarrow} \sigma_{bs}$. Thus,

$$\frac{j\hat{\beta}^b - \hat{\beta}j}{\widehat{SE}_{bs}} = \frac{\rho_n(\hat{\beta}^b - \hat{\beta})}{\hat{\sigma}_{bs}}$$

converges to the same limit distribution as $\rho_n(\hat{\beta} - \beta)/\sigma_{bs}$. Based on the uniform asymptotic normality showed in Section 3.6.1, $\rho_n(\hat{\beta} - \beta)/\sigma_{bs}$ converges to the standard normal distribution. Therefore, the bootstrap critical value $CV_{1-\alpha}^{bs}$ asymptotically converges to the $1 - \alpha$ quantile of the absolute value of the standard normal distribution. The bootstrap confidence interval $[\hat{\beta} - CV_{1-\alpha}^{bs} \widehat{SE}_{bs}, \hat{\beta} + CV_{1-\alpha}^{bs} \widehat{SE}_{bs}]$ asymptotically converges to the interval $[\hat{\beta} - CV_{1-\alpha}^{N(0,1)} SE_{\hat{\beta}}, \hat{\beta} + CV_{1-\alpha}^{N(0,1)} SE_{\hat{\beta}}]$.

Bootstrap algorithm:

I draw the bootstrap sample using the panel nonparametric bootstrap method suggested in Hansen (2020). I randomly draw individuals with replacement and keep the corresponding time series for each individual, keeping the time series dependence structure. This procedure essentially comes from the clustered bootstrap method in Cameron, Gel-

bach, and Miller (2008). It is also a special case of the two-step clustered bootstrapping method in Davison and Hinkley (1997). Davison and Hinkley (1997) suggest drawing individuals with replacement in the first step and resampling the time series data either with or without replacement in the second step. They show that the second-stage drawing without replacement is better than that with replacement.

Given any $\tau \in (0, 1)$, we can construct the bootstrap confidence interval for the true coefficient β at this rank value τ .³ The bootstrap method algorithm is as follows.

Step 1: Given a panel dataset $f(Y_{it}, \mathbf{X}_{it})$, $i = 1, \dots, n$, $t = 1, \dots, T$, compute the original estimator $\hat{\beta}$.

Step 2: Draw the bootstrap sample $\{(Y_{it}^b, \mathbf{X}_{it}^b), i = 1, \dots, n, t = 1, \dots, T\}$. Use the bootstrapped sample to compute the bootstrap estimator $\hat{\beta}^b$.

Step 3: Repeat the last step, bootstrapping for $b = 1, \dots, B$.

Step 4: Compute bootstrap standard error $\widehat{SE}_{bs} = \frac{\hat{\beta}_{0.75} - \hat{\beta}_{0.25}}{z_{0.75} - z_{0.25}}$.

Step 5: Compute the bootstrap critical value $CV_{1-\alpha}^{bs}$ as the $(1 - \alpha)$ -quantile of the Studentized $\left\{ \frac{\hat{\beta}^b - \hat{\beta}}{\widehat{SE}_{bs}}, b = 1, \dots, B \right\}$.

Step 6: Compute the bootstrap confidence interval $[\hat{\beta} - CV_{1-\alpha}^{bs} \widehat{SE}_{bs}, \hat{\beta} + CV_{1-\alpha}^{bs} \widehat{SE}_{bs}]$.

3.7 Empirical illustration

To apply the proposed method to an empirical example, I study the causal effect of a country's oil wealth on its military defense burden using the data from Cotet and Tsui (2013). There are many studies debating the effect of oil wealth on political conflict. On one hand, resource scarcity triggers conflict, so resource abundance should mitigate regime instability. On the other hand, political instability in the Middle East and some

³For simplicity of notation, I suppress the τ and denote the estimator as $\hat{\beta}$.

developing countries with rich oil resources makes people arrive at the “oil-fuels-war” conclusion. The political violence in a regime can be measured in multiple ways, for example, the onset of civil war. Additionally, political conflict may not necessarily lead to regime shift or war. Military defense burden is also considered as a barrier to political conflict. My empirical example studies the effect of oil wealth on military defense spending.

First, I provide point estimates of the effect of oil wealth on military defense spending (as a ratio of GDP) for countries at various rank levels τ . The rank variable represents a country’s unobserved general propensity of having military spending. Second, I present bootstrap confidence intervals for the effects. Finally, I show joint confidence sets for the oil wealth effect as a function of unobserved military spending propensity using the Dirichlet method.

Cotet and Tsui (2013) use the standard fixed-effect method to study the effect of oil wealth on political violence. They provide several measures of oil wealth, such as oil reserves or oil discoveries, and several measures related to political violence, such as the onset of civil war or the military defense spending as a ratio of GDP. They find that after controlling for the country fixed effect, oil wealth has no effect on civil conflict onset or military spending. However, the effect is significant among the subsample of non-democratic countries. The estimated effect has opposite signs between the democratic and non-democratic subsamples. Instead of effect heterogeneity varying by the observable democracy variable, I investigate effect heterogeneity across the unobserved military spending rank variable.

Consider the structural model

$$\log(DEFENSE_{it}) = \beta(U_i) \log(OILWEALTH_{it}) + \mathbf{X}_{it}^\theta(U_i) + V_{it}. \quad (3.42)$$

The outcome variable is measured as (the log of the percentage points of) the ratio

of military defense spending to GDP. The variable of interest is $\log(OILWEALTH_{it})$, the log of dollar-valued inflation-adjusted oil wealth per capita. The parameter $\beta(U_i)$ is the causal effect of oil wealth on military spending. The unobserved country rank variable U_i denotes country i 's general propensity to have military spending. The effect can differ across countries with different U_i . The covariate vector \mathbf{X}_{it} contains country characteristics including economic growth and population. V_{it} is the idiosyncratic error, which is assumed exogenous so that TS-OLS is consistent. The goal is to estimate $\beta(\tau)$ for various τ and to provide inference on the coefficient function $\beta(\cdot)$.

The data is from Cotet and Tsui (2013). My sample includes 45 countries over years 1988–2003.⁴ The military defense burden outcome variable is measured as a log of percentage points. For example, if a country's military spending is 3.2% of its GDP, then its outcome variable value is $\log(3.2)$. Oil wealth is measured as log oil value per capita.⁵ For example, a country with 6 million dollars per capita oil reserve wealth is counted as $\log(6,000,000) = 6.8$ as its $\log(OILWEALTH_{it})$ value. The effect $\beta(U_i)$ measures the elasticity given a 1% increase in oil wealth per capita, i.e., the percent effect on the percentage points of the military spending to GDP ratio.

The two other covariates included in my application are the country's economic growth and population density. Economic growth is measured as the country's annual GDP growth rate. For example, if a country's annual GDP growth is 3%, then the value of economic growth is 0.03. The population density is measured by the log of the country's total population.⁶

Table C.2 presents the coefficient estimates at rank values from $\tau = 0.1$ to 0.9

⁴I clean out observations with missing values on the outcome variable and explanatory variables as well as "stayers" according to Assumption 18. The original dataset has 222 countries from 1920–2008. My 45 countries are listed in Appendix C.2.3 with their time-averaged democracy levels. The sample used in this section is different from that used in Cotet and Tsui (2013) and does not represent the world. However, the sample selection issue is beyond the scope of this paper and left to future work.

⁵Cotet and Tsui (2013) instead use the log of oil wealth per capita *divided by* 100. My coefficient estimates can be multiplied by 100 to compare directly with those of Cotet and Tsui (2013).

⁶Cotet and Tsui (2013) use log of the country's population divided by 100 as the population density.

and their bootstrap standard errors, bootstrap critical values, and bootstrap confidence intervals. At $\tau = 0.5$, the coefficient is -0.18 , which means for a country who is at the median propensity for military spending, a 1% increase of oil wealth per capita causes a 0.18% decrease in the percentage points of the military spending as a ratio of GDP. The effect is not statistically significant at all rank levels.

Table C.2 shows the estimated coefficient function is nonlinear. For the countries who are very likely to have political violence (high τ), if they have extra oil discoveries, it will decrease their military spending ratio, meaning decreasing the military power to prevent civil war or political violence. For the countries who are least prone to military spending (low τ), having extra oil wealth increases their military spending ratio, meaning increasing the barrier of civil war. However, the overall pattern of the coefficient function is not very clear in this example. This is expected due to the limited number of observations, which causes large standard errors and imprecise estimation.

The resampling-based bootstrap critical value and bootstrap standard error converge to their true bootstrap-world limits as the number of bootstrap replications B increases. Table C.2 reports their limit values (i.e., for large B) at all rank values. The bootstrap confidence interval is wider than the traditional confidence interval, whose derivation is based on the standard normal distribution. The asymptotic normal distribution might be a poor approximation to capture the finite-sample distribution of the estimator. This helps illustrate the importance of using a bootstrap critical value instead of the traditional critical value 1.64 for building 90% confidence intervals.

Figure 3.4 illustrates the joint confidence set for the 45 countries' causal effects of oil wealth on military defense spending using the Dirichlet method. Each dot in this graph denotes a point estimate for one country. The vertical value of each dot is the country's estimated effect of oil wealth on military spending. I label the highest (0.75) and lowest (-1.11) estimates. The horizontal value of each dot is its corresponding estimated rank

Table 3.1: Empirical results for effect of oil wealth on military defense spending.

τ	Slope Estimates	Bootstrap S.E.	Bootstrap Critical Value	90% Bootstrap CI	
				Lower Bound	Upper Bound
0.1	0.18	0.13	4.41	0.38	0.75
0.2	0.43	0.32	2.13	1.11	0.26
0.3	0.06	0.23	2.15	0.43	0.54
0.4	0.14	0.20	1.91	0.53	0.25
0.5	0.18	0.25	1.72	0.60	0.25
0.6	0.12	0.22	2.39	0.39	0.64
0.7	0.19	0.22	1.99	0.64	0.25
0.8	0.01	0.24	1.82	0.44	0.42
0.9	1.01	0.40	3.68	2.47	0.46
OLS	0.04	0.04	1.64	0.02	0.11

900 bootstrap replications.

variable value. Each dot is contained in a rectangle, which denotes the confidence region for each country's coefficient estimate. The n rectangles form a joint confidence set for all 45 points on the function $\beta(\cdot)$, using a 90% confidence level.

Figure 3.4 shows that most of the confidence rectangles are relatively informative. However, for a few, the vertical height is very large. This is because the TS-OLS standard error estimates for the corresponding countries are very big, partly from the very limited time dimension ($T = 15$) in this example. With more time series observations, the standard errors would decrease, and the confidence rectangles would become smaller and more precise. In future work, I will pursue grouping similar individuals to reduce the TS-OLS standard error and improve inference.

3.8 Conclusion

In this paper, I propose a structural quantile-based random coefficient panel model with the goal of discovering heterogeneous causal effects while accounting for endogeneity. I combine the idea of quantile regression and random coefficient models by introducing

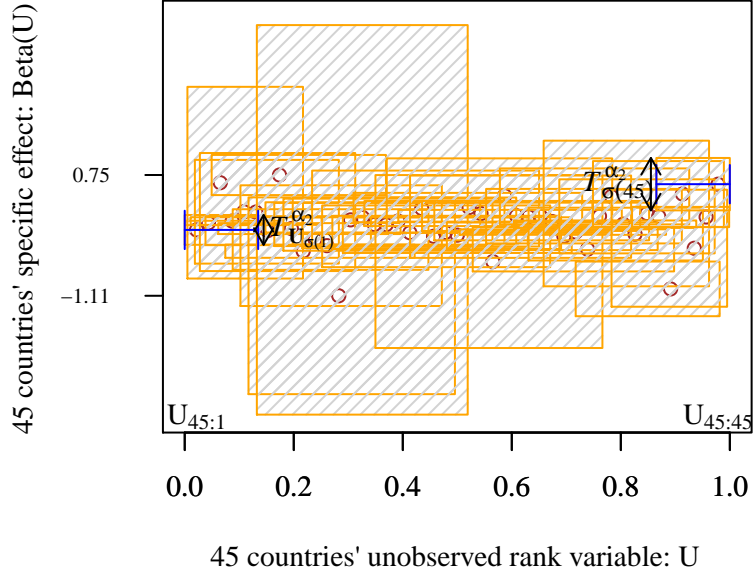


Figure 3.4: Joint confidence set for $\beta(\cdot)$.

a time-invariant rank variable and considering the coefficient vector as a function of the rank variable. I develop several theoretical results in both a simplified model and a general long-panel model. First, I establish identification and uniformly consistent estimation based on a sorting method. Second, I propose a novel inference method, building confidence sets or uniform confidence bands for the coefficient function based on a Dirichlet approach. Third, I establish the uniform asymptotic normality of the coefficient estimator and its bootstrap consistency. This facilitates a bootstrap confidence interval for the coefficient estimator at each specific rank.

In future work, I plan to study multiple extensions. First, I consider the partial identification. This could occur in a few cases, such as when T is less than K , or some individuals have missing values or lack variation in their time series data. In the case that we cannot obtain point identification, a worst-case bound of the coefficient estimator might help. Second, when the dataset lacks variation, I consider grouping similar individuals to create more variation, hoping to choose groups in an optimal way.

Appendix A

Appendix for Chapter 1

A.1 Proofs

Proof of Theorem 1. We show that the two conditions of Theorem 5.7 in van der Vaart (1998a) are satisfied. The first condition of Theorem 5.7 in van der Vaart (1998a) requires

$$\sup_{\beta \in B} |\hat{\mathbf{M}}_n(\beta, \tau) - \hat{\mathbf{W}} \hat{\mathbf{M}}_n(\beta, \tau) - \mathbf{M}(\beta, \tau) + \mathbf{W} \mathbf{M}(\beta, \tau)| \stackrel{p}{\rightarrow} 0. \quad (\text{A.1})$$

From the proof of Theorem 2.3 in de Castro, Galvao, and Kaplan (2017), we know that $\sup_{\beta \in B} \|\hat{\mathbf{M}}_n(\beta, \tau) - \mathbf{M}(\beta, \tau)\| = o_p(1)$. From Assumption 12, $\hat{\mathbf{W}} = \mathbf{W} + o_p(1)$, which does not depend on β .

Let $\|\cdot\|$ denote the standard Euclidean vector norm. For any matrix \mathbf{A} , denote the matrix norm as $\|\mathbf{A}\| = \|\mathbf{A}^{\top}\| = \sqrt{\text{tr}(\mathbf{A} \mathbf{A}^{\top})}$, which is the Euclidean vector norm if \mathbf{A} is a vector. Given this norm, the Cauchy–Schwarz inequality states that for any matrices \mathbf{A} and \mathbf{B} , $\|\mathbf{A} \mathbf{B}\| \leq \|\mathbf{A}\| \|\mathbf{B}\|$.

We now use the triangle inequality, Cauchy–Schwarz inequality, uniform convergence in probability of $\hat{\mathbf{M}}_n(\beta, \tau)$, and convergence in probability of $\hat{\mathbf{W}}$, to show the required

condition in (A.1):

$$\begin{aligned}
& \sup_{\beta 2B} |\hat{\mathbf{M}}_n(\cdot, \tau) - \hat{\mathbf{W}} \hat{\mathbf{M}}_n(\cdot, \tau) - \mathbf{M}(\cdot, \tau) + \mathbf{W} \mathbf{M}(\cdot, \tau)| \\
&= \sup_{\beta 2B} |(\mathbf{M}(\cdot, \tau) + (\hat{\mathbf{M}}(\cdot, \tau) - \mathbf{M}(\cdot, \tau))) - (\mathbf{W} + (\hat{\mathbf{W}} - \mathbf{W})) \\
&\quad (\mathbf{M}(\cdot, \tau) + (\hat{\mathbf{M}}(\cdot, \tau) - \mathbf{M}(\cdot, \tau))) \\
&\quad - \mathbf{M}(\cdot, \tau) + \mathbf{W} \mathbf{M}(\cdot, \tau)| \\
&= \sup_{\beta 2B} |\mathbf{M}(\cdot, \tau) - (\hat{\mathbf{W}} - \mathbf{W}) \mathbf{M}(\cdot, \tau) + \mathbf{M}(\cdot, \tau) - \mathbf{W} (\hat{\mathbf{M}}(\cdot, \tau) - \mathbf{M}(\cdot, \tau)) \\
&\quad + (\hat{\mathbf{M}}(\cdot, \tau) - \mathbf{M}(\cdot, \tau)) - \mathbf{W} \mathbf{M}(\cdot, \tau) + \mathbf{M}(\cdot, \tau) - (\hat{\mathbf{W}} - \mathbf{W}) (\hat{\mathbf{M}}(\cdot, \tau) - \mathbf{M}(\cdot, \tau)) \\
&\quad + (\hat{\mathbf{M}}(\cdot, \tau) - \mathbf{M}(\cdot, \tau)) - \mathbf{W} (\hat{\mathbf{M}}(\cdot, \tau) - \mathbf{M}(\cdot, \tau)) \\
&\quad + (\hat{\mathbf{M}}(\cdot, \tau) - \mathbf{M}(\cdot, \tau)) - (\hat{\mathbf{W}} - \mathbf{W}) \mathbf{M}(\cdot, \tau) \\
&\quad + (\hat{\mathbf{M}}(\cdot, \tau) - \mathbf{M}(\cdot, \tau)) - (\hat{\mathbf{W}} - \mathbf{W}) (\hat{\mathbf{M}}(\cdot, \tau) - \mathbf{M}(\cdot, \tau))| \\
&\leq \sup_{\beta 2B} |\mathbf{M}(\cdot, \tau) - (\hat{\mathbf{W}} - \mathbf{W}) \mathbf{M}(\cdot, \tau)|_j + \sup_{\beta 2B} |\mathbf{M}(\cdot, \tau) - \mathbf{W} (\hat{\mathbf{M}}(\cdot, \tau) - \mathbf{M}(\cdot, \tau))|_j \\
&\quad + \sup_{\beta 2B} |(\hat{\mathbf{M}}(\cdot, \tau) - \mathbf{M}(\cdot, \tau)) - \mathbf{W} \mathbf{M}(\cdot, \tau)|_j + \sup_{\beta 2B} |(\hat{\mathbf{W}} - \mathbf{W}) (\hat{\mathbf{M}}(\cdot, \tau) - \mathbf{M}(\cdot, \tau))|_j \\
&\quad + \sup_{\beta 2B} |(\hat{\mathbf{M}}(\cdot, \tau) - \mathbf{M}(\cdot, \tau)) - \mathbf{W} (\hat{\mathbf{M}}(\cdot, \tau) - \mathbf{M}(\cdot, \tau))|_j \\
&\quad + \sup_{\beta 2B} |(\hat{\mathbf{M}}(\cdot, \tau) - \mathbf{M}(\cdot, \tau)) - (\hat{\mathbf{W}} - \mathbf{W}) \mathbf{M}(\cdot, \tau)|_j \\
&\quad + \sup_{\beta 2B} |(\hat{\mathbf{M}}(\cdot, \tau) - \mathbf{M}(\cdot, \tau)) - (\hat{\mathbf{W}} - \mathbf{W}) (\hat{\mathbf{M}}(\cdot, \tau) - \mathbf{M}(\cdot, \tau))|_j \\
&= \sup_{\beta 2B} k |\mathbf{M}(\cdot, \tau) - (\hat{\mathbf{W}} - \mathbf{W}) \mathbf{M}(\cdot, \tau)|_k + \sup_{\beta 2B} k |\mathbf{M}(\cdot, \tau) - \mathbf{W} (\hat{\mathbf{M}}(\cdot, \tau) - \mathbf{M}(\cdot, \tau))|_k \\
&\quad + \sup_{\beta 2B} k |(\hat{\mathbf{M}}(\cdot, \tau) - \mathbf{M}(\cdot, \tau)) - \mathbf{W} \mathbf{M}(\cdot, \tau)|_k \\
&\quad + \sup_{\beta 2B} k |\mathbf{M}(\cdot, \tau) - (\hat{\mathbf{W}} - \mathbf{W}) (\hat{\mathbf{M}}(\cdot, \tau) - \mathbf{M}(\cdot, \tau))|_k \\
&\quad + \sup_{\beta 2B} k |(\hat{\mathbf{M}}(\cdot, \tau) - \mathbf{M}(\cdot, \tau)) - \mathbf{W} (\hat{\mathbf{M}}(\cdot, \tau) - \mathbf{M}(\cdot, \tau))|_k \\
&\quad + \sup_{\beta 2B} k |(\hat{\mathbf{M}}(\cdot, \tau) - \mathbf{M}(\cdot, \tau)) - (\hat{\mathbf{W}} - \mathbf{W}) \mathbf{M}(\cdot, \tau)|_k
\end{aligned}$$

$$\begin{aligned}
& + \sup_{\beta, 2B} k(\hat{\mathbf{M}}(\cdot, \tau) - \mathbf{M}(\cdot, \tau))^{\top} (\hat{\mathbf{W}} - \mathbf{W})(\hat{\mathbf{M}}(\cdot, \tau) - \mathbf{M}(\cdot, \tau))k \\
& \quad \text{by Cauchy-Schwarz inequality} \\
& \sup_{\beta, 2B} k\mathbf{M}(\cdot, \tau)^{\top} k k(\hat{\mathbf{W}} - \mathbf{W})k k\mathbf{M}(\cdot, \tau)k + \sup_{\beta, 2B} k\mathbf{M}(\cdot, \tau)^{\top} k k\mathbf{W}k k(\hat{\mathbf{M}}(\cdot, \tau) - \mathbf{M}(\cdot, \tau))k \\
& + \sup_{\beta, 2B} k(\hat{\mathbf{M}}(\cdot, \tau) - \mathbf{M}(\cdot, \tau))^{\top} k k\mathbf{W}k k\mathbf{M}(\cdot, \tau)k \\
& + \sup_{\beta, 2B} k\mathbf{M}(\cdot, \tau)^{\top} k k(\hat{\mathbf{W}} - \mathbf{W})k k(\hat{\mathbf{M}}(\cdot, \tau) - \mathbf{M}(\cdot, \tau))k \\
& + \sup_{\beta, 2B} k(\hat{\mathbf{M}}(\cdot, \tau) - \mathbf{M}(\cdot, \tau))^{\top} k k\mathbf{W}k k(\hat{\mathbf{M}}(\cdot, \tau) - \mathbf{M}(\cdot, \tau))k \\
& + \sup_{\beta, 2B} k(\hat{\mathbf{M}}(\cdot, \tau) - \mathbf{M}(\cdot, \tau))^{\top} k k(\hat{\mathbf{W}} - \mathbf{W})k k\mathbf{M}(\cdot, \tau)k \\
& + \sup_{\beta, 2B} k(\hat{\mathbf{M}}(\cdot, \tau) - \mathbf{M}(\cdot, \tau))^{\top} k k(\hat{\mathbf{W}} - \mathbf{W})k k(\hat{\mathbf{M}}(\cdot, \tau) - \mathbf{M}(\cdot, \tau))k \\
& \sup_{\beta, 2B} k\mathbf{M}(\cdot, \tau)^{\top} k k(\hat{\mathbf{W}} - \mathbf{W})k \overbrace{\sup_{\beta, 2B} k\mathbf{M}(\cdot, \tau)k}^{=O(1)} \\
& + \sup_{\beta, 2B} k\mathbf{M}(\cdot, \tau)^{\top} k k\mathbf{W}k \sup_{\beta, 2B} k(\hat{\mathbf{M}}(\cdot, \tau) - \mathbf{M}(\cdot, \tau))k \\
& + \sup_{\beta, 2B} k(\hat{\mathbf{M}}(\cdot, \tau) - \mathbf{M}(\cdot, \tau))^{\top} k \overbrace{k\mathbf{W}k}^{=O(1)} \sup_{\beta, 2B} k\mathbf{M}(\cdot, \tau)k \\
& + \sup_{\beta, 2B} k\mathbf{M}(\cdot, \tau)^{\top} k k(\hat{\mathbf{W}} - \mathbf{W})k \sup_{\beta, 2B} k(\hat{\mathbf{M}}(\cdot, \tau) - \mathbf{M}(\cdot, \tau))k \\
& + \sup_{\beta, 2B} k(\hat{\mathbf{M}}(\cdot, \tau) - \mathbf{M}(\cdot, \tau))^{\top} k k\mathbf{W}k \sup_{\beta, 2B} k(\hat{\mathbf{M}}(\cdot, \tau) - \mathbf{M}(\cdot, \tau))k \\
& + \sup_{\beta, 2B} k(\hat{\mathbf{M}}(\cdot, \tau) - \mathbf{M}(\cdot, \tau))^{\top} k k(\hat{\mathbf{W}} - \mathbf{W})k \sup_{\beta, 2B} k\mathbf{M}(\cdot, \tau)k \\
& + \sup_{\beta, 2B} k(\hat{\mathbf{M}}(\cdot, \tau) - \mathbf{M}(\cdot, \tau))^{\top} k k(\hat{\mathbf{W}} - \mathbf{W})k \sup_{\beta, 2B} k(\hat{\mathbf{M}}(\cdot, \tau) - \mathbf{M}(\cdot, \tau))k \\
& = o_p(1) + o_p(1) + o_p(1) + o_p(1) + o_p(1) + o_p(1) + o_p(1) = o_p(1).
\end{aligned}$$

Above, we know $k\mathbf{W}k = O(1)$ since \mathbf{W} is fixed, and we know $\sup_{\beta, 2B} k\mathbf{M}(\cdot, \tau)k = O(1)$ since $\mathbf{M}(\cdot, \tau)$ is continuous in β and B is a compact set.

The second condition of Theorem 5.7 in van der Vaart (1998a) is that $\tau_{0\tau}$ satisfies the well-separated minimum property. Since $\mathbf{M}(\cdot, \tau)^{\top} \mathbf{W} \mathbf{M}(\cdot, \tau)$ is continuous in

and $f : k \rightarrow_{0\tau} k \in \epsilon$, \mathcal{B}_g is a compact set, let $\hat{\tau}$ denote the minimizer: for any $\epsilon > 0$,

$$\inf_{:k \rightarrow_{0\tau} k \in \epsilon} \mathbf{M}(\hat{\tau}) \mathbf{W} \mathbf{M}(\hat{\tau}) = \mathbf{M}(\hat{\tau}) \mathbf{W} \mathbf{M}(\hat{\tau}). \quad (\text{A.2})$$

By Assumption 3, for any $\hat{\tau} \in \mathcal{B}_g$, $\mathbf{M}(\hat{\tau}) \neq \mathbf{0}$, so $\mathbf{M}(\hat{\tau}) \neq \mathbf{0}$. Since \mathbf{W} is positive definite by Assumption 12,

$$\mathbf{M}(\hat{\tau}) \mathbf{W} \mathbf{M}(\hat{\tau}) > 0.$$

Thus, for any $\epsilon > 0$,

$$\inf_{:k \rightarrow_{0\tau} k \in \epsilon} \mathbf{M}(\hat{\tau}) \mathbf{W} \mathbf{M}(\hat{\tau}) > 0 = \mathbf{M}(\hat{\tau}) \mathbf{W} \mathbf{M}(\hat{\tau}). \quad (\text{A.3})$$

Consistency follows by Theorem 5.7 in van der Vaart (1998a). \square

Proof of Theorem 2. From the definition of $\hat{\tau}_{\text{GMM}}$ in (1.12), we have the first-order condition of the objective function:

$$\left[\nabla_{\tau} \hat{\mathbf{M}}_n(\hat{\tau}_{\text{GMM}}, \tau) \right] \mathbf{W} \hat{\mathbf{M}}_n(\hat{\tau}_{\text{GMM}}, \tau) = \mathbf{0}. \quad (\text{A.4})$$

Let $\hat{\mathbf{M}}_n^{(k)}(\hat{\tau}, \tau)$ denote the k th element of $\hat{\mathbf{M}}_n(\hat{\tau}, \tau)$, so $\nabla_{\tau} \hat{\mathbf{M}}_n^{(k)}(\hat{\tau}_{\text{GMM}}, \tau)$ is a column vector. Let

$$\underline{\mathbf{M}}_n(\tau) = \left(\nabla_{\tau} \hat{\mathbf{M}}_n^{(1)}(\tilde{\tau}_{\tau,1}, \tau), \dots, \nabla_{\tau} \hat{\mathbf{M}}_n^{(d_\beta)}(\tilde{\tau}_{\tau,d_\beta}, \tau) \right)^\top, \quad (\text{A.5})$$

where $\tilde{\tau}_{\tau,k}$ lies between $\hat{\tau}_{\text{GMM}}$ and $\hat{\tau}_{\text{GMM}}$, for all $k = 1, \dots, d_\beta$. By the mean value theorem,

$$\hat{\mathbf{M}}_n(\hat{\tau}_{\text{GMM}}, \tau) = \hat{\mathbf{M}}_n(\hat{\tau}_{\text{GMM}}, \tau) + \underline{\mathbf{M}}_n(\tau)(\hat{\tau}_{\text{GMM}} - \hat{\tau}_{\text{GMM}}). \quad (\text{A.6})$$

Pre-multiplying this by $\left[\nabla_{\tau} \hat{\mathbf{M}}_n(\hat{\tau}_{\text{GMM}}, \tau) \right] \mathbf{W}$, and using (A.4) for the first equality,

$$\begin{aligned} \mathbf{0} &= \left[\nabla_{\tau} \hat{\mathbf{M}}_n(\hat{\tau}_{\text{GMM}}, \tau) \right] \mathbf{W} \hat{\mathbf{M}}_n(\hat{\tau}_{\text{GMM}}, \tau) \\ &= \left[\nabla_{\tau} \hat{\mathbf{M}}_n(\hat{\tau}_{\text{GMM}}, \tau) \right] \mathbf{W} \hat{\mathbf{M}}_n(\hat{\tau}_{\text{GMM}}, \tau) + \left[\nabla_{\tau} \hat{\mathbf{M}}_n(\hat{\tau}_{\text{GMM}}, \tau) \right] \mathbf{W} \underline{\mathbf{M}}_n(\tau)(\hat{\tau}_{\text{GMM}} - \hat{\tau}_{\text{GMM}}). \end{aligned}$$

Multiplying by $\rho_{\hat{n}}^-$ and rearranging,

$$\begin{aligned} & \rho_{\hat{n}}^- (\hat{\mathbf{G}}_{\text{GMM}}(\tau)) \\ &= \{[\nabla_{\hat{\mathbf{M}}_n(\hat{\mathbf{G}}_{\text{GMM}}, \tau)]^{\hat{\mathbf{W}}} \hat{\mathbf{M}}_n(\tau)\}^{-1} [\nabla_{\hat{\mathbf{M}}_n(\hat{\mathbf{G}}_{\text{GMM}}, \tau)]^{\hat{\mathbf{W}}} \overbrace{\rho_{\hat{n}}^- \hat{\mathbf{M}}_n(\tau)}^{=o_p(1)} \end{aligned} \quad (\text{A.7})$$

$$= f \hat{\mathbf{G}}^{\hat{\mathbf{W}}} \hat{\mathbf{W}} \hat{\mathbf{G}} g^{-1} \hat{\mathbf{G}}^{\hat{\mathbf{W}}} \hat{\mathbf{W}} \rho_{\hat{n}}^- \hat{\mathbf{M}}_n(\tau) + o_p(1), \quad (\text{A.8})$$

where $\hat{\mathbf{W}} = \mathbf{W} + o_p(1)$ by Assumption 12, and $\hat{\mathbf{M}}_n(\tau) = \mathbf{G} + o_p(1)$ and $\nabla_{\hat{\mathbf{M}}_n(\hat{\mathbf{G}}_{\text{GMM}}, \tau)} = \mathbf{G} + o_p(1)$ by Assumption 11. From Lemma 2.4 in de Castro, Galvao, and Kaplan (2017), $\rho_{\hat{n}}^- \hat{\mathbf{M}}_n(\tau) \xrightarrow{d} \mathbf{N}(\mathbf{0}, \underline{\Sigma}_{\tau})$. Applying the continuous mapping theorem yields the result stated in the theorem. \square

Proof of Theorem 3. The proof follows the same structure as in Section 9.1 and Theorem 9.2 of Newey and McFadden (1994).

Under H_0 in (1.21), $\tau_{0\tau}$ is the unique minimizer of $\mathbf{M}(\cdot, \tau)^{\underline{\Sigma}_{\tau}^{-1}} \mathbf{M}(\cdot, \tau)$ subject to $\mathbf{a}(\cdot) = \mathbf{0}$, i.e.,

$$\tau_{0\tau} = \arg \min_{\beta \in B: \mathbf{a}(\beta) = \mathbf{0}} \mathbf{M}(\beta, \tau)^{\underline{\Sigma}_{\tau}^{-1}} \mathbf{M}(\beta, \tau). \quad (\text{A.9})$$

This follows because $\tau_{0\tau}$ is the unique *unconstrained* minimizer (by Assumption 3), and under H_0 it satisfies the constraint, too.

Next, we show that

$$\tau_{0\tau} \xrightarrow{p} \tau_{0\tau}. \quad (\text{A.10})$$

The same arguments from Theorem 1 for $\tau_{0\tau} \xrightarrow{p} \tau_{0\tau}$ apply to $\tau_{0\tau}$; the only difference is that we have the subset $\mathcal{F} : \mathbf{a}(\beta) = \mathbf{0}, \beta \in Bg$ instead of just B . Consequently,

$$\begin{aligned} & \sup_{\beta \in B: \mathbf{a}(\beta) = \mathbf{0}} |\hat{\mathbf{M}}_n(\beta, \tau)^{\hat{\mathbf{W}}} \hat{\mathbf{M}}_n(\beta, \tau) - \mathbf{M}(\beta, \tau)^{\hat{\mathbf{W}}} \mathbf{M}(\beta, \tau)| \\ & \underbrace{\sup_{\beta \in B} |\hat{\mathbf{M}}_n(\beta, \tau)^{\hat{\mathbf{W}}} \hat{\mathbf{M}}_n(\beta, \tau) - \mathbf{M}(\beta, \tau)^{\hat{\mathbf{W}}} \mathbf{M}(\beta, \tau)|}_{\text{from (A.1)}} \xrightarrow{p} 0, \end{aligned}$$

so the first condition from Theorem 5.7 in van der Vaart (1998a) is satisfied. Similarly, for the second (and final) condition,

$$\inf_{\|\tau - \tau_0\| \leq \epsilon} \mathbf{M}(\tau) - \underline{\mathbf{W}}\mathbf{M}(\tau) \underset{\text{from (A.3)}}{>} 0 = \mathbf{M}(\tau_0) - \underline{\mathbf{W}}\mathbf{M}(\tau_0),$$

so Theorem 5.7 in van der Vaart (1998a) implies $\hat{\tau}_n \xrightarrow{P} \tau_0$.

Define the Lagrangian for τ_n :

$$L_n(\tau) = (1/2) \hat{\mathbf{M}}_n(\tau)^\top \underline{\Sigma}^{-1} \hat{\mathbf{M}}_n(\tau) - \mathbf{a}(\tau)^\top \lambda_n. \quad (\text{A.11})$$

The first-order condition of the Lagrangian scaled by ρ_n^- is

$$\begin{aligned} \mathbf{0} &= [\nabla \hat{\mathbf{M}}_n(\tau_n)^\top \underline{\Sigma}^{-1} \rho_n^- \hat{\mathbf{M}}_n(\tau_n) - \nabla [\mathbf{a}(\tau_n)^\top] \rho_n^- \lambda_n, \\ \mathbf{0} &= \rho_n^- \mathbf{a}(\tau_n), \end{aligned} \quad (\text{A.12})$$

where τ_n is the constrained estimator and λ_n is the corresponding vector of Lagrange multipliers. The second equation can be approximated using a Taylor expansion of $\rho_n^- \mathbf{a}(\tau_n)$ at τ_0 :

$$\rho_n^- \mathbf{a}(\tau_n) = \rho_n^- \mathbf{a}(\tau_0) + \underline{\mathbf{A}} \rho_n^- (\tau_n - \tau_0) + o_p(1) = \mathbf{0} + \underline{\mathbf{A}} \rho_n^- (\tau_n - \tau_0) + o_p(1). \quad (\text{A.13})$$

We now simplify the first equation in (A.12) with some probability limits and expansions. From the proof of Theorem 2.3 in de Castro, Galvao, and Kaplan (2017),

$$\sup_{\mathcal{B}} \|\hat{\mathbf{M}}_n(\tau_n) - \mathbf{M}(\tau_0)\| = o_p(1). \quad (\text{A.14})$$

Consequently,

$$\begin{aligned} \|\hat{\mathbf{M}}_n(\tau_n) - \mathbf{M}(\tau_0)\| &= \underbrace{\|\hat{\mathbf{M}}_n(\tau_n) - \mathbf{M}(\tau_n)\|}_{=o_p(1) \text{ by (A.14)}} + \underbrace{\|\mathbf{M}(\tau_n) - \mathbf{M}(\tau_0)\|}_{=o_p(1) \text{ by (A.10) and } \mathbf{M}(\cdot) \text{ continuity}} \\ &= \|\hat{\mathbf{M}}_n(\tau_n) - \mathbf{M}(\tau_n)\| + \|\mathbf{M}(\tau_n) - \mathbf{M}(\tau_0)\| \end{aligned}$$

$$= o_p(1).$$

Combining this result with Assumption 11, the continuous differentiability of $\mathbf{a}(\cdot)$, and $\underline{\Sigma} \stackrel{p}{\rightarrow} \underline{\Sigma}_\tau$,

$$\left[\nabla_{\mathbf{y}} \hat{\mathbf{M}}_n(\mathbf{y}_n, \tau) \right] \stackrel{p}{\rightarrow} \underline{\Sigma}^{-1} \stackrel{p}{\rightarrow} \underline{\mathbf{G}} \stackrel{p}{\rightarrow} \underline{\Sigma}_\tau^{-1}, \quad (\text{A.15})$$

$$\nabla_{\mathbf{y}} \mathbf{a}(\mathbf{y}_n, \tau) \stackrel{p}{\rightarrow} \underline{\mathbf{A}}. \quad (\text{A.16})$$

To further simplify the first equation in (A.12), let

$$\Upsilon_n = \underline{\Sigma}_\tau^{-1/2} \rho_{\hat{\mathbf{M}}_n(\mathbf{y}_{0\tau}, \tau)}, \quad (\text{A.17})$$

so

$$\rho_{\hat{\mathbf{M}}_n(\mathbf{y}_{0\tau}, \tau)} = \underline{\Sigma}_\tau^{1/2} \Upsilon_n. \quad (\text{A.18})$$

By Lemma 2.4 in de Castro, Galvao, and Kaplan (2017), $\rho_{\hat{\mathbf{M}}_n(\mathbf{y}_{0\tau}, \tau)} \stackrel{d}{\rightarrow} \mathbf{N}(\mathbf{0}, \underline{\Sigma}_\tau)$, so

$$\Upsilon_n \stackrel{d}{\rightarrow} \Upsilon \sim \mathbf{N}(\mathbf{0}, \mathbf{I}), \quad (\text{A.19})$$

where \mathbf{I} is the $d_\beta \times d_\beta$ identity matrix. Starting from (A.8), substituting $\hat{\mathbf{W}} = \underline{\Sigma}^{-1}$ and using (A.17),

$$\rho_{\hat{\mathbf{M}}_n(\mathbf{y}_{2s}, \tau)} = \underline{\mathbf{B}}^{-1} \underline{\mathbf{G}} \stackrel{p}{\rightarrow} \underline{\Sigma}_\tau^{-1/2} \Upsilon_n + o_p(1), \quad (\text{A.20})$$

$$\underline{\mathbf{B}} = \underline{\mathbf{G}} \stackrel{p}{\rightarrow} \underline{\Sigma}_\tau^{-1} \underline{\mathbf{G}}, \quad (\text{A.21})$$

where $\underline{\mathbf{B}}$ is a positive definite, symmetric, $d_\beta \times d_\beta$ matrix. A Taylor expansion of $\hat{\mathbf{M}}_n(\mathbf{y}, \tau)$ at $\mathbf{y}_{0\tau}$ is

$$\rho_{\hat{\mathbf{M}}_n(\mathbf{y}, \tau)} = \rho_{\hat{\mathbf{M}}_n(\mathbf{y}_{0\tau}, \tau)} + \underline{\mathbf{M}}_n(\tau) \rho_{\hat{\mathbf{M}}_n(\mathbf{y}_{0\tau}, \tau)}, \quad (\text{A.22})$$

where $\underline{\mathbf{M}}_n(\tau)$ is defined the same as in (A.5), with $\tilde{\tau}_{\tau,k}$ between $\mathbf{y}_{0\tau}$ and \mathbf{y} , for all

$k = 1, \dots, d_\beta$. Using (A.18) and (A.22),

$$\begin{aligned} \rho_{\hat{\mathbf{M}}_n}^-(n, \tau) &= \rho_{\hat{\mathbf{M}}_n}^-(0_\tau, \tau) + \underline{\mathbf{M}}_n(\tau) \rho_{\hat{\mathbf{M}}_n}^-(n, 0_\tau) \\ &= \underline{\Sigma}_\tau^{1/2} \boldsymbol{\Upsilon}_n + \underline{\mathbf{M}}_n(\tau) \rho_{\hat{\mathbf{M}}_n}^-(n, 0_\tau). \end{aligned} \quad (\text{A.23})$$

Finally, the first equation in the constrained first-order condition in (A.12) becomes

$$\begin{aligned} \mathbf{0} &= \overbrace{[\nabla_{\mathbf{a}(n)} \hat{\mathbf{M}}_n(n, \tau)]^{\triangleright \underline{\Sigma}}^{-1}}^{\text{use (A.15)}} \overbrace{\rho_{\hat{\mathbf{M}}_n}^-(n, \tau)}^{\text{use (A.23)}} \overbrace{\nabla [\mathbf{a}(n)]^{\triangleright}}^{\text{use (A.16)}} \rho_{\hat{\mathbf{M}}_n}^- \\ &= \underline{\mathbf{G}}^{\triangleright \underline{\Sigma}_\tau^{-1}} (\underline{\Sigma}_\tau^{1/2} \boldsymbol{\Upsilon}_n + (\underline{\mathbf{G}} + o_p(1)) \rho_{\hat{\mathbf{M}}_n}^-(n, 0_\tau)) \\ &\quad + \underline{\mathbf{A}}^{\triangleright} \rho_{\hat{\mathbf{M}}_n}^- + o_p(1) \\ &= \underline{\mathbf{G}}^{\triangleright \underline{\Sigma}_\tau^{-1/2}} \boldsymbol{\Upsilon}_n + \underline{\mathbf{G}}^{\triangleright \underline{\Sigma}_\tau^{-1}} \underline{\mathbf{G}} \rho_{\hat{\mathbf{M}}_n}^-(n, 0_\tau) + \underline{\mathbf{A}}^{\triangleright} \rho_{\hat{\mathbf{M}}_n}^- + o_p(1). \end{aligned} \quad (\text{A.24})$$

Combining (A.13) and (A.24), the simplified constrained first-order condition can be written in matrix form:

$$\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \underline{\mathbf{G}}^{\triangleright \underline{\Sigma}_\tau^{-1/2}} \boldsymbol{\Upsilon}_n \\ \mathbf{0} \end{bmatrix} + \begin{bmatrix} \underline{\mathbf{B}} & \underline{\mathbf{A}}^{\triangleright} \\ \underline{\mathbf{A}} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \rho_{\hat{\mathbf{M}}_n}^-(n, 0_\tau) \\ \rho_{\hat{\mathbf{M}}_n}^- \end{bmatrix} + o_p(1).$$

Let

$$\underline{\mathbf{H}} \quad \underline{\mathbf{I}} \quad \underline{\mathbf{B}}^{-1/2} \underline{\mathbf{A}}^{\triangleright} (\underline{\mathbf{A}} \underline{\mathbf{B}}^{-1} \underline{\mathbf{A}}^{\triangleright})^{-1} \underline{\mathbf{A}} \underline{\mathbf{B}}^{-1/2}, \quad (\text{A.25})$$

a $d_\beta \times d_\beta$, symmetric, idempotent orthogonalization matrix of rank $d_\beta - r$. Using the formula for partitioned matrix inverse (e.g., Newey and McFadden, 1994, eqn. (9.5)),

$$\begin{aligned} \begin{bmatrix} \rho_{\hat{\mathbf{M}}_n}^-(n, 0_\tau) \\ \rho_{\hat{\mathbf{M}}_n}^- \end{bmatrix} &= \begin{bmatrix} \underline{\mathbf{B}} & \underline{\mathbf{A}}^{\triangleright} \\ \underline{\mathbf{A}} & \mathbf{0} \end{bmatrix}^{-1} \begin{bmatrix} \underline{\mathbf{G}}^{\triangleright \underline{\Sigma}_\tau^{-1/2}} \boldsymbol{\Upsilon}_n \\ \mathbf{0} \end{bmatrix} + o_p(1) \\ &= \begin{bmatrix} \underline{\mathbf{B}}^{-1/2} \underline{\mathbf{H}} \underline{\mathbf{B}}^{-1/2} & \underline{\mathbf{B}}^{-1} \underline{\mathbf{A}}^{\triangleright} (\underline{\mathbf{A}} \underline{\mathbf{B}}^{-1} \underline{\mathbf{A}}^{\triangleright})^{-1} \\ (\underline{\mathbf{A}} \underline{\mathbf{B}}^{-1} \underline{\mathbf{A}}^{\triangleright})^{-1} \underline{\mathbf{A}} \underline{\mathbf{B}}^{-1} & (\underline{\mathbf{A}} \underline{\mathbf{B}}^{-1} \underline{\mathbf{A}}^{\triangleright})^{-1} \end{bmatrix} \begin{bmatrix} \underline{\mathbf{G}}^{\triangleright \underline{\Sigma}_\tau^{-1/2}} \boldsymbol{\Upsilon}_n \\ \mathbf{0} \end{bmatrix} \\ &\quad + o_p(1) \end{aligned}$$

$$\begin{aligned}
&= \begin{bmatrix} \underline{\mathbf{B}}^{-1/2} \underline{\mathbf{H}} \underline{\mathbf{B}}^{-1/2} \underline{\mathbf{G}}^{> \Sigma_\tau}{}^{1/2} \boldsymbol{\Upsilon}_n \\ (\underline{\mathbf{A}} \underline{\mathbf{B}}^{-1} \underline{\mathbf{A}}^{>})^{-1} \underline{\mathbf{A}} \underline{\mathbf{B}}^{-1} \underline{\mathbf{G}}^{> \Sigma_\tau}{}^{1/2} \boldsymbol{\Upsilon}_n \end{bmatrix} + o_p(1) \\
&= \begin{bmatrix} \underline{\mathbf{B}}^{-1/2} \underline{\mathbf{H}} \underline{\mathbf{B}}^{-1/2} \\ (\underline{\mathbf{A}} \underline{\mathbf{B}}^{-1} \underline{\mathbf{A}}^{>})^{-1} \underline{\mathbf{A}} \underline{\mathbf{B}}^{-1} \end{bmatrix} \underline{\mathbf{G}}^{> \Sigma_\tau}{}^{1/2} \boldsymbol{\Upsilon}_n + o_p(1).
\end{aligned}$$

Then

$$\begin{aligned}
\rho_{\bar{n}}^-(\hat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}_{0\tau}) &= \underline{\mathbf{B}}^{-1/2} \underline{\mathbf{H}} \underline{\mathbf{B}}^{-1/2} \underline{\mathbf{G}}^{> \Sigma_\tau}{}^{1/2} \boldsymbol{\Upsilon}_n + o_p(1) \\
&= \underline{\mathbf{B}}^{-1} \underline{\mathbf{G}}^{> \Sigma_\tau}{}^{1/2} \boldsymbol{\Upsilon}_n + \underline{\mathbf{B}}^{-1} \underline{\mathbf{A}}^{>} (\underline{\mathbf{A}} \underline{\mathbf{B}}^{-1} \underline{\mathbf{A}}^{>})^{-1} \underline{\mathbf{A}} \underline{\mathbf{B}}^{-1} \underline{\mathbf{G}}^{> \Sigma_\tau}{}^{1/2} \boldsymbol{\Upsilon}_n + o_p(1).
\end{aligned}$$

Subtract (A.20) to get

$$\rho_{\bar{n}}^-(\hat{\boldsymbol{\mu}}_n - \hat{\boldsymbol{\mu}}_{2s}) = \underline{\mathbf{B}}^{-1} \underline{\mathbf{A}}^{>} (\underline{\mathbf{A}} \underline{\mathbf{B}}^{-1} \underline{\mathbf{A}}^{>})^{-1} \underline{\mathbf{A}} \underline{\mathbf{B}}^{-1} \underline{\mathbf{G}}^{> \Sigma_\tau}{}^{1/2} \boldsymbol{\Upsilon}_n + o_p(1). \quad (\text{A.26})$$

We now look at the DM statistic, using the preceding results and a few more approximations and expansions. Since $\hat{\boldsymbol{\mu}}_n = \boldsymbol{\mu}_{0\tau} + o_p(1)$ and $\hat{\boldsymbol{\mu}}_{2s} = \boldsymbol{\mu}_{0\tau} + o_p(1)$, then $\hat{\boldsymbol{\mu}}_n = \hat{\boldsymbol{\mu}}_{2s} + o_p(1)$. Also $\underline{\mathbf{M}}_n(\tau) = \underline{\mathbf{G}} + o_p(1)$ and $\nabla^{>} \hat{\mathbf{M}}_n(\hat{\boldsymbol{\mu}}_{2s}, \tau) = \underline{\mathbf{G}} + o_p(1)$, so $\underline{\mathbf{M}}_n(\tau) = \nabla^{>} \hat{\mathbf{M}}_n(\hat{\boldsymbol{\mu}}_{2s}, \tau) + o_p(1)$. Using a Taylor expansion of $\rho_{\bar{n}}^- \hat{\mathbf{M}}_n(\hat{\boldsymbol{\mu}}_n, \tau)$ at $\hat{\boldsymbol{\mu}}_{2s}$ similar to (A.22),

$$\begin{aligned}
&\widehat{\text{DM}}_n - n \hat{\mathbf{M}}_n(\hat{\boldsymbol{\mu}}_n, \tau) > \Sigma^{-1} \hat{\mathbf{M}}_n(\hat{\boldsymbol{\mu}}_n, \tau) - n \hat{\mathbf{M}}_n(\hat{\boldsymbol{\mu}}_{2s}, \tau) > \Sigma^{-1} \hat{\mathbf{M}}_n(\hat{\boldsymbol{\mu}}_{2s}, \tau) \\
&= [\rho_{\bar{n}}^- \hat{\mathbf{M}}_n(\hat{\boldsymbol{\mu}}_{2s}, \tau) + \underline{\mathbf{M}}_n(\tau) \rho_{\bar{n}}^-(\hat{\boldsymbol{\mu}}_n - \hat{\boldsymbol{\mu}}_{2s})] > \Sigma^{-1} \\
&\quad [\rho_{\bar{n}}^- \hat{\mathbf{M}}_n(\hat{\boldsymbol{\mu}}_{2s}, \tau) + \underline{\mathbf{M}}_n(\tau) \rho_{\bar{n}}^-(\hat{\boldsymbol{\mu}}_n - \hat{\boldsymbol{\mu}}_{2s})] \\
&\quad \rho_{\bar{n}}^- \hat{\mathbf{M}}_n(\hat{\boldsymbol{\mu}}_n, \tau) > \rho_{\bar{n}}^- \Sigma^{-1} \hat{\mathbf{M}}_n(\hat{\boldsymbol{\mu}}_n, \tau) \\
&= \rho_{\bar{n}}^-(\hat{\boldsymbol{\mu}}_n - \hat{\boldsymbol{\mu}}_{2s}) > \underline{\mathbf{M}}_n(\tau) > \Sigma^{-1} \underline{\mathbf{M}}_n(\tau) \rho_{\bar{n}}^-(\hat{\boldsymbol{\mu}}_n - \hat{\boldsymbol{\mu}}_{2s}) \\
&\quad + 2 \rho_{\bar{n}}^-(\hat{\boldsymbol{\mu}}_n - \hat{\boldsymbol{\mu}}_{2s}) > \underline{\mathbf{M}}_n(\tau) > \Sigma^{-1} \rho_{\bar{n}}^- \hat{\mathbf{M}}_n(\hat{\boldsymbol{\mu}}_{2s}, \tau) + o_p(1) \\
&= \rho_{\bar{n}}^-(\hat{\boldsymbol{\mu}}_n - \hat{\boldsymbol{\mu}}_{2s}) > \underline{\mathbf{G}}^{> \Sigma_\tau}{}^{-1} \underline{\mathbf{G}} \rho_{\bar{n}}^-(\hat{\boldsymbol{\mu}}_n - \hat{\boldsymbol{\mu}}_{2s}) \\
&\quad + 2 \rho_{\bar{n}}^-(\hat{\boldsymbol{\mu}}_n - \hat{\boldsymbol{\mu}}_{2s}) > (\nabla^{>} \hat{\mathbf{M}}_n(\hat{\boldsymbol{\mu}}_{2s}, \tau) + o_p(1)) \Sigma^{-1} \rho_{\bar{n}}^- \hat{\mathbf{M}}_n(\hat{\boldsymbol{\mu}}_{2s}, \tau) + o_p(1)
\end{aligned}$$

$$\begin{aligned}
&= \rho_{\bar{n}}(\hat{\beta}_{2s})' \mathbf{G} > \underline{\Sigma}_\tau^{-1} \mathbf{G} \rho_{\bar{n}}(\hat{\beta}_{2s}) \\
&\quad + 2 \rho_{\bar{n}}(\hat{\beta}_{2s})' \underbrace{\mathbf{V} > \hat{\mathbf{M}}_n(\hat{\beta}_{2s}, \tau) \underline{\Sigma}_\tau^{-1} \rho_{\bar{n}}(\hat{\beta}_{2s})}_{=0 \text{ by (A.4)}} + o_p(1) \\
&= \rho_{\bar{n}}(\hat{\beta}_{2s})' \mathbf{G} > \underline{\Sigma}_\tau^{-1} \mathbf{G} \rho_{\bar{n}}(\hat{\beta}_{2s}) + o_p(1).
\end{aligned}$$

From (A.26),

$$\begin{aligned}
\underline{\Sigma}_\tau^{1/2} \mathbf{G} \rho_{\bar{n}}(\hat{\beta}_{2s}) &= \underline{\Sigma}_\tau^{1/2} \mathbf{G} \mathbf{B}^{-1} \mathbf{A} > (\mathbf{A} \mathbf{B}^{-1} \mathbf{A} >)^{-1} \mathbf{A} \mathbf{B}^{-1} \mathbf{G} > \underline{\Sigma}_\tau^{1/2} \boldsymbol{\Upsilon}_n \\
&\quad \stackrel{d}{=} \mathbf{V} \boldsymbol{\Upsilon}_n, \\
\mathbf{V} &= \underline{\Sigma}_\tau^{1/2} \mathbf{G} \mathbf{B}^{-1} \mathbf{A} > (\mathbf{A} \mathbf{B}^{-1} \mathbf{A} >)^{-1} \mathbf{A} \mathbf{B}^{-1} \mathbf{G} > \underline{\Sigma}_\tau^{1/2}.
\end{aligned}$$

By inspection, \mathbf{V} is idempotent, so $\widehat{\mathbf{D}}\widehat{\mathbf{M}}_n \stackrel{d}{=} \mathbf{V}$, where $\mathbf{N}(\mathbf{0}, \mathbf{I})$ from (A.19). Also, $\text{rank}(\mathbf{V}) = r$ since \mathbf{B} and $\underline{\Sigma}_\tau^{1/2}$ are positive definite and $\text{rank}(\mathbf{A}) = r - d_\beta = \text{rank}(\mathbf{G})$. Applying the identity for a quadratic form of a multivariate standard normal vector with an idempotent matrix of rank r , $\boldsymbol{\Upsilon}_n' \mathbf{V} \boldsymbol{\Upsilon}_n \sim \chi_r^2$. \square

Proof of Theorem 4. The proof follows the same steps as in Section 9.5 of Newey and McFadden (1994). As in (A.17), let

$$\boldsymbol{\Upsilon}_n = \underline{\Sigma}_\tau^{1/2} \rho_{\bar{n}}(\hat{\beta}_{0\tau}, \tau),$$

so $\rho_{\bar{n}}(\hat{\beta}_{0\tau}, \tau) = \underline{\Sigma}_\tau^{-1/2} \boldsymbol{\Upsilon}_n$. Using Lemma 2.4 in de Castro, Galvao, and Kaplan (2017), $\rho_{\bar{n}}(\hat{\beta}_{0\tau}, \tau) \stackrel{d}{=} \mathbf{N}(\mathbf{0}, \underline{\Sigma}_\tau)$, so

$$\boldsymbol{\Upsilon}_n \stackrel{d}{=} \boldsymbol{\Upsilon} \sim \mathbf{N}(\mathbf{0}, \mathbf{I}), \tag{A.27}$$

where \mathbf{I} is the $d_Z - d_\beta$ identity matrix. Define the symmetric, idempotent, rank $d_Z - d_\beta$ matrix

$$\mathbf{R} = \mathbf{I} - \underline{\Sigma}_\tau^{1/2} \mathbf{G} (\mathbf{G} > \underline{\Sigma}_\tau^{-1} \mathbf{G})^{-1} \mathbf{G} > \underline{\Sigma}_\tau^{-1/2}. \tag{A.28}$$

Specializing (A.7) in the proof of Theorem 2 to the two-step estimator,

$$\begin{aligned}
& \rho_{\bar{n}}^{-1}(\hat{\cdot}_{2s} \quad_{0\tau}) \\
&= \left\{ \left[\nabla_{>} \hat{\mathbf{M}}_n(\hat{\cdot}_{2s}, \tau) \right]_{>} \underline{\Sigma}^{-1} \underline{\mathbf{M}}_n(\tau) \right\}^1 \left[\nabla_{>} \hat{\mathbf{M}}_n(\hat{\cdot}_{2s}, \tau) \right]_{>} \underline{\Sigma}^{-1} \rho_{\bar{n}}^{-1} \hat{\mathbf{M}}_n(\quad_{0\tau}, \tau) \\
&= (\underline{\mathbf{G}}_{>} \underline{\Sigma}_{\tau}^{-1} \underline{\mathbf{G}})^1 \underline{\mathbf{G}}_{>} \underline{\Sigma}_{\tau}^{-1} \rho_{\bar{n}}^{-1} \hat{\mathbf{M}}_n(\quad_{0\tau}, \tau) + o_p(1) \\
&= (\underline{\mathbf{G}}_{>} \underline{\Sigma}_{\tau}^{-1} \underline{\mathbf{G}})^1 \underline{\mathbf{G}}_{>} \underline{\Sigma}_{\tau}^{-1} \underline{\Sigma}_{\tau}^{1/2} \boldsymbol{\Upsilon}_n + o_p(1), \tag{A.29}
\end{aligned}$$

using Assumption 11 and the assumption $\underline{\Sigma} = \underline{\Sigma}_{\tau} + o_p(1)$ from the statement of the theorem (and $\boldsymbol{\Upsilon}_n = O_p(1)$), where $\underline{\mathbf{M}}_n(\tau)$ is defined in (A.5) and $\nabla_{>}$ is the partial derivative wrt $\cdot_{>}$.

Using a Taylor expansion,

$$\begin{aligned}
\rho_{\bar{n}}^{-1} \hat{\mathbf{M}}_n(\hat{\cdot}_{2s}, \tau) &= \rho_{\bar{n}}^{-1} \hat{\mathbf{M}}_n(\quad_{0\tau}, \tau) + \overbrace{\underline{\mathbf{M}}_n(\tau)}^{= \underline{\mathbf{G}} + o_p(1)} \rho_{\bar{n}}^{-1}(\hat{\cdot}_{2s} \quad_{0\tau}) \\
&= \underline{\Sigma}_{\tau}^{1/2} \boldsymbol{\Upsilon}_n + [\underline{\mathbf{G}} + o_p(1)] \overbrace{\left[(\underline{\mathbf{G}}_{>} \underline{\Sigma}_{\tau}^{-1} \underline{\mathbf{G}})^1 \underline{\mathbf{G}}_{>} \underline{\Sigma}_{\tau}^{-1} \underline{\Sigma}_{\tau}^{1/2} \boldsymbol{\Upsilon}_n + o_p(1) \right]}^{\text{from (A.29)}} \\
&= \underline{\Sigma}_{\tau}^{1/2} \underline{\mathbf{R}} \boldsymbol{\Upsilon}_n + o_p(1).
\end{aligned}$$

The J -statistic is a quadratic form of this:

$$\begin{aligned}
\hat{J}_n &= \left[\rho_{\bar{n}}^{-1} \hat{\mathbf{M}}_n(\hat{\cdot}_{2s}, \tau) \right]_{>} \underline{\Sigma}^{-1} \left[\rho_{\bar{n}}^{-1} \hat{\mathbf{M}}_n(\hat{\cdot}_{2s}, \tau) \right] \\
&= \left[\underline{\Sigma}_{\tau}^{1/2} \underline{\mathbf{R}} \boldsymbol{\Upsilon}_n + o_p(1) \right]_{>} \left[\underline{\Sigma}_{\tau}^{-1} + o_p(1) \right] \left[\underline{\Sigma}_{\tau}^{1/2} \underline{\mathbf{R}} \boldsymbol{\Upsilon}_n + o_p(1) \right] \\
&= \boldsymbol{\Upsilon}_n_{>} \underline{\mathbf{R}} \underline{\Sigma}_{\tau}^{1/2} \underline{\Sigma}_{\tau}^{-1} \underline{\Sigma}_{\tau}^{1/2} \underline{\mathbf{R}} \boldsymbol{\Upsilon}_n + o_p(1) \\
&= \boldsymbol{\Upsilon}_n_{>} \overbrace{\underline{\mathbf{R}} \underline{\mathbf{R}}}^{= \underline{\mathbf{R}}} \boldsymbol{\Upsilon}_n + o_p(1) \\
&\stackrel{d}{=} \boldsymbol{\Upsilon}_{>} \underline{\mathbf{R}} \boldsymbol{\Upsilon} \\
&\chi_{d_Z \quad d_{\beta}}^2,
\end{aligned}$$

using the fact that both $\underline{\mathbf{R}}$ and $\underline{\Sigma}_{\tau}$ are symmetric and that $\underline{\mathbf{R}}$ is idempotent. The final $\chi_{d_Z \quad d_{\beta}}^2$ distribution comes from the identity for a quadratic form of a standard

multivariate Gaussian random vector with a symmetric, idempotent matrix of rank $d_Z = d_\beta$. □

Appendix B

Appendix for Chapter 2

B.1 Bandwidth for estimating Jacobian matrix

Kato (2012) provides an AMSE-optimal bandwidth for estimating the Jacobian matrix in conventional quantile regression. He shows that in the linear CQF model

$$Y = \mathbf{X}'\beta_0 + U \text{ with } Q_\tau(U | \mathbf{X}) = 0,$$

the AMSE-optimal bandwidth for estimating $E[f_0(0 | \mathbf{X})\mathbf{X}\mathbf{X}']$ is

$$h_{\text{opt}} = n^{-1/5} \left(\frac{4.5 \sum_{j,k=1}^{d_X} E[f_0(0 | \mathbf{X})X_j^2 X_k^2]}{\sum_{j,k=1}^{d_X} \left(E[f_0^{(2)}(0 | \mathbf{X})X_j X_k] \right)^2} \right)^{1/5}, \quad (\text{B.1})$$

where $f_0(0 | \mathbf{X})$ is the conditional PDF of U , conditional on the regressors \mathbf{X} ; and $f_0^{(2)}(0 | \mathbf{X})$ is the conditional PDF's second derivative. If additionally $U \perp \mathbf{X}$, then $f_0(0 | \mathbf{X}) = f_0(0)$, and formula (B.1) reduces to

$$h_{\text{opt}} = n^{-1/5} \left(\frac{4.5 \sum_{j,k=1}^{d_X} E[f_0(0)X_j^2 X_k^2]}{\sum_{j,k=1}^{d_X} \left(E[f_0^{(2)}(0)X_j X_k] \right)^2} \right)^{1/5}. \quad (\text{B.2})$$

where $f_0(\cdot)$ is the unconditional PDF of U and $f_0^{(2)}(\cdot)$ is its second derivative.

Kato (2012) further simplifies (B.2) when U is standard normal. That is, in the linear CQF model

$$Y = \mathbf{X}^\top \boldsymbol{\theta}_0 + U \text{ with } U | \mathbf{X} \sim \text{N}(0, 1),$$

the sample analog of the optimal bandwidth (B.1) reduces to

$$\hat{h}_{\text{opt}} = n^{-1/5} \left(\frac{4.5 \sum_{j,k=1}^{d_X} (n^{-1} \sum_{i=1}^n X_{ij}^2 X_{ik}^2)}{\alpha(\tau) \sum_{j,k=1}^{d_X} (n^{-1} \sum_{i=1}^n X_{ij} X_{ik})^2} \right)^{1/5}, \quad (\text{B.3})$$

where

$$\alpha(\tau) = \left[1 - \Phi(\tau) \right]^2 \phi(\tau), \quad (\text{B.4})$$

and $\Phi(\cdot)$ and $\phi(\cdot)$ are standard normal CDF and PDF.¹

I extend the optimal bandwidth formula in (B.3) to allow for general scale parameter.

Proposition 21. *In a linear CQF model*

$$Y = \mathbf{X}^\top \boldsymbol{\theta}_0 + U \text{ with } U | \mathbf{X} \sim \text{N}(\mu, \sigma^2), \text{ and } \mathbb{Q}_\tau(U | \mathbf{X}) = 0,$$

the sample analog of the AMSE-optimal bandwidth for estimating the population Jacobian matrix is

$$\hat{h}_{\text{opt}} = n^{-1/5} \left(\frac{4.5 \sum_{j,k=1}^{d_X} (n^{-1} \sum_{i=1}^n X_{ij}^2 X_{ik}^2)}{\hat{\alpha}(\tau) \sum_{j,k=1}^{d_X} (n^{-1} \sum_{i=1}^n X_{ij} X_{ik})^2} \right)^{1/5}, \quad (\text{B.5})$$

where

$$\hat{\alpha}(\tau) = \frac{1}{\delta^5} \left[1 - \Phi(\tau) \right]^2 \phi(\tau) \quad (\text{B.6})$$

and $\Phi(\cdot)$ and $\phi(\cdot)$ are standard normal CDF and PDF.

Proof. We can consider the $\text{N}(\mu, \sigma^2)$ distribution of U as the $\text{N}(0, \sigma^2)$ distribution with a shift of μ . We have

$$f_U(0) = \phi_{\text{N}(0, \sigma^2)} \left(\frac{0 - \mu}{\sigma} \right),$$

where $\phi_{\text{N}(0, \sigma^2)}$ and $\Phi_{\text{N}(0, \sigma^2)}$ are the $\text{N}(0, \sigma^2)$ PDF and CDF.

¹The version in Kato (2012), $\alpha(\tau) = [1 - \Phi(\tau)]^2 \phi(\tau)$, appears to have a typo; see the below proof of Proposition 21.

For a standard normal distribution,

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2},$$

which has the property that

$$\phi'(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \left(-\frac{2x}{2} \right) = -x\phi(x).$$

For a $N(0, \sigma^2)$ density function,

$$\phi_{N(0, \sigma^2)}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}},$$

its first derivative is

$$\phi'_{N(0, \sigma^2)}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \left(-\frac{2x}{2\sigma^2} \right) = \left(-\frac{x}{\sigma^2} \right) \phi_{N(0, \sigma^2)}(x).$$

By the chain rule, its second derivative is

$$\begin{aligned} \phi^{(2)}_{N(0, \sigma^2)}(x) &= \left(-\frac{1}{\sigma^2} \right) \phi_{N(0, \sigma^2)}(x) + \left(-\frac{x}{\sigma^2} \right) \phi'_{N(0, \sigma^2)}(x) \\ &= \left(-\frac{1}{\sigma^2} \right) \phi_{N(0, \sigma^2)}(x) + \left(-\frac{x}{\sigma^2} \right) \left(-\frac{x}{\sigma^2} \right) \phi_{N(0, \sigma^2)}(x) \\ &= \frac{1}{\sigma^2} \left(\frac{x^2}{\sigma^2} - 1 \right) \phi_{N(0, \sigma^2)}(x). \end{aligned}$$

Therefore,

$$f_U^{(2)}(0) = \frac{1}{\sigma^2} \left(\frac{\left(\frac{1}{\sigma^2} N(0, \sigma^2)(\tau) \right)^2}{\sigma^2} - 1 \right) \phi_{N(0, \sigma^2)} \left(\frac{1}{\sigma^2} N(0, \sigma^2)(\tau) \right).$$

Since

$$\frac{1}{\sigma^2} N(0, \sigma^2)(\tau) = \frac{1}{\sigma} N(0, 1)(\tau) \text{ and } \phi_{N(0, \sigma^2)}(x) = \frac{1}{\sigma} \phi_{N(0, 1)}(x/\sigma),$$

$f_U(0)$ and $f_U^{(2)}(0)$ further reduce to

$$\begin{aligned} f_U(0) &= \frac{1}{\sigma} \phi_{N(0, 1)} \left(\frac{1}{\sigma} N(0, \sigma^2)(\tau) \right) \\ &= \frac{1}{\sigma} \phi_{N(0, 1)} \left(N(0, 1)(\tau) \right) \end{aligned}$$

$$= \frac{1}{\sigma} \phi_{N(0,1)} \left(\frac{N(0,1)}{\sigma}(\tau) \right),$$

and

$$\begin{aligned} f_U^{(2)}(0) &= \frac{1}{\sigma^2} \left(\frac{\left(\frac{N(0,\sigma^2)}{\sigma}(\tau) \right)^2}{\sigma^2} - 1 \right) \frac{1}{\sigma} \phi_{N(0,1)} \left(\frac{N(0,\sigma^2)}{\sigma}(\tau) / \sigma \right) \\ &= \frac{1}{\sigma^2} \left(\frac{\left(\sigma \frac{N(0,1)}{\sigma}(\tau) \right)^2}{\sigma^2} - 1 \right) \frac{1}{\sigma} \phi_{N(0,1)} \left(\frac{N(0,1)}{\sigma}(\tau) \right) \\ &= \frac{1}{\sigma^3} \left(\left(\frac{N(0,1)}{\sigma}(\tau) \right)^2 - 1 \right) \phi_{N(0,1)} \left(\frac{N(0,1)}{\sigma}(\tau) \right). \end{aligned}$$

Then the optimal bandwidth in (B.1) reduces to

$$\begin{aligned} h_{\text{opt}} &= n^{-1/5} \left(\frac{4.5 \sum_{j,k=1}^d \mathbb{E}[f_0(0) X_j^2 X_k^2]}{\sum_{j,k=1}^d \left(\mathbb{E}[f_0^{(2)}(0) X_j X_k] \right)^2} \right)^{1/5} \\ &= n^{-1/5} \left(\frac{4.5 \sum_{j,k=1}^d \mathbb{E}[f_0(0) X_j^2 X_k^2]}{\sum_{j,k=1}^d \left(\mathbb{E}[f_0^{(2)}(0) X_j X_k] \right)^2} \right)^{1/5} \\ &= n^{-1/5} \left(\frac{4.5 f_0(0) \sum_{j,k=1}^d \mathbb{E}[X_j^2 X_k^2]}{\left(f_0^{(2)}(0) \right)^2 \sum_{j,k=1}^d \left(\mathbb{E}[X_j X_k] \right)^2} \right)^{1/5} \\ &= n^{-1/5} \left(\frac{4.5 \sum_{j,k=1}^d \mathbb{E}[X_j^2 X_k^2]}{\alpha(\tau) \sum_{j,k=1}^d \left(\mathbb{E}[X_j X_k] \right)^2} \right)^{1/5}, \end{aligned}$$

where

$$\begin{aligned} \alpha(\tau) &= \frac{\left[f_0^{(2)}(0) \right]^2}{f_0(0)} = \frac{\left\{ \frac{1}{\sigma^3} \left[\left(\frac{N(0,1)}{\sigma}(\tau) \right)^2 - 1 \right] \phi_{N(0,1)} \left(\frac{N(0,1)}{\sigma}(\tau) \right) \right\}^2}{\frac{1}{\sigma} \phi_{N(0,1)} \left(\frac{N(0,1)}{\sigma}(\tau) \right)} \\ &= \frac{1}{\sigma^5} \left[\left(\frac{N(0,1)}{\sigma}(\tau) \right)^2 - 1 \right]^2 \phi_{N(0,1)} \left(\frac{N(0,1)}{\sigma}(\tau) \right). \end{aligned}$$

In the special case $\sigma^2 = 1$, it reduces to $\alpha(\tau) = \left[1 - \left(\frac{N(0,1)}{\sigma}(\tau) \right)^2 \right]^2 \phi_{N(0,1)} \left(\frac{N(0,1)}{\sigma}(\tau) \right)$. The sample analog replaces σ with $\hat{\sigma}$ and replaces the expectations with sample averages. \square

The optimal bandwidth formula (B.5) is very helpful in practice, especially with multiple regressors since nonparametric conditional density derivative estimators converge slowly. In these cases, using the optimal bandwidth formula (B.1) might suffer large estimation error.

Kato's (2012) optimal bandwidth formula is under the assumption that regressors are exogenous. That is, X represents both regressor vector and the full instrument vector. I modify Kato's optimal bandwidth formula by replacing $X_j X_k$ term with $X_j Z_k$ for $j = 1, \dots, d_X$ and $k = 1, \dots, d_Z$ to adapt to IVQR. I also replace the conditional density functions that are conditioning on regressors to that conditioning on both the regressors and the instruments, since the IVQR object of interest is $E[f_0(0 | \mathbf{X}, \mathbf{Z}) \mathbf{X} \mathbf{Z}']$; e.g., see (16) in Kaplan and Sun (2017b).

Following this modification for IVQR, the general AMSE-optimal bandwidth formula (B.1) for estimating the Jacobian matrix becomes

$$h_{\text{opt}} = n^{-1/5} \left(\frac{4.5 \sum_{j=1}^{d_X} \sum_{k=1}^{d_Z} E[f_0(0 | \mathbf{X}, \mathbf{Z}) X_j^2 Z_k^2]}{\sum_{j=1}^{d_X} \sum_{k=1}^{d_Z} \left(E[f_0^{(2)}(0 | \mathbf{X}, \mathbf{Z}) X_j Z_k] \right)^2} \right)^{1/5}, \quad (\text{B.7})$$

and the AMSE-optimal plug-in bandwidth formula (B.5) in the general Gaussian model becomes

$$\hat{h}_{\text{opt}} = n^{-1/5} \left(\frac{4.5 \sum_{j=1}^{d_X} \sum_{k=1}^{d_Z} (n^{-1} \sum_{i=1}^n X_{ij}^2 Z_{ik}^2)}{\hat{\Delta}(\tau) \sum_{j=1}^{d_X} \sum_{k=1}^{d_Z} (n^{-1} \sum_{i=1}^n X_{ij} Z_{ik})^2} \right)^{1/5}, \quad (\text{B.8})$$

with the same $\hat{\Delta}(\tau)$ as in (B.6).

For simulation model 3, I use optimal bandwidth formula (B.7) to estimate the Jacobian matrix. Simulation model 3 is a small model with one binary treatment variable and one binary instrument, so the conditional PDF and its second derivative are relatively easy to estimate nonparametrically. The optimal bandwidth formula (B.7) could probably be improved, but it seems to work well in the simulations; any further improvement to the bandwidth would only further improve the performance of the averaging

GMM estimators.

For simulation models 1 and 2, I instead use the optimal bandwidth formula (B.8). These two simulation models are large models with multiple continuous endogenous regressors and instruments. Fully nonparametric estimation of the conditional density and its second derivative is possible but would have large estimation error given my sample sizes.

B.2 Averaging GMM method estimator details

The averaging estimator applying CLS framework is defined in four steps.

Step 1: Translate conservative estimator

I use de Castro et al.'s (2019) smoothed two-step GMM method for computation to obtain the IVQR estimator as the conservative GMM estimator, translating the CLS averaging GMM framework to the IVQR model.

I replace the IVQR moment functions

$$\mathbf{g}_{1i}(\beta, \tau) = \mathbf{Z}_i' [1 \quad Y_i - \mathbf{X}_i' \beta - \tau] \quad (\text{B.9})$$

with smoothed moment functions

$$\mathbf{g}_{1i}(\beta, \tau) = \mathbf{Z}_i' [\mathcal{F}((Y_i - \mathbf{X}_i' \beta - \tau)/h_n) - \tau], \quad (\text{B.10})$$

where h_n is the sequence of smoothing bandwidth and $\mathcal{F}(\cdot)$ is the same smoothed indicator function as defined in de Castro et al. (2019).²

$${}^2\mathcal{F}(u) = \begin{cases} 0 & \text{if } u \leq -1 \\ 0.5 + \frac{105}{64}(u + \frac{5}{3}u^3 + \frac{7}{5}u^5 - \frac{3}{7}u^7) & \text{if } -1 < u < 1 \\ 1 & \text{if } 1 \leq u. \end{cases}$$

This $\mathcal{F}(\cdot)$ is the integral of a fourth-order kernel function, which Kaplan and Sun (2017b) show leads to higher-order improvements in the linear iid IVQR setting.

The IVQR estimator is the conservative GMM estimator,

$$\hat{\tau}_{IVQR} = \hat{\tau}_1 = \arg \min_{\tau \in \mathcal{B}} \hat{\mathbf{M}}_{1n}(\tau)' \hat{\Sigma}_1^{-1} \hat{\mathbf{M}}_{1n}(\tau), \quad (\text{B.11})$$

where $\hat{\mathbf{M}}_{1n}(\tau) = 1/n \sum_{i=1}^n \mathbf{g}_{1i}(\tau)$ and $\hat{\Sigma}_1$ is some consistent estimator of the (long-run) variance of the IVQR sample moments from the first step. With iid sampling,

$$\hat{\Sigma}_1 = \frac{1}{n} \sum_{i=1}^n \mathbf{g}_{1i}(\tau) \mathbf{g}_{1i}(\tau)' - \hat{\mathbf{M}}_{1n}(\tau) \hat{\mathbf{M}}_{1n}(\tau)', \quad (\text{B.12})$$

where τ_0 is some initial consistent estimator of τ_0 . Specifically, my τ_0 is de Castro et al.'s (2019) method of moments estimator with instrument vector equal to the linear projection of \mathbf{X} onto \mathbf{Z} .

Step 2: Propose additional moments

I propose two types of additional moments based on intuitions in Section 2.3.

The first type is the conventional QR moments. Write the unconditional QR moments (2.6) in two separate parts,

$$\mathbf{0}_{d_{X_{exog}}} = E f \mathbf{X}_{exog,i} [1 f Y_i - \mathbf{X}_i' \theta_{\tau} - g(\tau)] g, \quad (\text{B.13})$$

$$\mathbf{0}_{d_D} = E f \mathbf{D}_i [1 f Y_i - \mathbf{X}_i' \theta_{\tau} - g(\tau)] g. \quad (\text{B.14})$$

I use (B.14) but not (B.13), so the additional moment functions are

$$\mathbf{g}_i(\tau) = \mathbf{D}_i [1 f Y_i - \mathbf{X}_i' \theta_{\tau} - g(\tau)]. \quad (\text{B.15})$$

The first part (B.13) is already contained in the IVQR population moments (1.1), therefore the “conservative moments,” as the exogenous regressors \mathbf{X}_{exog} are contained in the full instruments \mathbf{Z} . I use the second part (B.14), the QR moments with potentially endogenous regressors \mathbf{D} , as the “additional moments” to compute the “aggressive moments,” the “aggressive estimator,” and the “averaging estimator.” (For computation, I

again use a smoothed version of the moment function, like (B.10) but with \mathbf{D}_i replacing \mathbf{Z}_i .) I call this the IVQR-QR type of averaging.

The second type of additional moments use the 2SLS slope moment functions

$$\mathbf{g}_i(\beta, \tau) = (\mathbf{Z}_{1,i} \quad \mathbf{Z}_1)(Y_i - \mathbf{X}_i^\theta \beta), \quad (\text{B.16})$$

where \mathbf{Z}_1 is the instruments without the intercept term, i.e., $\mathbf{Z} = (1, \mathbf{Z}_1)$. I use the demeaned instruments $(\mathbf{Z}_{1,i} \quad \mathbf{Z}_1)$ in these additional moments, where $\mathbf{Z}_1 = 1/n \sum_{i=1}^n \mathbf{Z}_{1,i}$, to represent the 2SLS slope condition $\text{Cov}(\mathbf{Z}_1, Y - \mathbf{X}^\theta \beta) = \mathbf{0}$. With the 2SLS slope moments as the additional moments, I obtain the corresponding aggressive and averaging estimators. I call this the IVQR-2SLS type of averaging.

Step 3: Form aggressive estimator

Incorporating the additional moments (either the potentially endogenous QR moments or the 2SLS slope moments) to the IVQR moments, I obtain the aggressive moments. Define the aggressive GMM estimator as

$$\hat{\beta}_2 = \arg \min_{\beta} \hat{\mathbf{M}}_{2n}(\beta, \tau)^\theta \underline{\Sigma}_2^{-1} \hat{\mathbf{M}}_{2n}(\beta, \tau), \quad (\text{B.17})$$

where

$$\hat{\mathbf{M}}_{2n}(\beta, \tau) = \frac{1}{n} \sum_{i=1}^n \mathbf{g}_{2i}(\beta, \tau), \quad \mathbf{g}_{2i}(\beta, \tau) = (\mathbf{g}_{1i}(\beta, \tau)^\theta, \mathbf{g}_i(\beta, \tau)^\theta)^\theta. \quad (\text{B.18})$$

The aggressive moments contain both the conservative moments and additional moments. The GMM weighting matrix $\underline{\Sigma}_2^{-1}$ is constructed in the same way as $\underline{\Sigma}_1^{-1}$, except with aggressive moments (using \mathbf{g}_2) instead of conservative moments (using \mathbf{g}_1). The subscript “2” denotes “aggressive” here.

Step 4: Define averaging estimator

Following CLS, define the averaging GMM estimator as

$$\hat{\omega}_{\text{AVG}} = (1 - \hat{\omega}) \hat{\omega}_1 + \hat{\omega} \hat{\omega}_2. \quad (\text{B.19})$$

I use the empirical averaging weight $\hat{\omega}$ in CLS, which is the sample analog of their optimal weight:

$$\hat{\omega} = \frac{\text{tr}(\underline{\Upsilon}(\hat{\Sigma}_1, \hat{\Sigma}_2))}{n(\hat{\omega}_1, \hat{\omega}_2)' \underline{\Upsilon}(\hat{\omega}_1, \hat{\omega}_2) + \text{tr}(\underline{\Upsilon}(\hat{\Sigma}_1, \hat{\Sigma}_2))} \quad (\text{B.20})$$

where $\underline{\Upsilon}$ is related to the loss function (details below), and

$$\hat{\Sigma}_k = (\hat{\mathbf{G}}_k' \hat{\Sigma}_k^{-1} \hat{\mathbf{G}}_k)^{-1}, \quad (\text{B.21})$$

$$\hat{\mathbf{G}}_k = \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \theta} \mathbf{g}_{ki}(\hat{\omega}_k, \tau), \text{ for } k = 1, 2. \quad (\text{B.22})$$

The middle part $\hat{\Sigma}_k$ is the estimator of the covariance matrix of the conservative sample moments and aggressive sample moments for $k = 1$ and $k = 2$, respectively; its inverse $\hat{\Sigma}_k^{-1}$ is the efficient two-step GMM weighting matrix. $\hat{\mathbf{G}}_k$ is the estimator of the Jacobian matrix of the conservative moments and aggressive moments $k = 1$ and $k = 2$, respectively.

Both $\hat{\Sigma}_k$ and $\hat{\mathbf{G}}_k$, for $k = 1$ and $k = 2$, are evaluated at the conservative GMM estimator $\hat{\omega}_1$. Therefore, they are consistent regardless of misspecification of the additional moments. The diagonal matrix $\underline{\Upsilon}$ measures how much we weight each element in the parameter vector in the (scaled) loss function $(\hat{\omega}_1, \hat{\omega}_2)' \underline{\Upsilon}(\hat{\omega}_1, \hat{\omega}_2)$, as in (3.9) of CLS. When $\underline{\Upsilon}$ is the identity matrix, the expected loss (i.e., risk) becomes the sum of MSEs of each component of the estimator vector, $\sum_{j=1}^{d_X} E[(\hat{\beta}_j - \beta_{0\tau j})^2]$.

B.3 Uniform dominance condition

The uniform dominance condition in this paper is referring to the key sufficient conditions under which the averaging estimator uniformly dominates the estimator to be improved. The uniform dominance means the averaging estimator has a smaller MSE, thus is preferred, regardless of the DGPs.

CLS Theorem 5.2 gives key sufficient conditions under which the averaging GMM estimator uniformly dominates the conservative GMM estimator.³ One of the CLS key sufficient conditions is reduced to at least four parameters in some special case (certain $\mathbf{\Upsilon}$ in the loss function, or homoskedastic error with MSE loss function.) CLS footnote 10 gives a review about the uniform dominance condition in the Gaussian location model in other studies (Green and Strawderman, 1991; Judge and Mittelhammer, 2004; Kim and White, 2001; Mittelhammer and Judge, 2005). In most of the studies, the uniform dominance condition requires more than four parameters in the model. Hansen (2017b) averages the 2SLS and OLS to improve on 2SLS. He shows it requires at least 3 endogenous regressors for his averaging estimator to uniformly dominate the 2SLS estimator.

In a small model with only one endogenous regressor, the averaging estimator is not uniformly better. Simulation evidence of this is seen in my simulation model 3, CLS simulation model S3, and simulations in Hansen (2017b). CLS simulation model S3 has 1 endogenous regressor and 6 exogenous regressors, which suggests the uniform dominance condition is more related with the number of *endogenous* regressors than the total number of regressors.

A common understanding about the uniform dominance condition in averaging esti-

³The uniform dominance condition (UDC) in CLS Theorem 5.2 ($\text{tr}(\mathbf{\Upsilon}(\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2)) > 0$ and $\text{tr}(\mathbf{\Upsilon}(\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2)) > 4\rho_{\max}(\mathbf{\Upsilon}(\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2))$) can be checked for any specific DGP, but not at a high level for models, because the UDC is related to the number of regressors and to the choice of loss function (through $\mathbf{\Upsilon}$). (It requires the matrix $\mathbf{\Upsilon}(\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2)$ has at least 4 positive eigenvalues, thus at least 4 parameters.) Note that violation of UDC does not necessarily mean averaging estimator cannot uniformly dominate the conservative estimator, since UDC is a *sufficient* condition.

mation is that the averaging estimator has uniform dominance property in large models with multiple endogenous regressors.

Even in the case the averaging estimator does not uniformly dominate the original estimator, the averaging estimator may still be preferred if the averaging estimator improves most of the time, depending on the user’s preference.

In practice, it is common that the (endogenous) treatment variables we are interested in have interaction terms with other variables. In such cases, multiple endogenous regressors are involved in the model, and the uniform dominance condition is likely satisfied. Then, using the averaging methods proposed in this paper can improve estimation efficiency over the IVQR estimator.

B.4 Simulation model 3 details

B.4.1 Simulation DGP

This Job Training Partnership Act-based simulation DGP generalizes DGP 1 in de Castro et al. (2019) to a class of DGPs that allow different combinations of endogeneity, heterogeneity, and fat-tail levels.

Consider a structural random coefficient model that describes the impact of a job training program D_i on individual i ’s earnings Y_i ,

$$Y_i = \beta(U_i) + \gamma(U_i) D_i, \quad (\text{B.23})$$

The unobserved scalar $U_i \sim \text{Unif}(0, 1)$ incorporates other earnings determinants like ability. The individual-specific intercept and slope depend on U_i , through functions $\beta(\cdot)$ and $\gamma(\cdot)$.

The job training offer (eligibility) Z_i is completely randomized with $\text{P}(Z_i = 1) = \text{P}(Z_i = 0) = 1/2$. $Z_i = 1$ means individual i is offered the job training, and $Z_i = 0$

means individual i is not offered the training. This Z_i is a valid instrument for D_i .

When the individual is offered the job training program, he can choose to take this training program or not. If individual i actually takes the training, then $D_i = 1$, otherwise $D_i = 0$. There is self-selection based on the unobserved innate ability. I model the take-up decision as

$$P(D_i = 1 \mid Z_i = 1, U_i) = 0.5 + c_1(U_i - 0.5), \quad P(D_i = 1 \mid Z_i = 0, U_i) = 0. \quad (\text{B.24})$$

It describes the conditional probability of the individual taking the job training program, conditional on the fact that the individual is offered, and given his innate ability.

The constant $c_1 \in [0, 1]$ captures the endogeneity level of the observed choice variable D_i . When $c_1 = 0$, the conditional probability (B.24) becomes a constant 0.5. It means the individual's innate ability is irrelevant to the decision to take the job training offer:

$$P(D_i = 1 \mid Z_i = 1, U_i) = P(D_i = 1 \mid Z_i = 1) = 0.5. \quad (\text{B.25})$$

It implies

$$P(D_i = 1 \mid U_i) = P(D_i = 1). \quad (\text{B.26})$$

The treatment variable D_i is uncorrelated with the individual's unobserved innate ability. There is no endogeneity in the model. When $c_1 = 1$, the conditional probability function becomes

$$P(D_i = 1 \mid Z_i = 1, U_i) = U_i, \quad (\text{B.27})$$

which depends more strongly on innate ability than for any other $c_1 < 1$.

Each c_1 value corresponds to a DGP with a distinct endogeneity level. As c_1 increases, innate ability plays an increasing role in the self-selection process, i.e., increasing endogeneity of D_i . I consider $c_1 = 0$ and $c_1 = 1$ as the two endpoints, denoting the DGPs with no endogeneity and most endogeneity, respectively.

One of the main features of quantile regression is to capture slope heterogeneity. I

use the individual-specific random coefficient model to represent the slope heterogeneity. It assumes the treatment effect of training program on each individual is heterogeneous. The slope term, which measures the treatment effect of the training program, is a function of the individual's specific unobserved innate ability. Let the slope function be

$$\gamma(U_i) = 100c_2U_i^4. \quad (\text{B.28})$$

The nonnegative constant c_2 indicates the degree of treatment effect heterogeneity. Each c_2 value corresponds to a different DGP with a specific heterogeneity level. When $c_2 = 0$, the slope term becomes a constant (zero) for all individuals. That means the treatment effect is the same for both high and low ability individuals. There is no heterogeneity. The larger c_2 is, the larger is the difference of the treatment effect between lower and higher ability individuals, i.e., larger heterogeneity. In the simulation model, c_2 increases from 0 to 1 to indicate a class of DGPs with different (increasing) heterogeneity levels.

The intercept function is set as $\beta(U_i) = 60 + Q(U_i)$, with two possible $Q(\cdot)$. First, $Q(U_i)$ follows a χ_3^2 distribution. It represents that without training, higher-ability people have higher earnings, and lower-ability people have lower earnings. Second, $Q(U_i)$ follows a t -distribution with $c_3 - 1$ degrees of freedom. When considering the second case (varying c_3), the slope term is set to be a constant zero, i.e., $c_2 = 0$, no heterogeneous training effects.

When $c_3 = 1$, the distribution of earning is very fat-tailed (Cauchy) with respect to individuals' innate ability. As c_3 increases, the distribution of earning becomes less fat-tailed, approaching a normal distribution. Each c_3 value represents a specific DGP with a different fat-tail level of the earnings distribution with respect to individuals' ability levels.

Overall, in simulation model 1, I consider 11 possible values of each simulation

parameter: $c_1 \in \{0, 0.1, \dots, 1\}$, $c_2 \in \{0, 0.1, \dots, 1\}$, and $c_3 \in \{1, 1.5, \dots, 4.5, 5, 6, 7\}$. In the first case of the χ_3^2 intercept term, there is no c_3 involved. In total, it has $11 \times 11 = 121$ DGPs of the $(c_1, c_2, c_3 = \text{NA})$ combinations. In the second case of the t -distributed intercept term, different c_3 values are involved, but c_2 is set to zero. It also has $11 \times 11 = 121$ DGPs of the $(c_1, c_2 = 0, c_3)$ combinations. Therefore, this simulation model has $121 + 121 = 242$ DGPs altogether.

I first run simulations with all 242 DGPs. Based on the simulation results, I select 14 representative DGPs with distinct combinations of endogeneity, heterogeneity, and fat-tail levels that span the ranges of each. I label the endogeneity level as “No,” “Some,” or “Much,” based on the simulation results with respect to c_1 values. Similarly, I label heterogeneity level in four categories: “No,” “Little,” “Some,” and “Much,” with respect to c_2 ; and fat-tail level in four categories: “NA,” “Little,” “Some,” and “Much,” based on c_3 .⁴

B.4.2 Simulation results

Table B.1 presents the finite-sample rRMSE of the estimators proposed in this paper relative to that of the IVQR estimator at the median, $\tau = 0.5$.

Table B.1 shows no averaging estimator uniformly dominates IVQR in simulation model 3. That is, no column has relative rRMSE below 1 for each DGP. But, they do not suffer much in the least favorable cases, even when 2SLS and/or QR (or their aggressive GMM counterparts) have relative rRMSE well above 1.

As expected, the IVQR-2SLS averaging estimator performs better than (or almost the same as) the IVQR estimator in the DGPs with no slope heterogeneity. The efficiency

⁴Specifically, based on the simulation results, I label $c_1 = 0$ as “no endogeneity”; $c_1 \in [0.1, 0.3]$ as “some endogeneity”; and $c_1 \in [0.5, 1]$ as “very much endogeneity” in a DGP. I label $c_2 = 0$ as “no heterogeneity”; $c_2 = 0.1$ as “little heterogeneity”; $c_2 \in [0.3, 0.6]$ as “some heterogeneity”; $c_2 = 0.9$ as “very much heterogeneity.” I label $c_3 = 1$ as “very much fat tail”; $c_3 = 3$ as “some fat tail”; and the case $c_3 \in [5, 7]$ as “little fat tail.”

Table B.1: Finite-sample relative rRMSE in JTPA-based simulation model 3, $\tau = 0.5$.

DGP	Endog	Hetero	Fattail	IVQR.2SLS			IVQR.QR			BS	CON
				AVG	AGG	2SLS	AVG	AGG	QR		
1	No	No	NA	0.95	0.87	0.95	0.88	0.76	0.60	0.72	0.31
2	Some	No	NA	0.97	1.02	1.05	0.96	1.05	1.06	0.97	0.32
3	Some	No	NA	0.97	0.93	1.01	1.13	1.74	1.77	0.99	0.32
4	Much	No	NA	0.97	0.94	0.92	1.04	2.56	2.71	0.99	0.34
5	No	Some	NA	1.13	4.62	4.10	0.87	0.72	0.57	0.92	1.39
6	Some	Some	NA	1.13	5.04	4.45	0.92	0.88	0.78	0.98	1.35
7	Much	Some	NA	1.08	5.80	5.10	1.10	2.81	2.95	1.14	1.20
8	Much	Some	NA	1.08	7.64	6.87	1.04	3.68	3.53	1.20	1.50
9	Some	Much	NA	1.06	5.66	5.18	0.95	0.86	0.83	0.95	2.63
10	No	No	Some	0.95	1.00	1.17	0.89	0.78	0.54	0.70	0.18
11	No	No	Much	1.01	1.00	28.51	0.77	0.69	0.53	0.78	0.19
12	Much	No	Little	0.97	0.89	0.93	1.09	2.56	2.62	1.03	0.17
13	Much	No	Some	0.97	1.08	1.27	1.03	2.75	2.73	1.08	0.16
14	Much	No	Much	1.01	1.03	29.68	1.03	2.75	2.56	1.06	0.20

400 replications. 50 bootstraps. Sample size is 1000.

Columns 1–4 describe the DGP. Columns 5–11 report the relative rRMSE (i.e., rRMSE divided by the IVQR rRMSE) of the IVQR-2SLS averaging estimator, IVQR-2SLS aggressive estimator, 2SLS estimator, IVQR-QR averaging estimator, IVQR-QR aggressive estimator, QR estimator, and bootstrap averaging estimator, respectively. Column 12 reports the absolute rRMSE of the IVQR estimator (whose relative rRMSE is 1 by definition). Columns 5, 8, and 11 are the three averaging estimators of primary interest. Columns 6, 7, 9, and 10 are reported for reference to better see the mechanisms of the averaging methods.

gain is not much (up to 5%). In the DGPs with slope heterogeneity, the relative rRMSE of IVQR-2SLS averaging is around 1.06 to 1.13, while that of IVQR-2SLS aggressive estimator and 2SLS estimator is around 5 to 7. This indicates that although IVQR-2SLS averaging estimator is worse than IVQR, its rRMSE is much closer to 1 compared with the 2SLS estimator and aggressive estimator, due to putting most of the weight on the conservative estimator in these cases.

The IVQR-QR averaging estimator performs around 11–23% better than IVQR estimator in the DGPs with no endogeneity and no heterogeneity. As the endogeneity level of treatment variable increases, the QR moments become more misspecified, and the performance of the IVQR-QR averaging estimator becomes less favorable. In the DGPs with much endogeneity, the IVQR-QR averaging estimator performs worse than the IVQR estimator by around 3–13%. Compared with the relative rRMSE of the QR estimator and that of the IVQR-QR aggressive estimator, the averaging estimator is much closer to 1 and saved from even worse performance by putting most of the weight on the conservative estimator.

The bootstrap averaging estimator that averages among IVQR, 2SLS, and QR performs better than IVQR in all but one DGP in which either 2SLS or QR (or both) has relative rRMSE below 1. In the DGPs in which both 2SLS and QR are worse than IVQR, the bootstrap averaging estimator performs worse than IVQR, but not by as much. For example, in DGPs 7, 8, 13, and 14, the bootstrap averaging estimator has relative rRMSE from 1.06 to 1.20, compared to 1.27 to 29.68 for 2SLS and 2.56 to 3.53 for QR.

B.4.3 Results at other quantiles

Results at other quantiles are shown in Appendix B.5.3. It includes results at $\tau = 0.2$ up to $\tau = 0.8$.

Table B.2: Relative rRMSE of bootstrap averaging in simulation model 3.

DGP	$\tau = 0.2$		$\tau = 0.3$		$\tau = 0.4$		$\tau = 0.6$		$\tau = 0.7$		$\tau = 0.8$	
	BS	CON	BS	CON	BS	CON	BS	CON	BS	CON	BS	CON
1	0.72	0.20	0.78	0.22	0.75	0.26	0.73	0.37	0.67	0.44	0.69	0.59
2	0.81	0.24	0.91	0.24	0.94	0.28	0.90	0.39	0.95	0.40	0.78	0.53
3	0.97	0.27	1.03	0.29	1.03	0.34	1.05	0.38	0.95	0.41	0.86	0.59
4	1.00	0.29	1.04	0.28	0.99	0.32	1.02	0.33	0.92	0.41	0.87	0.47
5	0.93	0.26	0.88	0.44	0.90	0.85	0.96	2.28	0.81	3.00	0.93	3.82
6	0.88	0.29	0.91	0.47	0.90	0.75	0.93	2.01	0.75	2.85	0.92	3.82
7	1.18	0.33	1.14	0.52	1.11	0.77	1.10	1.63	0.81	1.97	0.99	2.18
8	1.24	0.39	1.19	0.66	1.11	0.95	1.29	2.08	0.80	2.82	1.02	3.09
9	0.96	0.34	1.01	0.71	1.00	1.34	0.96	4.10	0.72	5.97	0.94	7.51
10	0.72	0.26	0.80	0.19	0.75	0.19	0.74	0.19	0.80	0.19	0.75	0.23
11	0.84	0.42	0.83	0.28	0.77	0.21	0.74	0.21	0.75	0.32	0.78	0.47
12	0.81	0.30	0.91	0.23	0.97	0.18	1.08	0.15	1.06	0.16	1.07	0.15
13	0.90	0.33	0.97	0.25	1.06	0.19	1.04	0.16	1.02	0.17	1.11	0.18
14	0.83	0.79	0.97	0.41	0.91	0.25	1.08	0.20	1.08	0.23	1.11	0.33

400 replications, 50 bootstraps, sample size 1000. “CON” is the rRMSE of the conservative IVQR estimator; “BS” is the bootstrap averaging relative rRMSE (bootstrap rRMSE divided by IVQR rRMSE).

The “lucky quantile” phenomenon for 2SLS and IVQR-2SLS averaging again applies in simulation model 3. In simulation model 3, the slope is $\gamma(U) = 100c_2U^4$ with $U \sim \text{Unif}(0, 1)$.⁵ The 2SLS population slope is $E[100c_2U^4] + 2c_1c_2 E[100U^5] - 100c_1c_2 E[U^4] = 20c_2 + \frac{40}{3}c_1c_2 \geq (20c_2, 33.4c_2)$. The τ -IVQR slope is $100c_2\tau^4$; since $0.7^4 = 0.24$, the slope is very close to the 2SLS slope when $\tau = 0.7$ ($24c_2$ vs. $(20c_2, 33.4c_2)$).⁶ At a slightly smaller τ (not run in simulations), the IVQR and 2SLS population slopes are identical. So among $\tau = 0.2, 0.3, \dots, 0.8$, the “lucky quantile” is $\tau = 0.7$ in simulation model 3. This is true regardless of c_2 ; it depends mainly on how close τ^4 is to $E(U^4)$. Simulation results confirm that at the 0.7-quantile, 2SLS has relative rRMSE much below 1 even

⁵When both D and Z are binary, and Z affect D monotonically, the 2SLS estimator will converge to $\frac{E[Y|Z=1] - E[Y|Z=0]}{E[D|Z=1] - E[D|Z=0]} = \frac{E[\gamma(U)D|Z=1] - E[\gamma(U)D|Z=0]}{E[E[\gamma(U)D|Z=1,U]] - 0} = \frac{E[\gamma(U)D|Z=1,U]}{E[E[\gamma(U)D|Z=1,U]]} = \frac{E[\gamma(U)[0.5+c_1(U_i-0.5)]g]}{E[0.5+c_1(U_i-0.5)]} = \frac{E[100c_2U^4[0.5+c_1(U_i-0.5)]g]}{E[0.5+c_1(U_i-0.5)]} = E[100c_2U^4] + 2c_1c_2 E[100U^5] - 100c_1c_2 E[U^4] = 20c_2 + \frac{40}{3}c_1c_2 \geq (20c_2, 33.4c_2)$, since $E[U^4] = 1/5$; $E[U^5] = 1/6$ for a standard uniform distributed U , and $c_1 \geq [0, 1]$.

⁶ $24c_2 \geq (20c_2, 33.4c_2)$. $0.6^4 = 0.13$; $0.8^4 = 0.41$; $13c_2$ and $41c_2$ are outside the range $(20c_2, 33.4c_2)$ for the 2SLS population slope.

in the DGPs 5–9 with some or much slope heterogeneity. The IVQR-2SLS averaging estimator also has relative rRMSE below 1 in all the DGPs.

As τ moves away from the “lucky quantile,” the 2SLS estimator begins to show the pattern that its relative rRMSE is below 1 in DGPs with no slope heterogeneity, but above 1 in the DGPs with some or much slope heterogeneity. In DGPs 5–9 with slope heterogeneity, the IVQR-2SLS aggressive estimator’s relative rRMSE can be as large as 12, but that of averaging estimator is still close to 1 (bounded by 1.19). The averaging estimator’s rRMSE is less than 20% worse than IVQR in the least favorable situation.

For the IVQR-QR averaging estimator, the patterns at other quantiles are almost the same as at the median.

Table B.2 presents the bootstrap averaging results at different quantiles. Although bootstrap averaging does not uniformly dominate the conservative IVQR estimator, it offers significant improvement in a variety of DGPs with relatively small downside. Of the 98 relative rRMSEs reported in the BS columns in Tables B.1 and B.2, only five are 1.15 or above, with the largest being 1.29, compared to 30 relative rRMSEs below 0.85.

B.5 Additional simulation results

B.5.1 Simulation model 1 results at different quantile levels

Gaussian error case

Table B.3: Relative rRMSE in simulation model 1 at $\tau = 0.2$.

DGP	Endog	IVQR.2SLS			IVQR.QR			BS	SEE
		AVG	AGG	2SLS	AVG	AGG	QR		
1	0	0.91	0.90	0.70	0.63	0.47	0.46	0.56	0.69
2	0.05	0.90	0.88	0.66	0.66	0.53	0.52	0.57	0.66
3	0.10	0.95	0.95	0.73	0.78	0.81	0.80	0.72	0.75
4	0.15	0.92	0.94	0.68	0.83	1.02	0.99	0.74	0.67
5	0.20	0.95	0.93	0.67	0.89	1.28	1.27	0.77	0.67
6	0.25	0.94	0.91	0.72	0.91	1.61	1.58	0.77	0.73
7	0.30	0.93	0.97	0.68	0.95	1.83	1.79	0.77	0.69
8	0.35	0.95	0.88	0.67	0.96	2.16	2.13	0.77	0.67
9	0.40	0.91	0.93	0.66	0.96	2.56	2.48	0.74	0.66

200 replications. 50 bootstraps. Sample size is 1000.

Table B.4: Relative rRMSE in simulation model 1 at $\tau = 0.3$.

DGP	Endog	IVQR.2SLS			IVQR.QR			BS	SEE
		AVG	AGG	2SLS	AVG	AGG	QR		
1	0	0.93	0.93	0.75	0.63	0.45	0.45	0.59	0.75
2	0.05	0.99	0.96	0.78	0.71	0.60	0.58	0.66	0.79
3	0.10	0.92	0.92	0.78	0.79	0.82	0.81	0.74	0.78
4	0.15	0.93	0.93	0.66	0.84	1.03	1.01	0.73	0.67
5	0.20	0.95	0.97	0.80	0.95	1.47	1.45	0.87	0.80
6	0.25	0.91	0.85	0.70	0.95	1.59	1.57	0.82	0.71
7	0.30	0.95	0.95	0.72	0.96	2.04	2.00	0.80	0.72
8	0.35	0.94	0.96	0.78	0.98	2.48	2.44	0.86	0.79
9	0.40	0.93	0.90	0.71	0.97	2.59	2.53	0.80	0.72

200 replications. 50 bootstraps. Sample size is 1000.

Table B.5: Relative rRMSE in simulation model 1 at $\tau = 0.4$.

DGP	Endog	IVQR.2SLS			IVQR.QR			BS	SEE
		AVG	AGG	2SLS	AVG	AGG	QR		
1	0	0.94	0.94	0.77	0.64	0.48	0.47	0.58	0.85
2	0.05	0.91	0.89	0.76	0.67	0.57	0.56	0.62	0.84
3	0.10	0.89	0.92	0.77	0.76	0.80	0.80	0.68	0.83
4	0.15	0.92	0.96	0.82	0.86	1.24	1.21	0.81	0.89
5	0.20	0.93	0.93	0.76	0.88	1.44	1.44	0.81	0.84
6	0.25	0.89	0.89	0.74	0.94	1.67	1.64	0.82	0.82
7	0.30	0.94	0.91	0.82	0.96	2.19	2.16	0.89	0.88
8	0.35	0.92	0.93	0.74	0.97	2.40	2.36	0.82	0.81
9	0.40	0.91	0.92	0.80	1.01	2.96	2.91	0.89	0.86

200 replications. 50 bootstraps. Sample size is 1000.

Table B.6: Relative rRMSE in simulation model 1 at $\tau = 0.6$.

DGP	Endog	IVQR.2SLS			IVQR.QR			BS	SEE
		AVG	AGG	2SLS	AVG	AGG	QR		
1	0	0.93	0.97	0.79	0.63	0.43	0.42	0.57	0.82
2	0.05	0.91	0.90	0.75	0.67	0.57	0.54	0.63	0.83
3	0.10	0.92	0.94	0.74	0.76	0.80	0.79	0.67	0.82
4	0.15	0.91	0.94	0.78	0.86	1.17	1.16	0.80	0.84
5	0.20	0.93	0.93	0.74	0.93	1.44	1.43	0.82	0.87
6	0.25	0.90	0.89	0.78	0.96	1.78	1.77	0.84	0.86
7	0.30	0.96	0.98	0.81	0.99	2.18	2.14	0.91	0.90
8	0.35	0.92	0.92	0.75	0.97	2.45	2.40	0.83	0.84
9	0.40	0.94	0.93	0.73	0.98	2.85	2.80	0.81	0.80

200 replications. 50 bootstraps. Sample size is 1000.

Table B.7: Relative rRMSE in simulation model 1 at $\tau = 0.7$.

DGP	Endog	IVQR.2SLS			IVQR.QR			BS	SEE
		AVG	AGG	2SLS	AVG	AGG	QR		
1	0	0.94	0.93	0.73	0.67	0.45	0.43	0.64	0.74
2	0.05	0.93	0.93	0.70	0.68	0.56	0.55	0.63	0.70
3	0.10	0.94	0.94	0.70	0.76	0.79	0.79	0.72	0.69
4	0.15	0.91	0.93	0.76	0.89	1.14	1.12	0.77	0.76
5	0.20	0.93	0.92	0.78	0.95	1.49	1.48	0.84	0.77
6	0.25	0.96	0.99	0.74	0.94	1.67	1.67	0.83	0.75
7	0.30	0.91	0.91	0.76	0.96	2.16	2.13	0.82	0.76
8	0.35	0.96	0.98	0.77	1.01	2.59	2.56	0.85	0.76
9	0.40	0.91	0.89	0.72	0.98	2.65	2.58	0.80	0.73

200 replications. 50 bootstraps. Sample size is 1000.

Table B.8: Relative rRMSE in simulation model 1 at $\tau = 0.8$.

DGP	Endog	IVQR.2SLS			IVQR.QR			BS	SEE
		AVG	AGG	2SLS	AVG	AGG	QR		
1	0	0.94	0.90	0.70	0.64	0.45	0.43	0.57	0.69
2	0.05	0.99	1.00	0.75	0.71	0.59	0.58	0.66	0.74
3	0.10	0.94	0.92	0.70	0.75	0.74	0.73	0.67	0.71
4	0.15	0.93	0.94	0.66	0.82	1.01	0.99	0.75	0.67
5	0.20	0.93	0.91	0.73	0.91	1.31	1.29	0.79	0.74
6	0.25	0.93	0.96	0.71	0.93	1.58	1.56	0.82	0.71
7	0.30	0.91	0.94	0.69	0.95	1.89	1.89	0.78	0.70
8	0.35	0.90	0.89	0.72	0.94	2.20	2.14	0.78	0.73
9	0.40	0.95	0.97	0.75	1.00	2.73	2.67	0.83	0.74

200 replications. 50 bootstraps. Sample size is 1000.

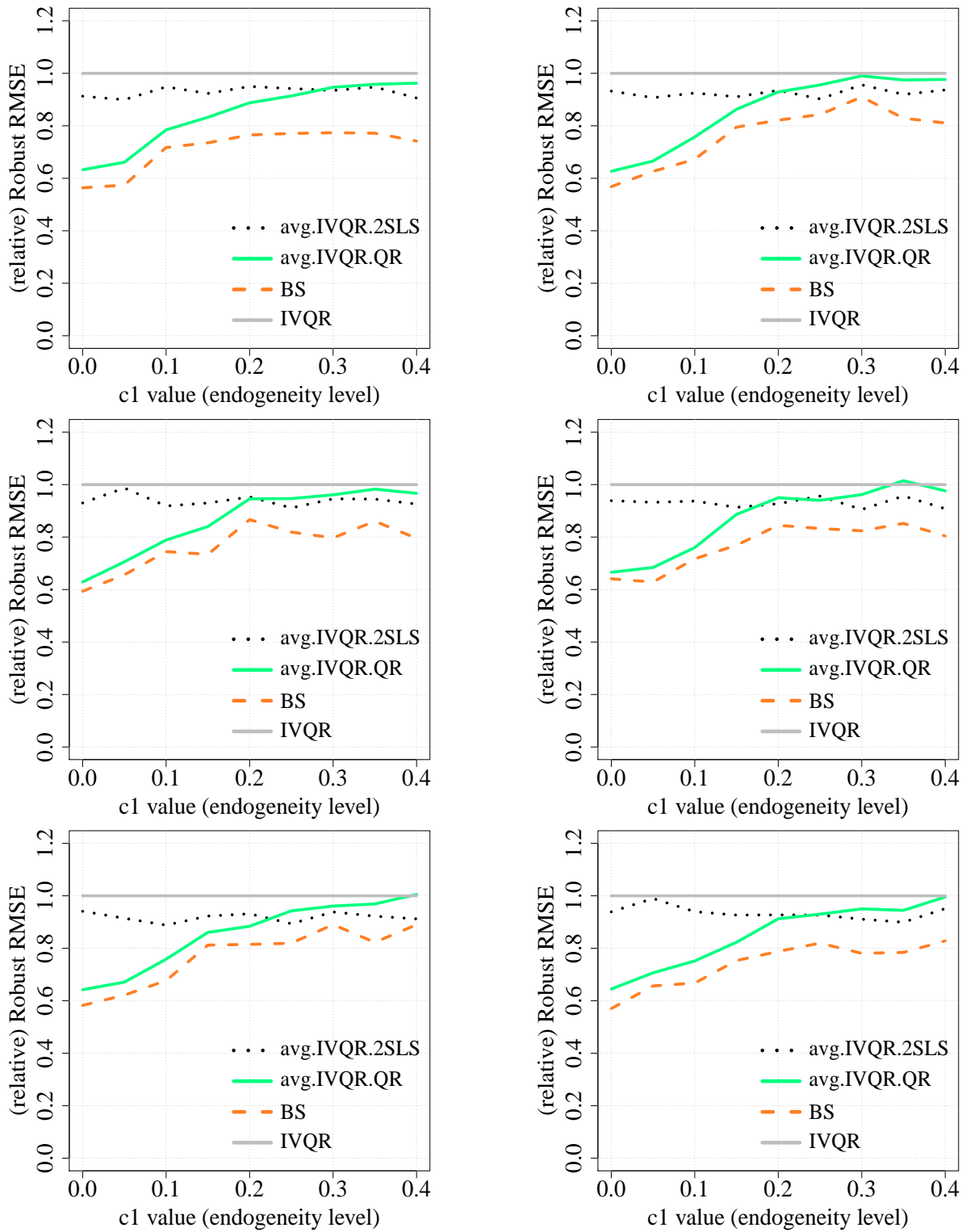


Figure B.1: Relative rRMSE in simulation model 1 (Gaussian error) at $\tau = 0.2$ (left top), $\tau = 0.3$ (left middle), $\tau = 0.4$ (left bottom), $\tau = 0.6$ (right top), $\tau = 0.7$ (right middle), $\tau = 0.8$ (right bottom), based on 200 replications and 50 bootstraps. Sample size $n = 1000$.

Non-Gaussian error case

Table B.9: Relative rRMSE in simulation model 1 (non-Gaussian error) at $\tau = 0.5$ (median).

DGP	Endog	IVQR.2SLS			IVQR.QR			BS	SEE
		AVG	AGG	2SLS	AVG	AGG	QR		
1	0	0.93	0.98	0.83	0.65	0.43	0.42	0.58	0.87
2	0.05	0.95	0.98	0.84	0.68	0.55	0.55	0.61	0.87
3	0.10	0.94	0.99	0.88	0.79	0.83	0.82	0.73	0.87
4	0.15	0.91	0.95	0.85	0.85	1.11	1.09	0.78	0.85
5	0.20	0.95	1.02	0.89	0.91	1.53	1.52	0.86	0.90
6	0.25	0.96	1.01	0.90	0.93	1.74	1.73	0.88	0.92
7	0.30	0.94	0.98	0.88	0.98	2.09	2.07	0.89	0.87
8	0.35	0.97	1.02	0.96	0.97	2.48	2.43	0.91	0.89
9	0.40	0.96	0.97	0.87	0.94	2.79	2.71	0.89	0.89

200 replications. 50 bootstraps. Sample size is 1000.

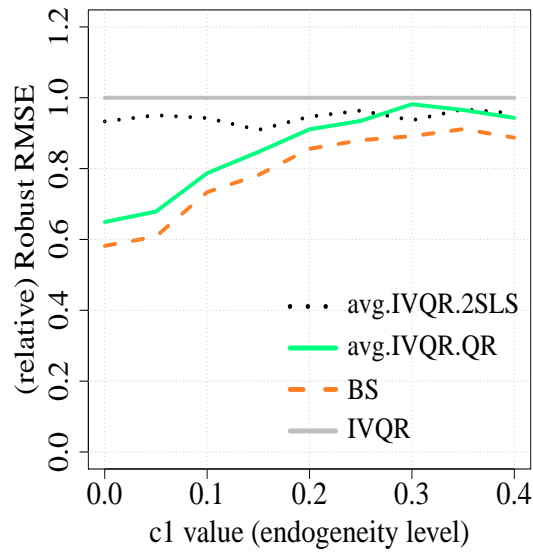


Figure B.2: Relative rRMSE in simulation model 1 (non-Gaussian error) at $\tau = 0.5$ based on 200 replications and 50 bootstraps. Sample size $n = 1000$.

Table B.10: Relative rRMSE in simulation model 1 (non-Gaussian error) at $\tau = 0.2$.

DGP	Endog	IVQR.2SLS			IVQR.QR			BS	SEE
		AVG	AGG	2SLS	AVG	AGG	QR		
1	0	0.99	0.99	1.16	0.61	0.46	0.44	0.64	1.04
2	0.05	1.03	1.06	1.37	0.70	0.62	0.59	0.71	1.20
3	0.10	0.99	1.02	1.28	0.72	0.79	0.76	0.79	1.09
4	0.15	0.98	1.00	1.21	0.82	1.02	0.99	0.85	1.09
5	0.20	0.98	1.03	1.21	0.88	1.22	1.18	0.93	1.06
6	0.25	1.02	1.03	1.30	0.96	1.67	1.63	0.94	1.16
7	0.30	1.01	1.04	1.27	0.95	1.95	1.91	0.94	1.10
8	0.35	0.98	1.00	1.24	0.96	2.35	2.25	0.97	1.09
9	0.40	0.98	0.99	1.33	0.98	2.72	2.60	0.99	1.12

200 replications. 50 bootstraps. Sample size is 1000.

Table B.11: Relative rRMSE in simulation model 1 (non-Gaussian error) at $\tau = 0.3$.

DGP	Endog	IVQR.2SLS			IVQR.QR			BS	SEE
		AVG	AGG	2SLS	AVG	AGG	QR		
1	0	0.97	1.05	1.13	0.63	0.46	0.44	0.63	0.82
2	0.05	1.00	1.05	1.11	0.69	0.59	0.57	0.72	0.91
3	0.10	0.99	1.03	1.08	0.74	0.73	0.72	0.75	0.85
4	0.15	0.98	1.04	1.11	0.85	1.09	1.08	0.85	0.85
5	0.20	1.01	1.05	1.16	0.93	1.49	1.49	0.94	0.93
6	0.25	1.00	1.06	1.11	0.94	1.66	1.65	0.91	0.84
7	0.30	0.99	1.06	1.06	0.96	2.00	1.95	0.96	0.86
8	0.35	0.99	1.07	1.20	0.97	2.45	2.40	1.00	0.89
9	0.40	0.99	1.03	1.09	0.95	2.70	2.61	0.98	0.87

200 replications. 50 bootstraps. Sample size is 1000.

Table B.12: Relative rRMSE in simulation model 1 (non-Gaussian error) at $\tau = 0.4$.

DGP	Endog	IVQR.2SLS			IVQR.QR			BS	SEE
		AVG	AGG	2SLS	AVG	AGG	QR		
1	0	0.94	0.99	1.00	0.64	0.47	0.46	0.61	0.89
2	0.05	0.98	1.06	1.01	0.69	0.56	0.55	0.67	0.87
3	0.10	0.98	1.01	0.94	0.77	0.81	0.78	0.75	0.86
4	0.15	0.98	1.10	1.04	0.86	1.14	1.14	0.85	0.82
5	0.20	0.96	1.07	1.05	0.93	1.47	1.46	0.91	0.86
6	0.25	0.98	1.02	0.96	0.98	1.64	1.63	0.95	0.84
7	0.30	0.97	1.06	1.01	0.94	2.04	2.01	0.93	0.86
8	0.35	0.99	1.04	1.03	0.95	2.33	2.28	0.93	0.84
9	0.40	0.93	1.03	0.98	0.97	2.81	2.72	0.92	0.83

200 replications. 50 bootstraps. Sample size is 1000.

Table B.13: Relative rRMSE in simulation model 1 (non-Gaussian error) at $\tau = 0.6$.

DGP	Endog	IVQR.2SLS			IVQR.QR			BS	SEE
		AVG	AGG	2SLS	AVG	AGG	QR		
1	0	0.90	0.93	0.77	0.65	0.48	0.46	0.59	0.90
2	0.05	0.85	0.85	0.69	0.64	0.54	0.49	0.59	0.79
3	0.10	0.89	0.90	0.74	0.77	0.84	0.84	0.70	0.84
4	0.15	0.90	0.91	0.80	0.85	1.21	1.19	0.80	0.90
5	0.20	0.89	0.86	0.77	0.90	1.40	1.39	0.84	0.86
6	0.25	0.92	0.95	0.78	0.95	1.76	1.74	0.89	0.86
7	0.30	0.92	0.92	0.75	0.97	2.08	2.04	0.88	0.91
8	0.35	0.86	0.82	0.74	0.94	2.27	2.23	0.81	0.81
9	0.40	0.87	0.85	0.78	1.00	2.69	2.63	0.87	0.88

200 replications. 50 bootstraps. Sample size is 1000.

Table B.14: Relative rRMSE in simulation model 1 (non-Gaussian error) at $\tau = 0.7$.

DGP	Endog	IVQR.2SLS			IVQR.QR			BS	SEE
		AVG	AGG	2SLS	AVG	AGG	QR		
1	0	0.82	0.77	0.61	0.63	0.45	0.43	0.56	0.85
2	0.05	0.88	0.85	0.69	0.69	0.56	0.57	0.63	0.93
3	0.10	0.82	0.75	0.63	0.77	0.76	0.76	0.69	0.89
4	0.15	0.80	0.77	0.63	0.81	1.00	0.98	0.70	0.87
5	0.20	0.82	0.82	0.64	0.90	1.38	1.36	0.78	0.89
6	0.25	0.85	0.83	0.68	0.96	1.77	1.74	0.78	0.86
7	0.30	0.84	0.84	0.66	0.99	2.12	2.07	0.85	0.92
8	0.35	0.79	0.74	0.60	0.94	2.03	1.99	0.75	0.87
9	0.40	0.84	0.81	0.65	0.96	2.50	2.42	0.77	0.90

200 replications. 50 bootstraps. Sample size is 1000.

Table B.15: Relative rRMSE in simulation model 1 (non-Gaussian error) at $\tau = 0.8$.

DGP	Endog	IVQR.2SLS			IVQR.QR			BS	SEE
		AVG	AGG	2SLS	AVG	AGG	QR		
1	0	0.82	0.73	0.51	0.64	0.47	0.44	0.52	0.56
2	0.05	0.80	0.72	0.53	0.67	0.53	0.52	0.56	0.58
3	0.10	0.83	0.77	0.53	0.77	0.80	0.77	0.64	0.59
4	0.15	0.83	0.76	0.53	0.84	1.01	1.00	0.68	0.58
5	0.20	0.81	0.73	0.51	0.87	1.28	1.25	0.65	0.57
6	0.25	0.84	0.76	0.51	0.92	1.46	1.43	0.72	0.56
7	0.30	0.80	0.73	0.49	0.94	1.70	1.67	0.68	0.55
8	0.35	0.84	0.76	0.53	1.00	2.08	2.03	0.73	0.59
9	0.40	0.84	0.74	0.54	0.99	2.26	2.17	0.69	0.59

200 replications. 50 bootstraps. Sample size is 1000.

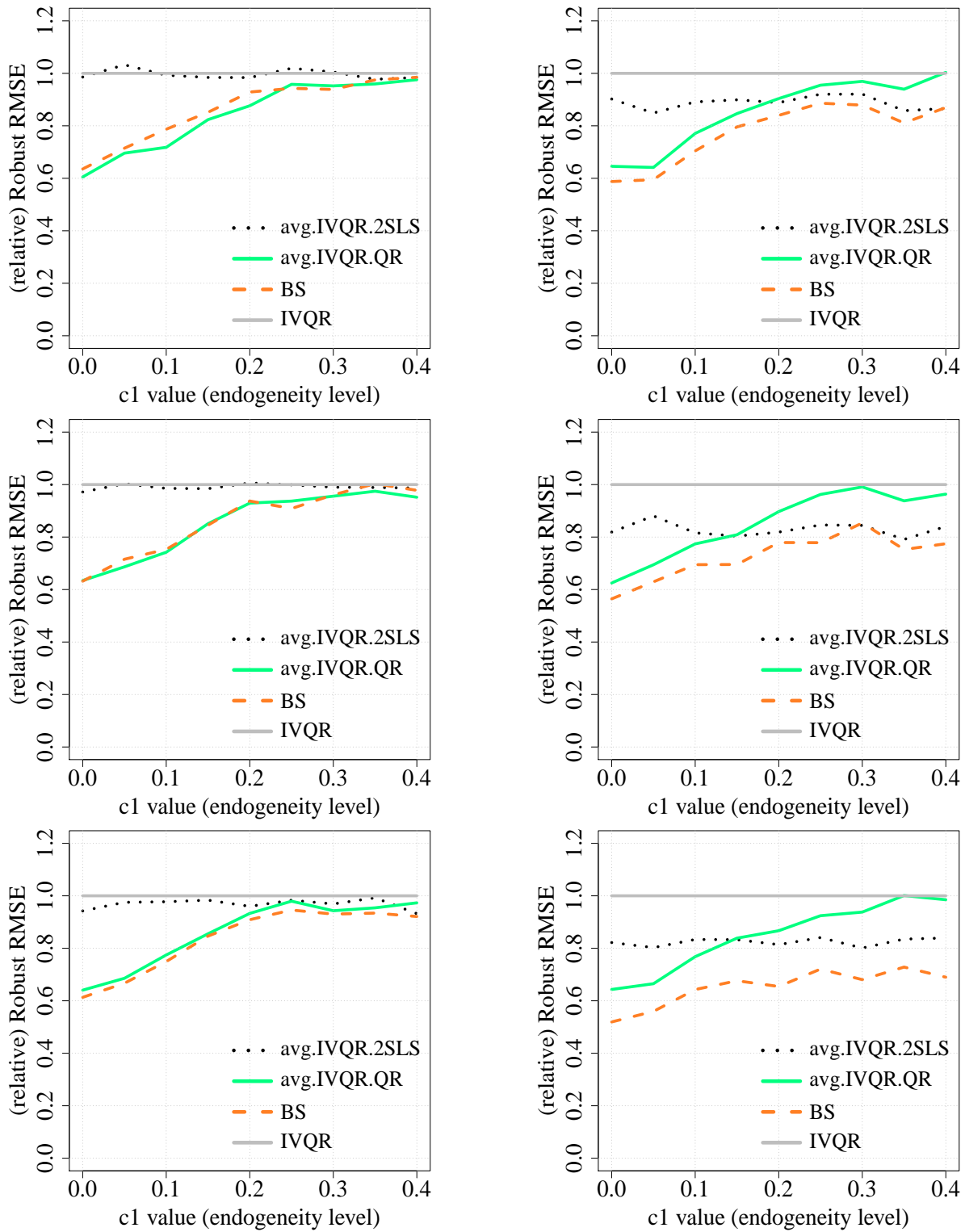


Figure B.3: Relative rRMSE in simulation model 1 (non-Gaussian error) at $\tau = 0.2$ (left top), $\tau = 0.3$ (left middle), $\tau = 0.4$ (left bottom), $\tau = 0.6$ (right top), $\tau = 0.7$ (right middle), $\tau = 0.8$ (right bottom), based on 200 replications and 50 bootstraps. Sample size $n = 1000$.

B.5.2 Simulation model 2 results at different quantile levels

Table B.16 and Table B.17 summarize the upper and lower bounds of relative rRMSE of three averaging estimators at quantile levels from $\tau = 0.2$ to $\tau = 0.8$ with different combinations of endogeneity and heterogeneity level.

Figure B.4, Figure B.5, Figure B.6, Figure B.7, and Figure B.8 provide figure of three averaging estimators' performance as endogeneity and/or heterogeneity level increases at quantile levels from $\tau = 0.2$ to $\tau = 0.8$. The case $\tau = 0.5$ to $\tau = 0.7$ are presented in the main text. Additional tables corresponding to these figures in Appendix B.5.2 are available in the supplemental appendix.

B.5.3 Simulation model 3 results at different quantile levels

Table B.18 and Table B.19 provide the table of the relative rRMSE of the IVQR-2SLS averaging estimators and IVQR-QR averaging estimators at quantile levels from $\tau = 0.2$ to $\tau = 0.8$.

Table B.16: The lower and upper bounds of relative rRMSE in simulation model 2 at three fixed endogeneity and varying heterogeneity.

		Endogeneity Level					
		No ($c_0 = 0$)		Some ($c_0 = 0.2$)		Much ($c_0 = 0.4$)	
		Lower	Upper	Lower	Upper	Lower	Upper
$\tau = 0.2$	$\hat{\theta}_{AVG.2SLS}$	0.970	1.021	0.943	1.044	0.959	1.033
	$\hat{\theta}_{AVG.QR}$	0.822	0.931	0.993	1.018	0.989	1.002
	$\hat{\theta}_{BS}$	0.795	1.065	0.790	1.075	0.815	1.072
$\tau = 0.3$	$\hat{\theta}_{AVG.2SLS}$	0.968	1.021	0.935	1.008	0.955	1.025
	$\hat{\theta}_{AVG.QR}$	0.824	0.890	0.998	1.006	0.991	1.001
	$\hat{\theta}_{BS}$	0.822	1.019	0.787	1.038	0.817	1.063
$\tau = 0.4$	$\hat{\theta}_{AVG.2SLS}$	0.986	1.011	0.949	1.020	0.961	1.011
	$\hat{\theta}_{AVG.QR}$	0.811	0.893	0.996	1.005	0.993	1.001
	$\hat{\theta}_{BS}$	0.796	0.989	0.814	1.054	0.879	1.051
$\tau = 0.5$	$\hat{\theta}_{AVG.2SLS}$	0.922	1.006	0.960	1.024	0.959	0.997
	$\hat{\theta}_{AVG.QR}$	0.831	0.919	0.997	1.002	0.998	1.001
	$\hat{\theta}_{BS}$	0.770	0.929	0.850	1.014	0.860	1.006
$\tau = 0.6$	$\hat{\theta}_{AVG.2SLS}$	0.942	0.999	0.925	0.994	0.955	1.015
	$\hat{\theta}_{AVG.QR}$	0.843	0.929	0.992	1.000	0.996	1.001
	$\hat{\theta}_{BS}$	0.792	0.879	0.798	0.972	0.827	0.978
$\tau = 0.7$	$\hat{\theta}_{AVG.2SLS}$	0.903	0.985	0.914	0.973	0.900	1.002
	$\hat{\theta}_{AVG.QR}$	0.863	0.949	0.988	0.995	0.998	1.000
	$\hat{\theta}_{BS}$	0.794	0.898	0.666	0.841	0.743	0.843
$\tau = 0.8$	$\hat{\theta}_{AVG.2SLS}$	0.928	0.976	0.919	0.982	0.921	0.983
	$\hat{\theta}_{AVG.QR}$	0.877	0.977	0.989	0.996	0.997	1.000
	$\hat{\theta}_{BS}$	0.751	0.943	0.595	0.822	0.660	0.797

200 replications. 50 bootstraps. Sample size is 1000.

Table B.17: The lower and upper bounds of relative rRMSE in simulation model 2 at three fixed heterogeneity and varying endogeneity.

		Heterogeneity Level					
		No ($c_h = 0$)		Some ($c_h = 0.5$)		Much ($c_h = 1$)	
		Lower	Upper	Lower	Upper	Lower	Upper
$\tau = 0.2$	$\hat{\theta}_{AVG.2SLS}$	0.940	0.989	0.970	1.023	0.997	1.035
	$\hat{\theta}_{AVG.QR}$	0.860	1.003	0.841	1.013	0.867	1.014
	$\hat{\theta}_{BS}$	0.776	0.825	1.009	1.075	1.013	1.070
$\tau = 0.3$	$\hat{\theta}_{AVG.2SLS}$	0.930	1.002	0.979	1.006	0.982	1.026
	$\hat{\theta}_{AVG.QR}$	0.850	1.000	0.834	1.004	0.834	1.002
	$\hat{\theta}_{BS}$	0.784	0.858	0.999	1.054	0.970	1.031
$\tau = 0.4$	$\hat{\theta}_{AVG.2SLS}$	0.919	1.012	0.976	1.020	0.985	1.018
	$\hat{\theta}_{AVG.QR}$	0.841	1.010	0.845	1.004	0.881	1.003
	$\hat{\theta}_{BS}$	0.805	0.940	0.964	1.032	0.978	1.033
$\tau = 0.5$	$\hat{\theta}_{AVG.2SLS}$	0.906	0.991	0.973	1.022	0.981	1.016
	$\hat{\theta}_{AVG.QR}$	0.857	1.001	0.876	1.001	0.888	1.001
	$\hat{\theta}_{BS}$	0.803	0.883	0.877	1.004	0.886	1.018
$\tau = 0.6$	$\hat{\theta}_{AVG.2SLS}$	0.890	0.997	0.966	1.003	0.962	1.016
	$\hat{\theta}_{AVG.QR}$	0.843	1.000	0.908	1.000	0.895	1.001
	$\hat{\theta}_{BS}$	0.791	0.894	0.846	0.975	0.752	0.955
$\tau = 0.7$	$\hat{\theta}_{AVG.2SLS}$	0.938	0.992	0.896	0.959	0.962	1.005
	$\hat{\theta}_{AVG.QR}$	0.843	1.000	0.913	0.999	0.930	0.999
	$\hat{\theta}_{BS}$	0.801	0.882	0.683	0.852	0.648	0.880
$\tau = 0.8$	$\hat{\theta}_{AVG.2SLS}$	0.948	1.005	0.930	0.955	0.956	0.989
	$\hat{\theta}_{AVG.QR}$	0.868	1.001	0.969	1.000	0.976	1.000
	$\hat{\theta}_{BS}$	0.766	0.831	0.663	0.908	0.574	0.944

200 replications. 50 bootstraps. Sample size is 1000.

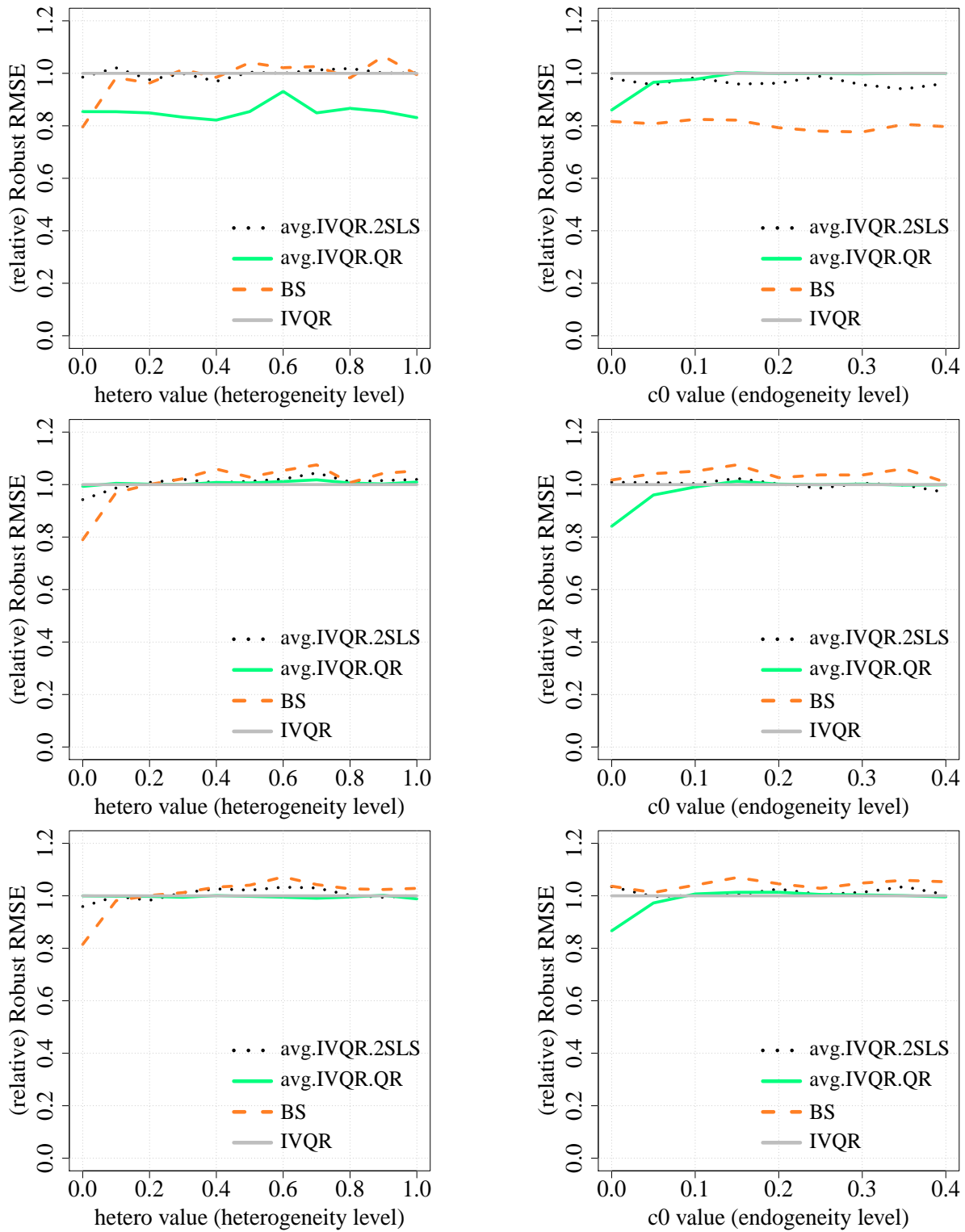


Figure B.4: Relative rRMSE in simulation model 2 at $\tau = 0.2$ quantile level in 6 cases: fixed endogeneity level $c_0 = 0$ (left top), $c_0 = 0.2$ (left middle), $c_0 = 0.4$ (left bottom) and varying heterogeneity; and fixed heterogeneity level $c_h = 0$ (right top), $c_h = 0.5$ (right middle), $c_h = 1$ (right bottom) and varying endogeneity, based on 200 replications and 50 bootstraps. Sample size $n = 1000$.

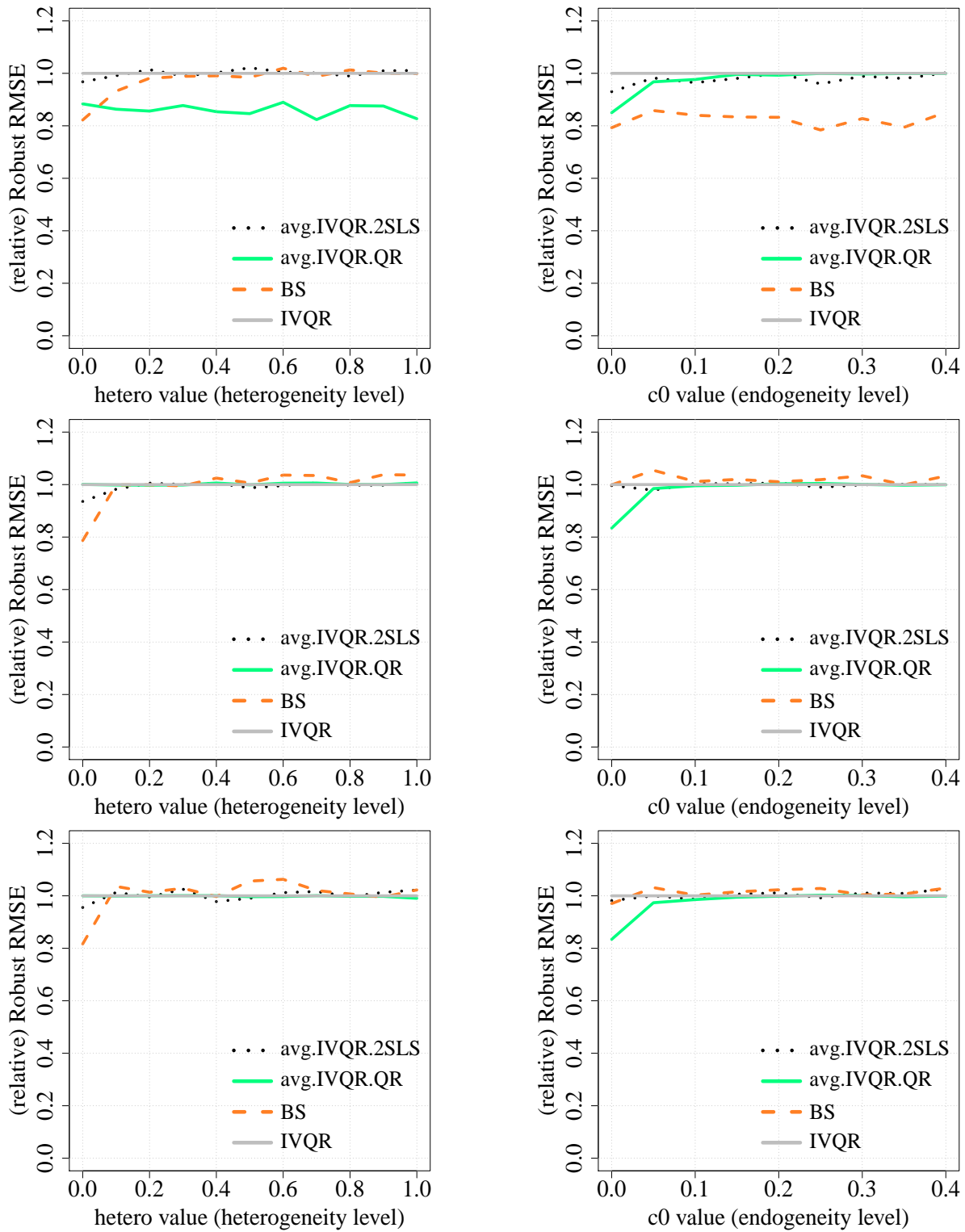


Figure B.5: Relative rRMSE in simulation model 2 at $\tau = 0.3$ quantile level in 6 cases: fixed endogeneity level $c_0 = 0$ (left top), $c_0 = 0.2$ (left middle), $c_0 = 0.4$ (left bottom) and varying heterogeneity; and fixed heterogeneity level $c_h = 0$ (right top), $c_h = 0.5$ (right middle), $c_h = 1$ (right bottom) and varying endogeneity, based on 200 replications and 50 bootstraps. Sample size $n = 1000$.

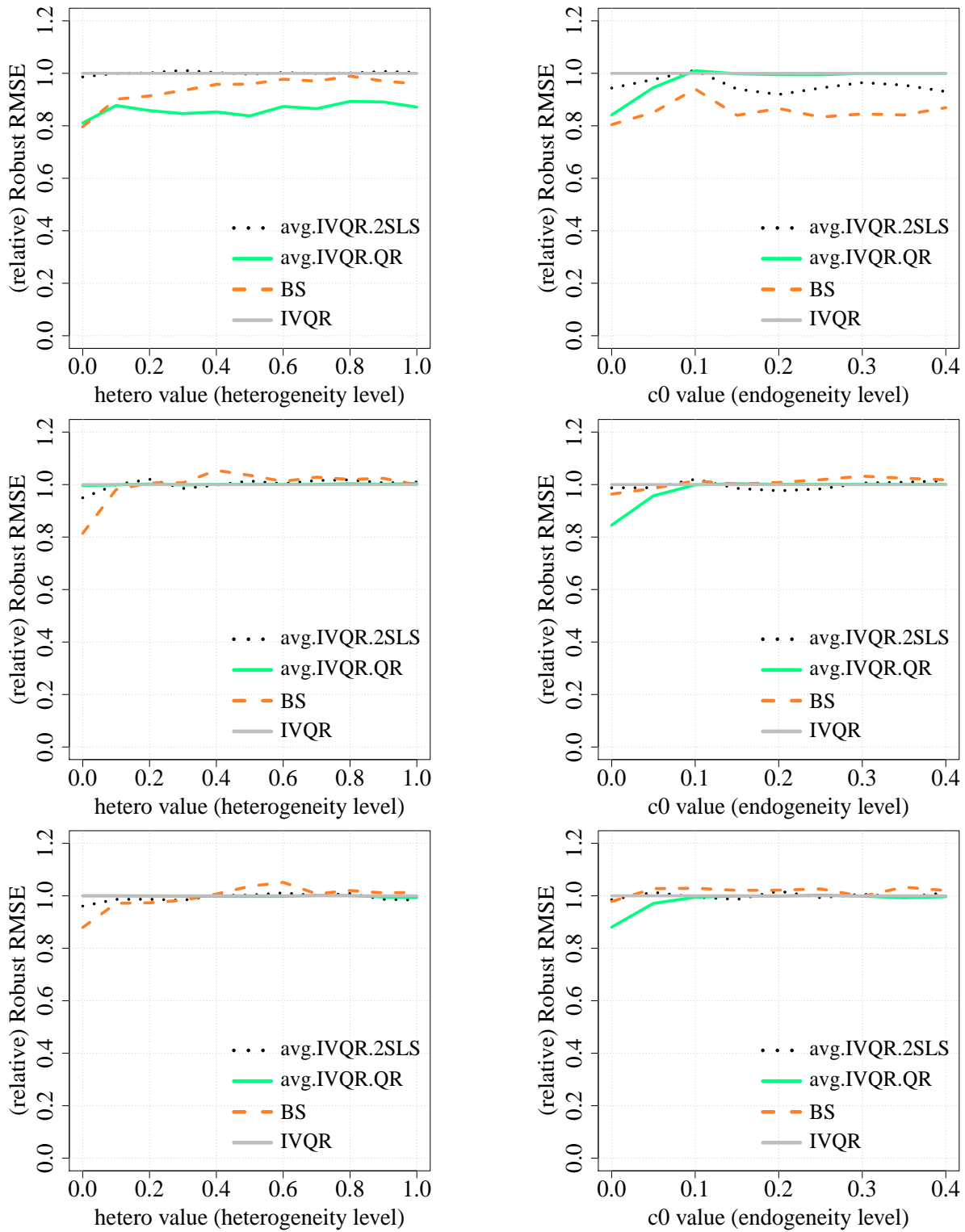


Figure B.6: Relative rRMSE in simulation model 2 at $\tau = 0.4$ quantile level in 6 cases: fixed endogeneity level $c_0 = 0$ (left top), $c_0 = 0.2$ (left middle), $c_0 = 0.4$ (left bottom) and varying heterogeneity; and fixed heterogeneity level $c_h = 0$ (right top), $c_h = 0.5$ (right middle), $c_h = 1$ (right bottom) and varying endogeneity, based on 200 replications and 50 bootstraps. Sample size $n = 1000$.

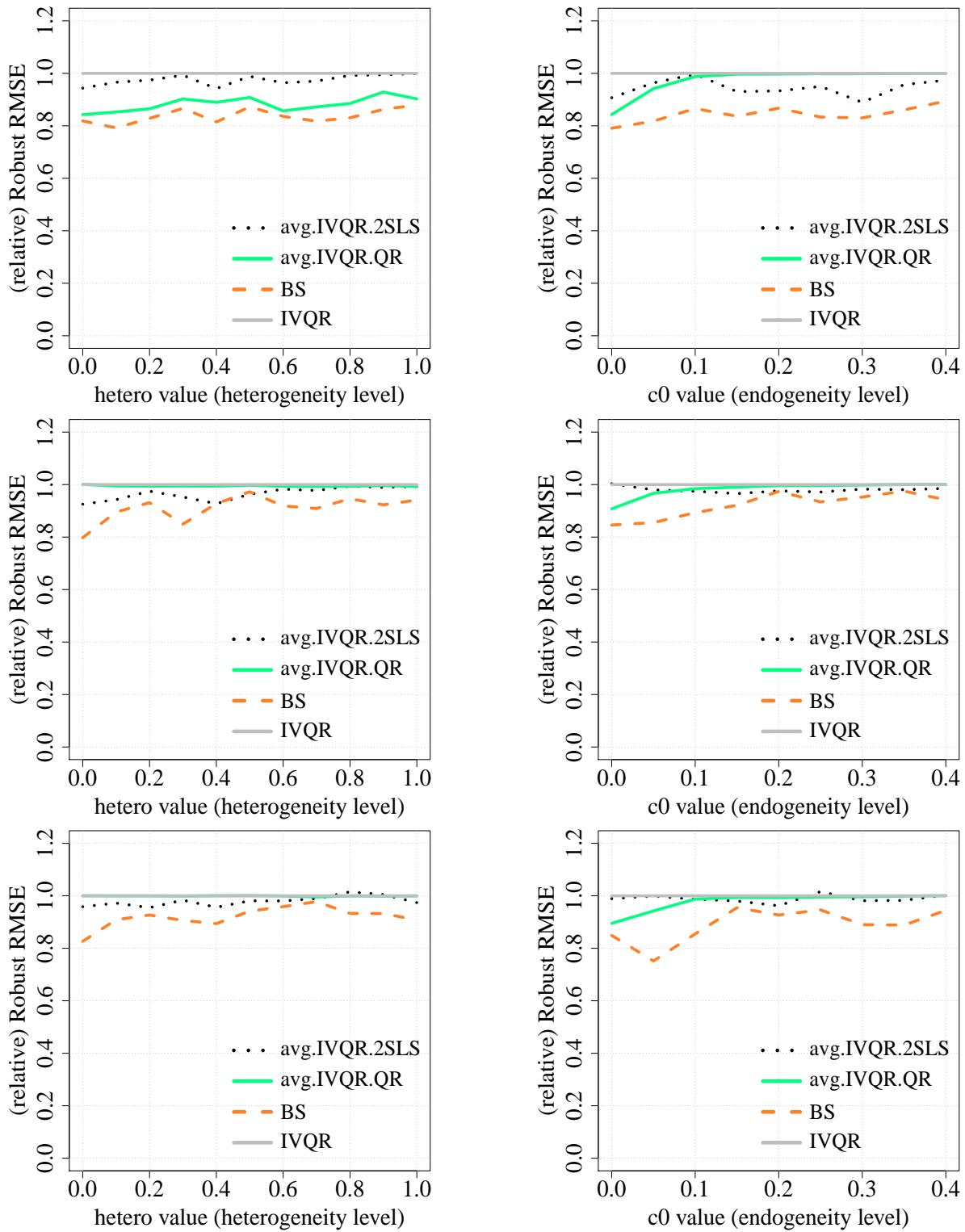


Figure B.7: Relative rRMSE in simulation model 2 at $\tau = 0.6$ quantile level in 6 cases: fixed endogeneity level $c_0 = 0$ (left top), $c_0 = 0.2$ (left middle), $c_0 = 0.4$ (left bottom) and varying heterogeneity; and fixed heterogeneity level $c_h = 0$ (right top), $c_h = 0.5$ (right middle), $c_h = 1$ (right bottom) and varying endogeneity, based on 200 replications and 50 bootstraps. Sample size $n = 1000$.

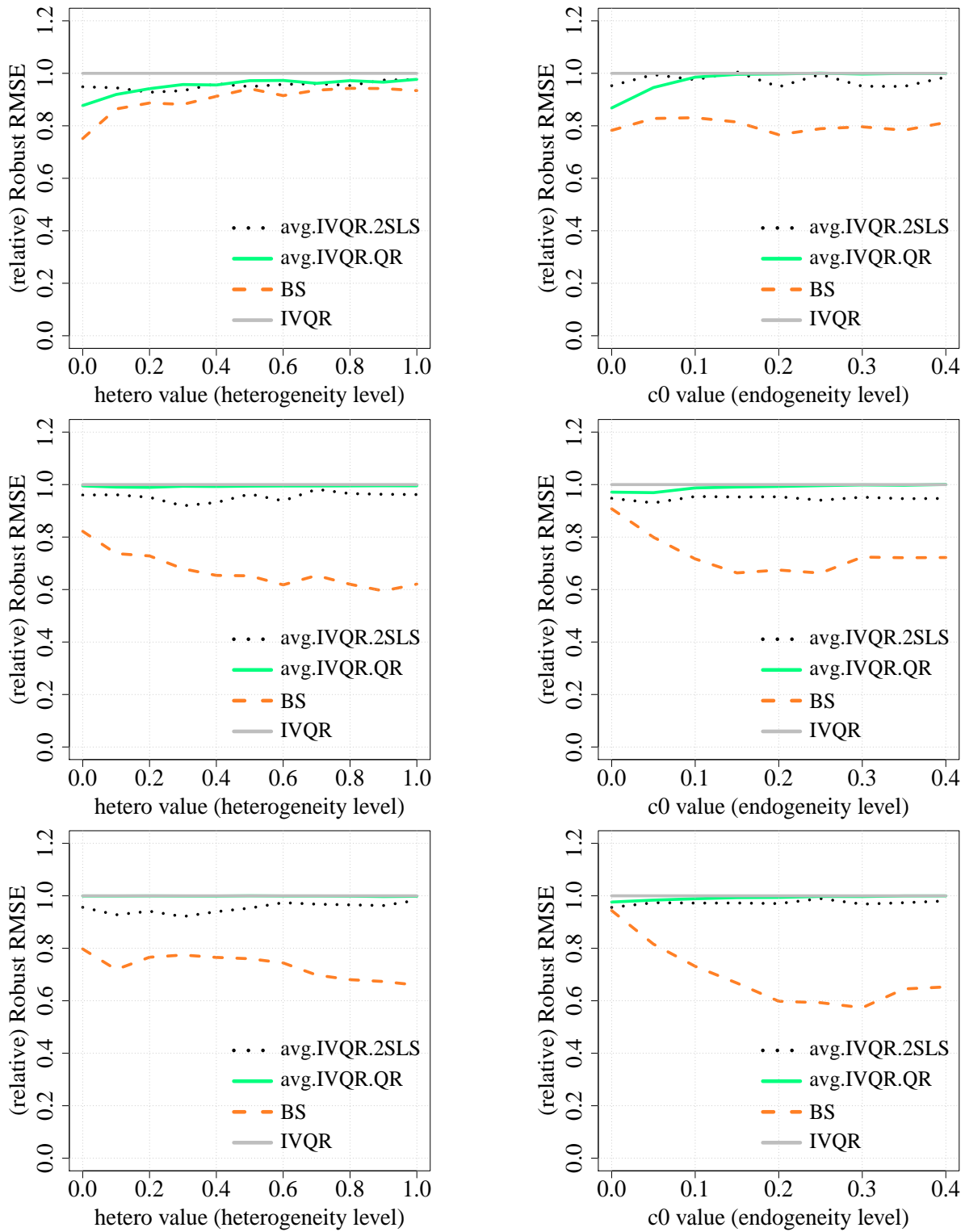


Figure B.8: Relative rRMSE in simulation model 2 at $\tau = 0.8$ quantile level in 6 cases: fixed endogeneity level $c_0 = 0$ (left top), $c_0 = 0.2$ (left middle), $c_0 = 0.4$ (left bottom) and varying heterogeneity; and fixed heterogeneity level $c_h = 0$ (right top), $c_h = 0.5$ (right middle), $c_h = 1$ (right bottom) and varying endogeneity, based on 200 replications and 50 bootstraps. Sample size $n = 1000$.

Table B.18: Relative rRMSE of IVQR-2SLS averaging estimators in simulation model 3.

DGP	$\tau = 0.2$				$\tau = 0.3$				$\tau = 0.4$			
	IVQR.2SLS				IVQR.2SLS				IVQR.2SLS			
	AVG	AGG	2SLS	CON	AVG	AGG	2SLS	CON	AVG	AGG	2SLS	CON
1	0.94	0.87	0.60	0.20	1.00	1.02	1.33	0.22	0.97	0.97	1.13	0.26
2	0.91	0.87	0.80	0.24	0.99	0.99	1.38	0.24	1.01	1.03	1.22	0.28
3	0.99	1.26	1.19	0.27	1.01	0.98	1.12	0.29	0.98	0.95	0.96	0.34
4	0.98	1.68	1.98	0.29	0.98	1.00	1.11	0.28	0.98	0.93	0.96	0.32
5	0.93	0.83	0.61	0.26	1.01	4.37	18.06	0.44	1.04	8.00	8.48	0.85
6	0.95	0.86	0.71	0.29	1.00	7.19	17.31	0.47	1.01	9.46	9.91	0.75
7	1.07	1.98	2.32	0.33	1.03	4.80	14.87	0.52	1.04	8.19	9.34	0.77
8	1.11	2.23	2.52	0.39	1.01	5.69	19.54	0.66	1.02	10.93	12.75	0.95
9	0.88	0.82	0.73	0.34	1.03	12.04	26.19	0.71	1.03	9.96	12.61	1.34
10	0.88	0.81	0.52	0.26	0.99	1.02	1.09	0.19	1.01	1.01	1.11	0.19
11	0.93	0.86	0.64	0.42	1.00	1.00	20.01	0.28	0.99	1.01	25.91	0.21
12	0.99	1.47	1.68	0.30	0.91	0.85	0.69	0.23	0.94	0.89	0.87	0.18
13	0.98	1.51	1.65	0.33	0.95	0.89	0.80	0.25	0.97	1.00	1.05	0.19
14	0.85	1.16	1.17	0.79	1.01	1.00	14.90	0.41	0.99	1.01	23.90	0.25
DGP	$\tau = 0.6$				$\tau = 0.7$				$\tau = 0.8$			
	IVQR.2SLS				IVQR.2SLS				IVQR.2SLS			
	AVG	AGG	2SLS	CON	AVG	AGG	2SLS	CON	AVG	AGG	2SLS	CON
1	0.89	0.77	0.79	0.37	0.77	0.66	0.67	0.44	0.73	0.52	0.54	0.59
2	0.97	0.90	0.87	0.39	0.97	0.85	0.84	0.40	0.79	0.58	0.50	0.53
3	0.91	0.84	0.85	0.38	0.85	0.82	0.80	0.41	0.79	0.60	0.55	0.59
4	0.99	0.98	0.95	0.33	0.87	0.83	0.76	0.41	0.80	0.70	0.70	0.47
5	1.16	1.50	1.35	2.28	0.78	0.63	0.56	3.00	1.10	2.35	2.19	3.82
6	1.17	1.97	1.68	2.01	0.69	0.52	0.52	2.85	1.03	2.23	2.08	3.82
7	1.12	3.07	2.55	1.63	0.81	0.61	0.56	1.97	1.12	2.23	2.00	2.18
8	1.19	3.89	3.33	2.08	0.81	0.71	0.62	2.82	1.13	2.55	2.35	3.09
9	1.19	2.11	1.87	4.10	0.73	0.52	0.51	5.97	1.10	2.53	2.40	7.51
10	0.98	0.90	1.08	0.19	0.99	0.95	1.09	0.19	0.92	0.91	0.90	0.23
11	1.00	1.01	26.26	0.21	0.99	1.02	17.33	0.32	1.01	1.02	11.24	0.47
12	0.99	0.96	1.03	0.15	0.99	0.98	1.01	0.16	0.97	0.91	1.01	0.15
13	1.01	1.02	1.28	0.16	0.99	0.98	1.22	0.17	0.97	1.06	1.34	0.18
14	0.98	0.99	29.94	0.20	1.00	0.98	26.13	0.23	0.99	1.00	18.66	0.33

400 replications. 50 bootstraps. Sample size is 1000.

Table B.19: Relative rRMSE of IVQR-QR averaging estimators in simulation model 3.

DGP	$\tau = 0.2$				$\tau = 0.3$				$\tau = 0.4$			
	IVQR.QR				IVQR.QR				IVQR.QR			
	AVG	AGG	QR	CON	AVG	AGG	QR	CON	AVG	AGG	QR	CON
1	0.94	0.87	0.60	0.20	0.95	0.92	0.64	0.22	0.89	0.87	0.64	0.26
2	0.91	0.87	0.80	0.24	1.00	1.10	1.02	0.24	1.00	1.00	1.08	0.28
3	0.99	1.26	1.19	0.27	1.05	1.35	1.47	0.29	1.05	1.28	1.47	0.34
4	0.98	1.68	1.98	0.29	1.05	1.80	2.51	0.28	0.97	2.01	2.52	0.32
5	0.93	0.83	0.61	0.26	0.88	0.81	0.58	0.44	0.87	0.75	0.57	0.85
6	0.95	0.86	0.71	0.29	0.97	0.90	0.81	0.47	0.92	0.88	0.86	0.75
7	1.07	1.98	2.32	0.33	1.02	2.06	2.74	0.52	1.05	2.75	3.14	0.77
8	1.11	2.23	2.52	0.39	1.11	2.36	2.93	0.66	1.10	3.34	3.60	0.95
9	0.88	0.82	0.73	0.34	0.97	0.90	0.84	0.71	1.01	1.03	0.94	1.34
10	0.88	0.81	0.52	0.26	0.92	0.86	0.58	0.19	0.84	0.82	0.50	0.19
11	0.93	0.86	0.64	0.42	0.93	0.84	0.56	0.28	0.93	0.84	0.57	0.21
12	0.99	1.47	1.68	0.30	0.95	1.45	2.10	0.23	1.01	1.72	2.54	0.18
13	0.98	1.51	1.65	0.33	0.98	1.60	1.92	0.25	1.05	1.88	2.39	0.19
14	0.85	1.16	1.17	0.79	1.00	1.39	1.61	0.41	0.93	1.78	2.16	0.25
DGP	$\tau = 0.6$				$\tau = 0.7$				$\tau = 0.8$			
	IVQR.QR				IVQR.QR				IVQR.QR			
	AVG	AGG	QR	CON	AVG	AGG	QR	CON	AVG	AGG	QR	CON
1	0.91	0.88	0.56	0.37	0.89	0.82	0.54	0.44	0.95	0.88	0.62	0.59
2	1.04	1.11	0.98	0.39	1.04	1.16	1.08	0.40	0.96	1.01	0.97	0.53
3	1.05	1.42	1.73	0.38	1.05	1.47	1.75	0.41	1.01	1.22	1.38	0.59
4	1.06	2.10	3.06	0.33	1.01	1.59	2.62	0.41	1.05	1.73	2.51	0.47
5	0.82	0.71	0.58	2.28	0.87	0.79	0.59	3.00	0.87	0.78	0.64	3.82
6	0.89	0.91	0.82	2.01	0.96	0.94	0.81	2.85	0.89	0.88	0.68	3.82
7	1.01	2.40	2.86	1.63	1.01	2.38	2.79	1.97	1.02	2.08	2.60	2.18
8	1.10	3.34	3.37	2.08	1.05	2.76	3.00	2.82	1.00	2.35	2.72	3.09
9	0.97	0.89	0.84	4.10	0.92	0.87	0.80	5.97	0.89	0.80	0.81	7.51
10	0.91	0.79	0.52	0.19	0.94	0.91	0.59	0.19	0.93	0.83	0.60	0.23
11	0.91	0.80	0.55	0.21	0.92	0.79	0.50	0.32	0.78	0.62	0.74	0.47
12	1.03	1.93	2.81	0.15	1.00	1.34	2.69	0.16	1.03	2.03	2.78	0.15
13	0.99	2.11	2.80	0.16	0.98	1.53	2.74	0.17	1.05	2.01	2.87	0.18
14	1.01	1.98	2.88	0.20	1.03	2.11	3.04	0.23	1.05	2.37	3.17	0.33

400 replications. 50 bootstraps. Sample size is 1000.

Appendix C

Appendix for Chapter 3

C.1 Proofs

C.1.1 Proof of Theorem 5: identification in simplified model

Proof. (i) Under Assumption 18, without loss of generality assume $T = K$ and thus $\underline{\mathbf{X}}_i$ is a squared matrix. (If $T > K$, simply select K rows, which has a row rank K , out of $\underline{\mathbf{X}}_i$. Let $\underline{\mathbf{X}}_i$ be the new $K \times K$ squared matrix.) We can identify all the individual-specific coefficients from its time dimension variations; with probability one, $\underline{\mathbf{X}}_i^{-1}$ exists, so

$$(U_i) = \underline{\mathbf{X}}_i^{-1} \mathbf{Y}_i. \quad (\text{C.1})$$

(ii) Based on Assumption 16 the monotonicity of $\mathbf{x}^\theta(u)$ in u and Assumption 15 the standard uniform distribution of U ,

$$P(\mathbf{x}^\theta(U) \leq \mathbf{x}^\theta(\tau)) = P(U \leq \tau) \quad \text{By Assumption 16} \quad (\text{C.2})$$

$$= \tau. \quad \text{By Assumption 15} \quad (\text{C.3})$$

Therefore, τ is a solution to $P(\mathbf{x}^\theta(U) \leq \mathbf{x}^\theta(\tau)) = \tau$. The uniqueness comes from the fact that $\mathbf{x}^\theta(u)$ is strictly increasing in u from Assumption 16 and that $\mathbf{x}^\theta(u)$ is

continuous in u from Assumption 17. □

C.1.2 Proof of Theorem 6: identification in general model

Proof. The proof will be the same as the Theorem 5 proof of (ii) under Assumptions 14–17. □

C.1.3 Proof of Theorem 7: uniform consistency in simplified model

Proof. Under Assumptions 14–16 and 18 in the simplified model, every individual's coefficient parameter can be perfectly estimated.

$$\hat{\beta}(U_i) = \beta(U_i) = \mathbf{X}_i^{-1} \mathbf{Y}_i.$$

Define the estimator $\hat{\beta}(\tau)$ of the true coefficient at any rank $\tau \in (0, 1)$ as the $\lceil n\tau \rceil$ th order statistic of the n individual coefficients, $\beta(U_1), \beta(U_2), \dots, \beta(U_n)$,

$$\hat{\beta}(\tau) = \beta(U_{n:\lceil n\tau \rceil}).$$

Let $F_U(\cdot)$ denote the distribution function of the standard uniform distributed random variable U . The function $F_U(\cdot)$ is continuous on $[0, 1]$. It can be written out explicitly as

$$F_U(u) = \begin{cases} 0, & u < 0 \\ u, & 0 \leq u \leq 1 \\ 1, & u > 1. \end{cases} \quad (\text{C.4})$$

Denote the τ -quantile of F_U as

$$Q_U(\tau) = F_U^{-1}(\tau) = \inf \{x : F_U(x) \geq \tau\}. \quad (\text{C.5})$$

Both the distribution function and the quantile function of the standard uniform distributed random variable U are identity functions on $(0, 1)$.

$$F_U(u) = u, \text{ for all } u \in (0, 1); \quad (\text{C.6})$$

$$Q_U(\tau) = \tau, \text{ for all } \tau \in (0, 1). \quad (\text{C.7})$$

Thus, $F_U = I$ and $Q_U = I$, where I is the identity function. Define the empirical distribution function of a size n sample of the standard uniform distributed U_1, \dots, U_n as

$$F_{n,U}(u) = \frac{1}{n} \sum_{i=1}^n 1_{\{U_i \leq u\}} \quad u \in \mathcal{G}. \quad (\text{C.8})$$

Define the sample quantile function as the inverse of the empirical distribution function

$$Q_U(\tau) \quad F_{n,U}^{-1}(\tau) = \inf \{x : F_{n,U}(x) \geq \tau\}. \quad (\text{C.9})$$

It is well known that the uniform sample quantile function is uniformly convergent. (See Shorack and Wellner (1986, Thm. 3, p. 95))

$$\sup_{\tau \in (0,1)} |F_{n,U}^{-1}(\tau) - I(\tau)| \xrightarrow{p} 0, \text{ as } n \rightarrow \infty. \quad (\text{C.10})$$

Moreover, the standard uniform distributed U has property that

$$F_{n,U}^{-1}(\tau) = U_{n:n}(\tau). \quad (\text{C.11})$$

Therefore,

$$\sup_{\tau \in (0,1)} |U_{n:n}(\tau) - \tau| = \sup_{\tau \in (0,1)} |F_{n,U}^{-1}(\tau) - I(\tau)| \xrightarrow{p} 0, \text{ as } n \rightarrow \infty. \quad (\text{C.12})$$

That is, the sequence of functions $g_n(\tau) = U_{n:n}(\tau)$ uniformly converges to function $g(\tau) = \tau$ on $(0, 1)$. Equivalently, with probability 1,

$$\forall \epsilon > 0, \exists N_\epsilon \in \mathbb{N} \text{ such that } \forall n > N_\epsilon, \forall \tau \in (0, 1), |U_{n:n}(\tau) - \tau| < \epsilon. \quad (\text{C.13})$$

The uniform continuity assumption Assumption 20 states that for all $k = 1, \dots, K$, $\delta_\epsilon > 0$, $\exists N_{\delta_\epsilon} > 0$, such that $\forall x, y \in (0, 1), |x - y| < \delta_\epsilon \Rightarrow |\beta_k(x) - \beta_k(y)| < \epsilon$. Therefore, for all $k = 1, \dots, K$,

$\exists \delta_\epsilon > 0, \exists N_{\delta_\epsilon} \in \mathbb{N}$ such that $\forall n > N_{\delta_\epsilon}$ and $\forall \tau \in (0, 1)$, $|U_{n:d\tau n e} - \tau| < \delta_\epsilon$, with probability 1,

then

$$|\beta_k(U_{n:d\tau n e}) - \beta_k(\tau)| < \epsilon. \quad (\text{C.14})$$

The sequence of functions $\beta_k(U_{n:d\tau n e})$ uniformly converges almost surely to the corresponding function $\beta_k(\tau)$ as n goes to infinity. For all $k = 1, \dots, K$,

$$\sup_{\tau \in (0, 1)} |\hat{\beta}_k(\tau) - \beta_k(\tau)| = \sup_{\tau \in (0, 1)} |\beta_k(U_{n:d\tau n e}) - \beta_k(\tau)| \xrightarrow{a.s.} 0, \text{ as } n \rightarrow \infty. \quad (\text{C.15})$$

To incorporate the uniform convergence of all the scalar-valued coefficient element function $\beta_k(\cdot)$ into that of the overall vector-valued coefficient function $\hat{\beta}(\cdot)$, consider the supremum norm at first.

$$\|\hat{\beta}(\tau) - \beta(\tau)\| = \|\hat{\beta}(\tau) - \beta(\tau)\|_1 = \max_{k=1, \dots, K} |\hat{\beta}_k(\tau) - \beta_k(\tau)|.$$

On the one hand,

$$\sup_{\tau \in (0, 1)} \max_{k=1, \dots, K} |\hat{\beta}_k(\tau) - \beta_k(\tau)| = \sup_{\tau \in (0, 1)} |\hat{\beta}_k(\tau) - \beta_k(\tau)|, \text{ for each } k = 1, \dots, K.$$

Thus,

$$\sup_{\tau \in (0, 1)} \max_{k=1, \dots, K} |\hat{\beta}_k(\tau) - \beta_k(\tau)| = \max_{k=1, \dots, K} \sup_{\tau \in (0, 1)} |\hat{\beta}_k(\tau) - \beta_k(\tau)|.$$

On the other hand,

$$\sup_{\tau \in (0, 1)} |\hat{\beta}_k(\tau) - \beta_k(\tau)| = |\hat{\beta}_k(\tau) - \beta_k(\tau)|, \text{ for all } 0 < \tau < 1.$$

Thus,

$$\max_{1 \leq k \leq K} \sup_{\tau \in (0,1)} \left| \hat{\beta}_k(\tau) - \beta_k(\tau) \right| = \max_{1 \leq k \leq K} \sup_{\tau \in (0,1)} \left| \hat{\beta}_k(\tau) - \beta_k(\tau) \right|, \text{ for all } k = 1, \dots, K \text{ and } 0 < \tau < 1.$$

Since $\sup_{\tau \in (0,1)} \max_{1 \leq k \leq K} \left| \hat{\beta}_k(\tau) - \beta_k(\tau) \right|$ is the smallest value v such that

$$v \geq \left| \hat{\beta}_k(\tau) - \beta_k(\tau) \right|, \text{ for all } k = 1, \dots, K \text{ and all } 0 < \tau < 1,$$

it should be no greater than other value who satisfy such condition, especially

$$\max_{1 \leq k \leq K} \sup_{\tau \in (0,1)} \left| \hat{\beta}_k(\tau) - \beta_k(\tau) \right| = v = \sup_{\tau \in (0,1)} \max_{1 \leq k \leq K} \left| \hat{\beta}_k(\tau) - \beta_k(\tau) \right|.$$

Altogether,

$$\sup_{\tau \in (0,1)} \max_{1 \leq k \leq K} \left| \hat{\beta}_k(\tau) - \beta_k(\tau) \right| = \max_{1 \leq k \leq K} \sup_{\tau \in (0,1)} \left| \hat{\beta}_k(\tau) - \beta_k(\tau) \right|, \quad (\text{C.16})$$

Therefore,

$$\begin{aligned} \sup_{\tau \in (0,1)} \left\| \hat{\beta}(\tau) - \beta(\tau) \right\|_1 &= \sup_{\tau \in (0,1)} \left\| \hat{\beta}(\tau) - \beta(\tau) \right\|_1 \\ &= \sup_{\tau \in (0,1)} \max_{1 \leq k \leq K} \left| \hat{\beta}_k(\tau) - \beta_k(\tau) \right| \\ &= \max_{1 \leq k \leq K} \sup_{\tau \in (0,1)} \left| \hat{\beta}_k(\tau) - \beta_k(\tau) \right| \\ &= \sup_{\tau \in (0,1)} \left| \hat{\beta}_k(\tau) - \beta_k(\tau) \right|, \text{ for some } k \text{ in } 1, \dots, K \\ &\stackrel{\text{q.s.}}{\rightarrow} 0, \text{ as } n \rightarrow \infty, \text{ by (C.15)}. \end{aligned} \quad (\text{C.17})$$

The vector-valued coefficient function estimator uniformly converges to the true coefficient function on $(0, 1)$.

The uniform consistency of the vector-valued coefficient function estimator can be generalized to any norm, since the coefficient vector is in a finite dimension and all the

norms are equivalent on a finite dimensional space. For example, consider a L_p -norm.

$$\left\| \hat{\beta}(\tau) - \beta(\tau) \right\|_{L_p} = \left(\sum_{k=1}^K \left| \hat{\beta}_k(\tau) - \beta_k(\tau) \right|^p \right)^{1/p}.$$

$$\begin{aligned} \sup_{\tau \in \mathcal{T}(0,1)} \left\| \hat{\beta}(\tau) - \beta(\tau) \right\|_{L_p} &= \sup_{\tau \in \mathcal{T}(0,1)} \left(\sum_{k=1}^K \left| \hat{\beta}_k(\tau) - \beta_k(\tau) \right|^p \right)^{1/p} \\ &= \sup_{\tau \in \mathcal{T}(0,1)} \left(K \max_{1 \leq k \leq K} \left| \hat{\beta}_k(\tau) - \beta_k(\tau) \right|^p \right)^{1/p} \\ &= K^{1/p} \sup_{\tau \in \mathcal{T}(0,1)} \left(\max_{1 \leq k \leq K} \left| \hat{\beta}_k(\tau) - \beta_k(\tau) \right|^p \right)^{1/p} \\ &= K^{1/p} \sup_{\tau \in \mathcal{T}(0,1)} \left(\left[\max_{1 \leq k \leq K} \left| \hat{\beta}_k(\tau) - \beta_k(\tau) \right| \right]^p \right)^{1/p} \\ &= K^{1/p} \sup_{\tau \in \mathcal{T}(0,1)} \max_{1 \leq k \leq K} \left| \hat{\beta}_k(\tau) - \beta_k(\tau) \right| \\ &= K^{1/p} \sup_{\tau \in \mathcal{T}(0,1)} \left\| \hat{\beta}(\tau) - \beta(\tau) \right\|_1 \\ &\stackrel{\text{a.s.}}{\rightarrow} 0, \text{ as } n \rightarrow \infty, \text{ by (C.17)}. \end{aligned} \tag{C.18}$$

Therefore,

$$\sup_{\tau \in \mathcal{T}(0,1)} \left\| \hat{\beta}(\tau) - \beta(\tau) \right\|_{L_p} \stackrel{\text{a.s.}}{\rightarrow} 0, \text{ as } n \rightarrow \infty. \tag{C.19}$$

The vector-valued coefficient function estimator again uniformly converges to the true coefficient function under L_p norm. \square

C.1.4 Proof of Proposition 8: consistency of the estimated permutation

Proof. Let W denote the event that

$$W = \left\{ \frac{\min_{i \neq j} |Y_{n:2} - Y_{n:1}|, \dots, |Y_{n:n} - Y_{n:n-1}|}{2} > \max_{1 \leq i \leq n} \left| \hat{Y}_{T,i} - Y_i \right| \right\}. \tag{C.20}$$

In the event, the left hand side random variable $\min_{n:2} Y_{n:1}, \dots, Y_{n:n} Y_{n:n} \mathcal{G}$ only relates with n rate. The right hand side random variable $\max_{1 \leq i \leq n} |\hat{Y}_{T,i} - Y_i|$ relates with both n rate and T rate. We want to show there exists a sequence of small $\epsilon_n > 0$ such that

$$\mathbb{P} \left(\frac{\min_{n:2} Y_{n:1}, \dots, Y_{n:n} Y_{n:n} \mathcal{G}}{2k_{\mathbf{x}} k} > \epsilon_n > \frac{|\hat{Y}_{T,i} - Y_i|}{k_{\mathbf{x}} k}, \delta_i = 1, \dots, n \right) \rightarrow 1 \quad (\text{C.21})$$

as $n, T \rightarrow \infty$ under the n and T rate relation. This is a sufficient but not necessary condition for (3.24).

Let

$$A = \left\{ \frac{\min_{n:2} Y_{n:1}, \dots, Y_{n:n} Y_{n:n} \mathcal{G}}{2k_{\mathbf{x}} k} > \epsilon_n \right\} \quad (\text{C.22})$$

and

$$B = \left\{ \epsilon_n > \frac{|\hat{Y}_{T,i} - Y_i|}{k_{\mathbf{x}} k}, \delta_i = 1, \dots, n \right\}. \quad (\text{C.23})$$

I want to show

$$\mathbb{P}(A) \rightarrow 1 \text{ as } n \rightarrow \infty,$$

and for small $\delta > 0$,

$$\mathbb{P}(B) \rightarrow 1 \text{ as } n, T \rightarrow \infty \text{ with } n = o(T^{\kappa/(3+\delta)}).$$

If these both hold, then

$$\mathbb{P}(A \text{ and } B) = \mathbb{P}(A) - \mathbb{P}(B) \rightarrow 1, \text{ as } n, T \rightarrow \infty, \text{ with } n = o(T^{\kappa/(3+\delta)}), \quad (\text{C.24})$$

where B is the complement of B .

If $\epsilon_n = c_0 n^{-2-\delta}$ for small $\delta > 0$, then $\mathbb{P}(A) \rightarrow 1$ as $n \rightarrow \infty$:

$$\begin{aligned} \mathbb{P}(A) &= \mathbb{P}(\min_{n:2} Y_{n:1}, \dots, Y_{n:n} Y_{n:n} \mathcal{G} > \epsilon_n 2k_{\mathbf{x}} k) \\ &= 1 - \mathbb{P}(\min_{n:2} Y_{n:1}, \dots, Y_{n:n} Y_{n:n} \mathcal{G} \leq \epsilon_n 2k_{\mathbf{x}} k) \end{aligned}$$

$$\begin{aligned}
& \leq \left\{ 1 - \left[1 - (n+1) \frac{\epsilon_n}{L} \frac{2k_{\mathbf{x}} k}{L} \right]^n \right\} \quad \text{by (C.41)} \\
& = [1 - (n+1)\epsilon_n C]^n, \quad \text{for some positive constant } C = 2k_{\mathbf{x}} k/L, \\
& = [1 - (n+1)n^{-2-\delta} C_1]^n \quad \text{for constant } C_1 = c_0 C \\
& \rightarrow 1 \text{ as } n \rightarrow \infty \text{ by Proposition 22.} \tag{C.25}
\end{aligned}$$

Notice

$$\begin{aligned}
P(B) &= P\left(k_{\mathbf{x}} k c_0 n^{-2-\delta} > \max_{1 \leq i \leq n} |\hat{Y}_{T,i} - Y_i|\right) \\
&= P\left(k_{\mathbf{x}} k c_0 n^{-2-\delta} > k_{\mathbf{x}} k \max_{1 \leq i \leq n} \|\hat{\Lambda}_T(U_i) - \Lambda(U_i)\|\right) \quad \text{by (C.31)} \\
&= P\left(c_0 n^{-2-\delta} > \max_{1 \leq i \leq n} \|\hat{\Lambda}_T(U_i) - \Lambda(U_i)\|\right)
\end{aligned}$$

It remains to show $P(B) \rightarrow 1$ as $n, T \rightarrow \infty$ under the n/T rate relation. It is sufficient to show $\max_{1 \leq i \leq n} k \|\hat{\Lambda}_T(U_i) - \Lambda(U_i)\| = o_p(n^{-2-\delta})$ as $n, T \rightarrow \infty$ under the n/T rate relation.

Let $Z_i = k \|\hat{\Lambda}_T(U_i) - \Lambda(U_i)\|$. By Assumption 23 and the weak law of large numbers,

$$\frac{1}{n} \sum_{i=1}^n Z_i = T^{-\kappa} \left(\frac{1}{n} \sum_{i=1}^n T^{\kappa} Z_i \right) = T^{-\kappa} [E(T^{\kappa} Z_i) + o_p(1)] = O_p(T^{-\kappa}).$$

Thus,

$$\max_{1 \leq i \leq n} Z_i = \sum_{i=1}^n Z_i = n \frac{1}{n} \sum_{i=1}^n Z_i = n O_p(T^{-\kappa}) = O_p(n T^{-\kappa}).$$

The n/T rate relation $n = o(T^{\kappa/(3+\delta)})$ implies $n^{3+\delta}/T^{\kappa} \rightarrow 0$ as $n, T \rightarrow \infty$, or equivalently $T^{-\kappa} = o(n^{-3-\delta})$ as $n, T \rightarrow \infty$. Therefore, $\max_{1 \leq i \leq n} k \|\hat{\Lambda}_T(U_i) - \Lambda(U_i)\| = O_p(n T^{-\kappa}) = o_p(n^{-2-\delta})$. \square

Proposition 22 proves a property which is used in the proof of Proposition 8.

Proposition 22. For some small positive $\delta > 0$,

(i)

$$\left[1 - (n+1) \frac{2k_{\mathbf{x}} k}{L} c_0 n^{-3-\delta} \right]^{n^2} \rightarrow 1, \text{ as } n \rightarrow \infty; \tag{C.26}$$

$$(ii) \quad \left[1 - (n+1) \frac{2k\mathbf{x}}{L} \frac{k}{c_0 n^3} \right]^{n^2} \neq e^{-C}, \text{ as } n \rightarrow \infty; \quad (C.27)$$

$$(iii) \quad \left[1 - (n+1) \frac{2k\mathbf{x}}{L} \frac{k}{c_0 n^{3+\delta}} \right]^{n^2} \neq 0, \text{ as } n \rightarrow \infty, \quad (C.28)$$

where c_0 is a constant, and $C = \frac{2k\mathbf{x}}{L} \frac{k}{c_0}$ is a constant.

Proof. We prove (i) here. The proof of (ii) and (iii) can be obtained in a similar way. The proof of (ii) and (iii) are identical to (i) until the last equality of computing $\lim_{n \rightarrow \infty} \ln a$, at which the results differ depending on $\delta > 0$, $\delta = 0$, or $\delta < 0$ in the derivation of (i).

Let

$$a = \left[1 - (n+1) \frac{2k\mathbf{x}}{L} \frac{k}{c_0 n^{3+\delta}} \right]^{n^2}, \quad (C.29)$$

then

$$\ln a = n^2 \ln \left[1 - (n+1) C n^{-3-\delta} \right], \quad (C.30)$$

where $C = (2k\mathbf{x} k/L)c_0$ is some constant. The limit of $\ln a$ is

$$\begin{aligned} \lim_{n \rightarrow \infty} \ln a &= \lim_{n \rightarrow \infty} \frac{\ln \left[1 - (n+1) C n^{-3-\delta} \right]}{n^{-2}}, \\ &= \lim_{n \rightarrow \infty} \frac{\frac{1}{1 - (n+1) C n^{-3-\delta}} (-C) \left[(2+\delta) n^{-3-\delta} + (3+\delta) n^{-4-\delta} \right]}{2n^{-3}} \quad \text{by L'Hospital's Rule} \\ &= \lim_{n \rightarrow \infty} \frac{\frac{C}{2} \frac{1}{n^\delta} \left[2 + \delta + \frac{3+\delta}{n} \right]}{1 - \frac{C(n+1)}{n^{3+\delta}}} \\ &= 0. \end{aligned}$$

The last equality is because $\frac{1}{n^\delta} \neq 0$, $\left[2 + \delta + \frac{3+\delta}{n} \right] \neq 2 + \delta$, and $\left[1 - \frac{C(n+1)}{n^{3+\delta}} \right] \neq 1$, as $n \rightarrow \infty$. Therefore, by the continuous mapping theorem,

$$\lim_{n \rightarrow \infty} a = \lim_{n \rightarrow \infty} e^{\ln a} = e^{\lim_{n \rightarrow \infty} \ln a} = e^0 = 1. \quad \square$$

C.1.5 Proof of Theorem 9: uniform consistency in general model

Proof. Main idea of proof

To obtain the consistency of the coefficient estimator at any rank in the large- T large- n general model, it requires to ensure two things. One is that in the fixed- n large- T setting, the estimated n individual coefficients should be consistent via its time-dimension observations. The other is that the ordering of the predicted outcome values \hat{Y} should correctly represent the ordering of the unobserved rank variable values.

The first requirement is met by Assumption 21. The second requirement is a non-trivial work to show. Let Y denote the true outcome values at $\mathbf{X} = \mathbf{x}$ computed using the unknown true coefficient. Let \hat{Y} denote the fitted outcome values at $\mathbf{X} = \mathbf{x}$ computed using the estimated individual coefficients. Assumption 16 indicates the ordering of outcome values Y can represent the ordering of unobserved rank variable U . However, we can only obtain the fitted outcome values \hat{Y} instead of the true outcome values Y in the general model. Since the estimated individual coefficient has estimation error in the general model, the ordering of the fitted outcome values \hat{Y} might be different from the ordering of true outcome values Y . If so, the ordering of \hat{Y} cannot represent the ordering of rank variable U .

The idea is to show the fitted outcome value \hat{Y} is close enough to the true outcome value Y . Technically, we hope to show the fitted outcome value \hat{Y} is within some ϵ -ball of the true value Y , with radius less than the smallest value of all $Y_{n:k+1} - Y_{n:k}$, $k = 1, \dots, n-1$. The ϵ -ball between \hat{Y} and Y measures the time-dimension estimation error and it is related with the convergence rate of T . The $fY_{n:k+1} - Y_{n:k} \mathcal{G}_{k=1}^n$'s are random variables with respect to the individual dimension n and they are related with the convergence rate of n . We hope to find certain rate relation between T and n , at which the ϵ -ball shrinks faster than the smallest value of all $Y_{n:k+1} - Y_{n:k}$, $k = 1, \dots, n-1$ does in order to guarantee the order of fitted outcome values \hat{Y} represent the order of

the true outcome values Y and the order of the unobserved rank variable values U .

Time dimension

Let $\hat{\mathbf{x}}_T(U_i)$ denote the TS-OLS estimator computed using its T time periods data for individual i . From Assumption 21 (i), we know $\delta \epsilon > 0$, with probability approaching 1,

$$\begin{aligned} \left| \hat{Y}_{T,i} - Y_i \right| &= \left| \mathbf{x}'^0 \hat{\mathbf{x}}_T(U_i) - \mathbf{x}'^0(U_i) \right| \\ &= \left| \mathbf{x}'^0 \left(\hat{\mathbf{x}}_T(U_i) - \mathbf{x}^0(U_i) \right) \right| \\ &= \|\mathbf{x}^0\| \left\| \hat{\mathbf{x}}_T(U_i) - \mathbf{x}^0(U_i) \right\| \text{ by the Cauchy-Schwarz inequality} \end{aligned} \quad (\text{C.31})$$

$$< k \mathbf{x}^0 k \epsilon, \quad (\text{C.32})$$

where $\|\mathbf{x}\|$ denotes the L_2 norm. That is, for any vector $\mathbf{a} = (a_1, \dots, a_K) \in \mathbb{R}^K$, $\|\mathbf{a}\| = \left(\sum_{i=1}^K a_i^2 \right)^{1/2}$.

Individual dimension

Consider the individual dimension random variable

$$\min_{Y_{n:2}, \dots, Y_{n:n}} \mathcal{G}(Y_{n:2}, \dots, Y_{n:n}).$$

Each element $Y_{n:k+1} - Y_{n:k}$, $k = 1, \dots, n-1$, is a random variable.

$$Y_{n:k+1} - Y_{n:k} = \mathbf{x}'^0(U_{n:k+1}) - \mathbf{x}'^0(U_{n:k}), k = 1, \dots, n-1. \quad (\text{C.33})$$

Take the first order Taylor expansion

$$Y_{n:k+1} - Y_{n:k} = \mathbf{x}'^0(\boldsymbol{\alpha}) \underbrace{(U_{n:k+1} - U_{n:k})}_{\text{Beta}(1,n)}, \quad (\text{C.34})$$

where $\mathbf{x}'^0(\boldsymbol{\alpha})$ is a $K-1$ vector for some $0 < \boldsymbol{\alpha} < 1$. Since $(U_{n:k+1} - U_{n:k}) \stackrel{p}{\rightarrow} 0$, each random variable $(Y_{n:k+1} - Y_{n:k}) \stackrel{p}{\rightarrow} 0$, as $n \rightarrow \infty$. Under Assumption 22, $\mathbf{x}'^0(\boldsymbol{\alpha})$ has a

positive lower bound L . Therefore,

$$\begin{aligned}
Y_{n:k+1} - Y_{n:k} &= \mathbf{x}^\theta(U_{n:k+1}, U_{n:k}) \\
&= \underbrace{\mathbf{x}^\theta(t)}_{L \text{ by Assumption 22}}(U_{n:k+1}, U_{n:k}) \\
&\geq L(U_{n:k+1} - U_{n:k}).
\end{aligned} \tag{C.35}$$

Furthermore,

$$\mathbb{P}(Y_{n:k+1} - Y_{n:k} \leq x) = \mathbb{P}(L(U_{n:k+1} - U_{n:k}) \leq x) = \mathbb{P}\left(U_{n:k+1} - U_{n:k} \leq \frac{x}{L}\right), \tag{C.36}$$

and

$$\begin{aligned}
\mathbb{P}(\min_{n:2} Y_{n:1}, \dots, Y_{n:n} - Y_{n:n-1} \leq g - x) &= \mathbb{P}\left(\bigwedge_{k=1}^{n-1} Y_{n:k+1} - Y_{n:k} \leq x + g\right) \\
&= \mathbb{P}\left(\bigwedge_{k=1}^{n-1} \left\{U_{n:k+1} - U_{n:k} \leq \frac{x}{L}\right\}\right) \\
&= \mathbb{P}\left(\min_{n:2} U_{n:1}, \dots, U_{n:n} - U_{n:n-1} \leq g + \frac{x}{L}\right).
\end{aligned}$$

The order of convergence for random variable $\min_{n:2} Y_{n:1}, \dots, Y_{n:n} - Y_{n:n-1} \leq g$ is bounded by the order of convergence for random variable $\min_{n:2} U_{n:1}, \dots, U_{n:n} - U_{n:n-1} \leq g$.

Let $D_k = U_{n:k+1} - U_{n:k}$, $k = 1, \dots, n-1$. Let $D_0 = U_{n:1}$ and $D_n = 1 - U_{n:n}$. We know from equation (3.10) that D_0, \dots, D_n jointly follows the Dirichlet distribution with $n+1$ parameters 1.

$$(D_0, \dots, D_n) \sim \text{Dir}(\underbrace{1, \dots, 1}_{n+1}).$$

Then we can use the distribution function of $\min_{n:2} D_0, \dots, D_n \leq g$ as the upper bound for the distribution function of the random variable $\min_{n:2} D_1, \dots, D_n \leq g$

$$\mathbb{P}(\min_{n:2} D_1, \dots, D_n \leq g - x) \leq \mathbb{P}(\min_{n:2} D_0, \dots, D_n \leq g - x). \tag{C.37}$$

We can derive the distribution function of $\min\{D_0, \dots, D_n\}$, that for any $x \in [0, \frac{1}{n+1}]$,

$$\begin{aligned} P(\min\{D_0, \dots, D_n\} > x) &= P(D_0 > x, \text{ and } \dots, \text{ and } D_n > x) \\ &= [1 - (n+1)x]^n. \end{aligned} \quad (\text{C.38})$$

$P(\min\{D_0, \dots, D_n\} > x) = 0$, for $x > \frac{1}{n+1}$; $P(\min\{D_0, \dots, D_n\} > x) = 1$, for $x \leq 0$.

The last equality in (C.38) is based on the fact that

$$\begin{aligned} &P(D_0 > x, \text{ and } \dots, \text{ and } D_n > x) \\ &= \int_x^1 \int_x^{1-nx} \int_x^{1-(n-1)x-d_0} \int_x^{1-(n-2)x-d_0-d_1} \dots \int_x^{1-x-d_0-d_1-\dots-d_{n-2}} f_{\text{Dir}}\left(d_0, d_1, \dots, d_{n-1}, 1 - \sum_{k=0}^{n-1} d_k\right) dd_{n-1} \dots dd_0 \\ &= \int_x^1 \int_x^{1-nx} \int_x^{1-(n-1)x-d_0} \int_x^{1-(n-2)x-d_0-d_1} \dots \int_x^{1-x-d_0-d_1-\dots-d_{n-2}} n! dd_{n-1} \dots dd_0 \\ &= n! \int_x^1 \int_x^{1-nx} \int_x^{1-(n-1)x-d_0} \int_x^{1-(n-2)x-d_0-d_1} \dots \int_x^{1-x-d_0-d_1-\dots-d_{n-2}} 1 dd_{n-1} \dots dd_0 \\ &= n! \frac{[1 - (n+1)x]^n}{n!} = [1 - (n+1)x]^n. \end{aligned} \quad (\text{C.39})$$

where $f_{\text{Dir}}(d_0, d_1, \dots, d_{n-1}, 1 - \sum_{k=0}^{n-1} d_k)$ denote the Dirichlet density function with $(n+1)$ parameter $(1, \dots, 1)$.

$$f_{\text{Dir}}\left(d_0, d_1, \dots, d_{n-1}, 1 - \sum_{k=0}^{n-1} d_k; \underbrace{1, \dots, 1}_{n+1}\right) = \frac{[\prod_{k=0}^{n-1} d_k^{1-1}] (1 - \sum_{k=0}^{n-1} d_k)^{1-1}}{\left[\frac{\prod_{k=0}^n (1)}{(n+1)}\right]} = (n+1) = n!, \quad (\text{C.40})$$

and the gamma function with any positive integer n is $\Gamma(n) = (n-1)!$.

The second to last equality in (C.39) is based on a calculus derivation using changing of variables.

$$\begin{aligned} &\int_x^1 \int_x^{1-nx} \int_x^{1-(n-1)x-d_0} \int_x^{1-(n-2)x-d_0-d_1} \dots \int_x^{1-2x-d_0-d_1-\dots-d_{n-3}} \int_x^{1-x-d_0-d_1-\dots-d_{n-2}} 1 dd_{n-1} dd_{n-2} \dots dd_0 \\ &= \int_x^1 \int_x^{1-nx} \int_x^{1-(n-1)x-d_0} \int_x^{1-(n-2)x-d_0-d_1} \dots \int_x^{1-3x-d_0-d_1-\dots-d_{n-4}} \int_x^{1-2x-d_0-d_1-\dots-d_{n-3}} (1-2x-d_0-d_1-\dots-d_{n-2}) dd_{n-2} dd_{n-3} \dots dd_0 \end{aligned}$$

$$\begin{aligned}
& \text{Let } s_2 = 1 - 2x \stackrel{d_0}{=} d_1 \dots d_{n-2} \int_x^1 \int_x^{1-nx} \int_x^{1-(n-1)x-d_0} \dots \int_x^{1-(n-2)x-d_0-d_1} \\
& \int_x^{1-3x-d_0-d_1} \dots \int_x^{1-d_{n-4}} \int_1^0 s_2 ds_2 dd_{n-3} \dots dd_0 \\
& = \frac{1}{2} \int_x^1 \int_x^{1-nx} \int_x^{1-(n-1)x-d_0} \int_x^{1-(n-2)x-d_0-d_1} \int_x^{1-3x-d_0-d_1} \dots \int_x^{1-d_{n-4}} (1-3x-d_0-d_1-\dots-d_{n-3})^2 dd_{n-3} \dots dd_0 \\
& \text{Let } s_3 = 1 - 3x \stackrel{d_0}{=} d_1 \dots d_{n-3} \frac{1}{2} \int_x^1 \int_x^{1-nx} \int_x^{1-(n-1)x-d_0} \int_x^{1-(n-2)x-d_0-d_1} \int_1^0 s_3^2 ds_3 dd_{n-4} \dots dd_0 \\
& = \frac{1}{2 \cdot 3} \int_x^1 \int_x^{1-nx} \int_x^{1-(n-1)x-d_0} \int_x^{1-(n-2)x-d_0-d_1} \int_x^{1-4x-d_0-d_1} \dots \int_x^{1-d_{n-5}} (1-4x-d_0-d_1-\dots-d_{n-4})^3 dd_{n-4} \dots dd_0 \\
& = \dots \\
& = \frac{1}{(n-1)!} \int_x^1 (1-nx-d_0)^{n-1} dd_0 \\
& \text{Let } s = 1 - nx - d_0 \stackrel{1}{=} \frac{1}{(n-1)!} \int_1^0 s^{n-1} ds \\
& = \frac{1}{n!} s^n \Big|_0^{1-(n+1)x} \\
& = \frac{[1-(n+1)x]^n}{n!}.
\end{aligned}$$

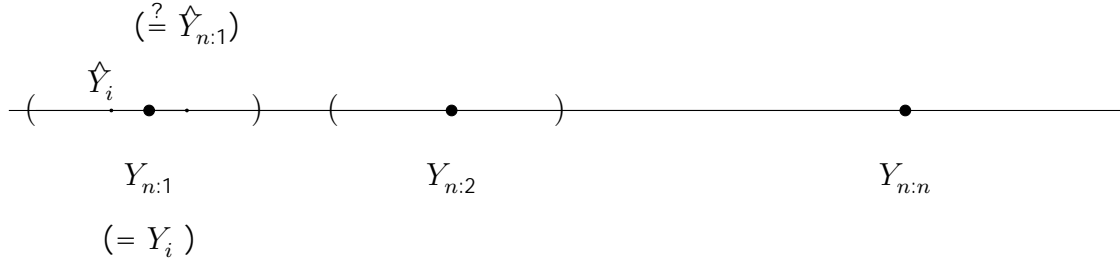
For the special case $x = 0$,

$$\int_0^1 \int_0^1 \int_0^{d_0} \int_0^{d_0-d_1} \dots \int_0^{d_0-d_1-\dots-d_{n-2}} f_{\text{Dir}}\left(d_0, d_1, \dots, d_{n-1}, 1 - \sum_{k=0}^{n-1} d_k\right) dd_{n-1} \dots dd_0 = n! \frac{1}{n!} = 1.$$

Altogether, we can derive the upper bound for the distribution function of the random variable $\min\{Y_{n:2}, Y_{n:1}, \dots, Y_{n:n} - Y_{n:n-1}\}g$. For $x \geq [0, \frac{L}{n+1}]$,

$$\begin{aligned}
P(\min\{Y_{n:2}, Y_{n:1}, \dots, Y_{n:n} - Y_{n:n-1}\}g \leq x) &= P\left(\min\{U_{n:2}, U_{n:1}, \dots, U_{n:n} - U_{n:n-1}\}g \leq \frac{x}{L}\right) \\
&= 1 - [1 - (n+1)x/L]^n. \tag{C.41}
\end{aligned}$$

Correct ordering under rate relation



Proposition 8 implies that when n and T are sufficiently large with the rate relation $n = o(T^{\kappa/(3+\delta)})$, the order of fitted outcome values \hat{Y} can represent the order of the true outcome values Y , which further represents the order of the unobserved rank variable values U . Therefore,

$$\hat{Y}_{n:k} = \mathbf{x}^{\theta \wedge}(U_{n:k}), k = 1, \dots, n. \quad (\text{C.42})$$

The order of the fitted outcome values indicates the order of the rank variable values. Therefore, the estimator $\hat{\tau}$, defined as the coefficient estimates associated with the $d_{n\tau e}$ th order statistic of the fitted outcome values, is actually the coefficient estimates for the person with the $d_{n\tau e}$ th order statistic of rank variable.

$$\hat{\tau} = \hat{\tau}(U_{n:d_{n\tau e}}). \quad (\text{C.43})$$

Consistency

Next we are going to show the estimator is uniform consistent in probability under the n/T rate relation. That is to show, as $n \rightarrow \infty, T \rightarrow \infty$, and $n = o(T^{\frac{\kappa}{3+\delta}})$, for any $\eta > 0$,

$$\mathbb{P}\left(\sup_{\tau \in (0,1)} \left\| \hat{\tau}(\tau) - \tau \right\| > \eta\right) \rightarrow 0. \quad (\text{C.44})$$

Let A denote the event which will result in the correct ordering. (It is the same as event W previously.)

$$\begin{aligned}
A &= \left\{ \frac{\min_{1 \leq i \leq n} f_{Y_{n:2}}(Y_{n:1}, \dots, Y_{n:n})}{2} > \left| \hat{Y}_{T,1} - Y_1 \right|, \text{ and } \dots, \right. \\
&\quad \left. \dots, \text{ and } \frac{\min_{1 \leq i \leq n} f_{Y_{n:2}}(Y_{n:1}, \dots, Y_{n:n})}{2} > \left| \hat{Y}_{T,n} - Y_n \right| \right\} \\
&= \left\{ \frac{\min_{1 \leq i \leq n} f_{Y_{n:2}}(Y_{n:1}, \dots, Y_{n:n})}{2} > \max_{1 \leq i \leq n} \left| \hat{Y}_{T,i} - Y_i \right| \right\}. \tag{C.45}
\end{aligned}$$

From Proposition 8, we know under the n/T rate relation, the event A eventually occurs with probability 1, i.e., as $n, T \rightarrow \infty$, with $n = o\left(T^{\frac{\kappa}{3+\delta}}\right)$,

$$P(A) \rightarrow 1. \tag{C.46}$$

Let A^c denote the complement of event A . Then $P(A^c) \rightarrow 0$, as $n, T \rightarrow \infty$ with $n = o\left(T^{\frac{\kappa}{3+\delta}}\right)$.

For any $\eta > 0$,

$$P\left(\sup_{\tau \in \mathcal{T}(0,1)} \left\| \hat{\tau}(\tau) - \tau \right\| > \eta\right) = \underbrace{P(A)}_{\rightarrow 0} + \underbrace{P(A^c)}_{\rightarrow 1} P\left(\sup_{\tau \in \mathcal{T}(0,1)} \left\| \hat{\tau}(\tau) - \tau \right\| > \eta \mid A^c\right). \tag{C.47}$$

In Appendix C.1.5 we showed under event A , the order of fitted outcome values \hat{Y} correctly represents the order of unobserved rank variable. Then

$$\begin{aligned}
0 & \leq P\left(\sup_{\tau \in \mathcal{T}(0,1)} \left\| \hat{\tau}(\tau) - \tau \right\| > \eta \mid A\right) \\
&= P\left(\sup_{\tau \in \mathcal{T}(0,1)} \left\| \hat{\tau}(U_{n:(\tau n e)}) - \tau \right\| > \eta \mid A\right) \\
&= P\left(\sup_{\tau \in \mathcal{T}(0,1)} \left\| \hat{\tau}(U_{n:(\tau n e)}) - (U_{n:(\tau n e)}) + (U_{n:(\tau n e)}) - \tau \right\| > \eta \mid A\right) \\
& \leq P\left(\sup_{\tau \in \mathcal{T}(0,1)} \left\| \hat{\tau}(U_{n:(\tau n e)}) - (U_{n:(\tau n e)}) \right\| + \sup_{\tau \in \mathcal{T}(0,1)} \left\| (U_{n:(\tau n e)}) - \tau \right\| > \eta \mid A\right)
\end{aligned}$$

$$\begin{aligned}
&= \mathbb{P} \left(\sup_{\tau \geq 2(0,1)} \left\| \begin{matrix} (U_{n:d\tau ne}) \\ (\tau) \end{matrix} \right\| > \eta \mid \sup_{\tau \geq 2(0,1)} \left\| \begin{matrix} (U_{n:d\tau ne}) \\ (U_{n:d\tau ne}) \end{matrix} \right\| \leq \eta \mid A \right) \\
&= \mathbb{P} \left(\sup_{\tau \geq 2(0,1)} \left\| \begin{matrix} (U_{n:d\tau ne}) \\ (U_{n:d\tau ne}) \end{matrix} \right\| \leq \eta \mid A \right) \\
&\quad + \mathbb{P} \left(\sup_{\tau \geq 2(0,1)} \left\| \begin{matrix} (U_{n:d\tau ne}) \\ (U_{n:d\tau ne}) \end{matrix} \right\| > \eta \text{ and} \right. \\
&\quad \left. \sup_{\tau \geq 2(0,1)} \left\| \begin{matrix} (U_{n:d\tau ne}) \\ (\tau) \end{matrix} \right\| > \eta \mid \sup_{\tau \geq 2(0,1)} \left\| \begin{matrix} (U_{n:d\tau ne}) \\ (U_{n:d\tau ne}) \end{matrix} \right\| \leq \eta \mid A \right) \\
&\quad + \mathbb{P} \left(\sup_{\tau \geq 2(0,1)} \left\| \begin{matrix} (U_{n:d\tau ne}) \\ (U_{n:d\tau ne}) \end{matrix} \right\| > \eta \mid A \right) \\
&\quad + \mathbb{P} \left(\sup_{\tau \geq 2(0,1)} \left\| \begin{matrix} (U_{n:d\tau ne}) \\ (\tau) \end{matrix} \right\| > \eta \mid \sup_{\tau \geq 2(0,1)} \left\| \begin{matrix} (U_{n:d\tau ne}) \\ (U_{n:d\tau ne}) \end{matrix} \right\| > 0 \mid A \right).
\end{aligned} \tag{C.48}$$

Let K denote the event $K = \left\{ \sup_{\tau \geq 2(0,1)} \left\| \begin{matrix} (U_{n:d\tau ne}) \\ (U_{n:d\tau ne}) \end{matrix} \right\| \leq \eta \right\}$. Assumption 21 (iii) assumes $\mathbb{P}(K) \neq 0$, as $n \rightarrow \infty, T \rightarrow \infty$. Then,

$$\mathbb{P}(K \mid A) = \frac{\mathbb{P}(K \text{ and } A)}{\mathbb{P}(A)} \stackrel{\neq 0}{\underset{\neq 1}{\frac{\mathbb{P}(K)}{\mathbb{P}(A)}}} \neq 0, \text{ as } n \rightarrow \infty, T \rightarrow \infty, \text{ with } n = o\left(T^{\frac{\kappa}{3+\delta}}\right). \tag{C.49}$$

Let $\gamma = \eta \mid \sup_{\tau \geq 2(0,1)} \left\| \begin{matrix} (U_{n:d\tau ne}) \\ (U_{n:d\tau ne}) \end{matrix} \right\| > 0$. Denote the event

$$E = \left\{ \sup_{\tau \geq 2(0,1)} \left\| \begin{matrix} (U_{n:d\tau ne}) \\ (\tau) \end{matrix} \right\| > \gamma \right\}.$$

From the proof of Theorem 7 (see (C.15)), we know

$$\mathbb{P}(E) \neq 0, \text{ as } n \rightarrow \infty. \tag{C.50}$$

The equality $\mathbb{P}(E) = \mathbb{P}(E \mid A) \mathbb{P}(A) + \mathbb{P}(E \mid A^c) \mathbb{P}(A^c)$ implies

$$\mathbb{P}(E \mid A) = \frac{\mathbb{P}(E) - \mathbb{P}(E \mid A^c) \mathbb{P}(A^c)}{\mathbb{P}(A)}. \tag{C.51}$$

Since $P(E) \neq 0$, $P(A) \neq 1$, and $P(A) \neq 0$ under the n/T rate relation,

$$P(E \neq A) \neq 0, \text{ as } n \neq 1, T \neq 1, \text{ with } n = o\left(T^{\frac{\kappa}{3+\delta}}\right). \quad (\text{C.52})$$

Therefore,

$$\begin{aligned} & 0 \leq P\left(\sup_{\tau \in (0,1)} \left\| \hat{\tau}(\tau) - \tau \right\| > \eta \neq A\right) \\ & \leq P\left(\underbrace{\sup_{\tau \in (0,1)} \left\| \hat{U}_{n:d\tau n e} - U_{n:d\tau n e} \right\| > \eta \neq A}_{=P(K \neq A) \neq 0, \text{ as } n \neq 1, T \neq 1, \text{ with } n = o\left(T^{\frac{\kappa}{3+\delta}}\right) \text{ by (C.49)}}\right) \\ & \quad + P\left(\underbrace{\sup_{\tau \in (0,1)} \left\| \hat{U}_{n:d\tau n e} - \tau \right\| > \eta \neq \sup_{\tau \in (0,1)} \left\| \hat{U}_{n:d\tau n e} - U_{n:d\tau n e} \right\| > 0 \neq A}_{=P(E \neq A) \neq 0, \text{ as } n \neq 1, T \neq 1, \text{ with } n = o\left(T^{\frac{\kappa}{3+\delta}}\right) \text{ by (C.52)}}\right) \\ & \neq 0. \end{aligned} \quad (\text{C.53})$$

Thus,

$$P\left(\sup_{\tau \in (0,1)} \left\| \hat{\tau}(\tau) - \tau \right\| > \eta \neq A\right) \neq 0, \text{ as } n \neq 1, T \neq 1, \text{ with } n = o\left(T^{\frac{\kappa}{3+\delta}}\right). \quad (\text{C.54})$$

And altogether in (C.47)

$$\begin{aligned} P\left(\sup_{\tau \in (0,1)} \left\| \hat{\tau}(\tau) - \tau \right\| > \eta\right) & \leq \underbrace{P(A)}_{\neq 0} + \underbrace{P(A)}_{\neq 1} \underbrace{P\left(\sup_{\tau \in (0,1)} \left\| \hat{\tau}(\tau) - \tau \right\| > \eta \neq A\right)}_{\neq 0} \\ & \neq 0. \end{aligned} \quad (\text{C.55})$$

Therefore, the coefficient function estimator is uniformly consistent to the true coefficient function on $(0, 1)$ as $n \neq 1$, $T \neq 1$, with n and T rate relation $n = o\left(T^{\kappa/(3+\delta)}\right)$. \square

C.1.6 Proof of Theorem 10 : Dirichlet method in simplified model—confidence set of generic coefficient function

Proof.

$$\begin{aligned}
& \mathbb{P}\left(f(U_{n:1}, \beta(U_{n:1})), \dots, (U_{n:n}, \beta(U_{n:n}))g \left\{ \left[\begin{smallmatrix} n \\ k=1 \end{smallmatrix} t_k \right] \right\}\right) \\
&= \mathbb{P}\left(f(U_{n:1}, \beta(U_{n:1})), \dots, (U_{n:n}, \beta(U_{n:n}))g \left\{ \left[\begin{smallmatrix} n \\ k=1 \end{smallmatrix} f(x, y) j \mathbf{a}_k \quad x \quad \mathbf{b}_k; y = \beta(U_{n:k})g \right\}\right) \\
&\quad \mathbb{P}\left((U_{n:1}, \beta(U_{n:1})) \geq t_1, \text{ and } \dots, (U_{n:n}, \beta(U_{n:n})) \geq t_n\right) \\
&= \mathbb{P}\left(U_{n:1} \geq (\mathbf{a}_1, \mathbf{b}_1), \text{ and } \dots, \text{ and } U_{n:n} \geq (\mathbf{a}_n, \mathbf{b}_n)\right) \\
&= 1 - \alpha \text{ by (3.26).}
\end{aligned}$$

The equality establishes when the n individual coefficients $\beta(U_{n:1}), \dots, \beta(U_{n:n})$ are distinct. \square

C.1.7 Proof of Corollary 11: Confidence set for vector-valued generic coefficient function in simplified model

Proof. Consider each element of the vector-valued coefficient function, $\beta_l(\cdot)$, for $l = 1, \dots, K$. From Theorem 10,

$$1 - \alpha = \mathbb{P}\left(f\left[\begin{smallmatrix} n \\ k=1 \end{smallmatrix} (U_{n:k}, \beta_l(U_{n:k}))g \left\{ \left[\begin{smallmatrix} n \\ k=1 \end{smallmatrix} t_k^l \right] \right\}\right). \quad (\text{C.56})$$

Since the uncertainty only comes from the $U_{n:1}, \dots, U_{n:n}$, the n points $\left[\begin{smallmatrix} n \\ k=1 \end{smallmatrix} (U_{n:k}, \beta_l(U_{n:k}))\right]$ on first element function is in the confidence set $\left[\begin{smallmatrix} n \\ k=1 \end{smallmatrix} t_k^1\right]$ if and only if the n points $\left[\begin{smallmatrix} n \\ k=1 \end{smallmatrix} (U_{n:k}, \beta_2(U_{n:k}))\right]$ on second element function is in the confidence set $\left[\begin{smallmatrix} n \\ k=1 \end{smallmatrix} t_k^2\right]$ etc. if and only if the n points $\left[\begin{smallmatrix} n \\ k=1 \end{smallmatrix} (U_{n:k}, \beta_l(U_{n:k}))\right]$ on l th element function is in the confidence set $\left[\begin{smallmatrix} n \\ k=1 \end{smallmatrix} t_k^l\right]$. Let event A_l denote $A_l = \left\{ \left[\begin{smallmatrix} n \\ k=1 \end{smallmatrix} (U_{n:k}, \beta_l(U_{n:k})) \right] \left[\begin{smallmatrix} n \\ k=1 \end{smallmatrix} t_k^l \right] \right\}$. Therefore,

A_1, A_2, \dots, A_K . Then

$$\begin{aligned} \mathbb{P}\left(f\left[\prod_{k=1}^n (U_{n:k}, (U_{n:k}))\right]g \left\{ \left[\prod_{k=1}^n \mathcal{I}_k\right\}\right) &= \mathbb{P}(A_1 \text{ and } A_2, \dots, \text{ and } A_K) \\ &= \mathbb{P}(A_1) \\ &= 1 - \alpha. \end{aligned}$$

This holds with equality when the n individual coefficients $\beta_k(U_{n:1}), \dots, \beta_k(U_{n:n})$ are distinct for each of the element coefficients $k = 1, \dots, K$. \square

C.1.8 Proof of Theorem 12: Dirichlet method in simplified model— uniform confidence band of monotone coefficient function

Proof. Under Assumptions 14–18, with a strict monotone coefficient function, Theorem 10 shows

$$1 - \alpha = \mathbb{P}\left(f\left[\prod_{k=1}^n (U_{n:k}, \beta(U_{n:k}))\right]g \left\{ \left[\prod_{k=1}^n t_k\right\}\right).$$

Let event A denote

$$A = \left\{ f\left[\prod_{k=1}^n (U_{n:k}, \beta(U_{n:k}))\right]g \left\{ \left[\prod_{k=1}^n t_k\right\}\right\}. \quad (\text{C.57})$$

We know

$$\begin{aligned} A &= \left\{ f(U_{n:1}, \beta(U_{n:1})), \dots, (U_{n:n}, \beta(U_{n:n}))g \left\{ \left[\prod_{k=1}^n f(x, y) \mid x \in \mathfrak{b}_k; y = \beta(U_{n:k})g\right\}\right\} \\ &= \left\{ (U_{n:1}, \beta(U_{n:1})) \geq t_1, \text{ and } \dots, \text{ and } (U_{n:n}, \beta(U_{n:n})) \geq t_n \right\}. \end{aligned} \quad (\text{C.58})$$

Under the additional assumption that the coefficient function is monotone, we know the event A implies the event $f(u, y) : y = \beta(u), 0 \leq u \leq 1, g \in S$. Conversely, suppose the coefficient function is all within S , from the fact that the rectangle S_k and S_{k+1} overlap on the interval t_{k+1} , there is no way that the coefficient function is inside the area S but outside the interval t_{k+1} . And in the simplified model, all the coefficients are perfectly

observed. That implies the $(U_{n:k+1}, \beta(U_{n:k+1}))$ is inside the interval t_{k+1} .

Altogether, under the monotone coefficient function condition, the event A and the event $f(u, y) : y = \beta(u), 0 \leq u \leq 1, g \in S$ are equivalent. Therefore,

$$P(f(u, y) : y = \beta(u), 0 \leq u \leq 1, g \in S) = P(A) = 1 - \alpha. \quad (\text{C.59})$$

The area S has exact coverage probability $1 - \alpha$ for the coefficient function $\beta(\cdot)$. \square

C.1.9 Proof of Theorem 13: Dirichlet method in general model— pointwise confidence set of individual coefficient on the generic coefficient function in fixed- n , large- T setting

Proof. Consider an individual whose individual rank variable is the k th smallest value in the size n sample. That is, the individual has rank variable $U_{n:k}$. Let A and B denote the event $A = \{U_{n:k} \in [a_k^{\alpha_1}, b_k^{\alpha_1}]\}$ and $B = \{\beta(U_{n:k}) \geq T_{\sigma(k)}^{\alpha_2}\}$. We know from (3.25) that $P(A) = 1 - \alpha_1$ and from the Assumption 24 that $\lim_{T \rightarrow \infty} P(B) = 1 - \alpha_2$.

Then

$$\begin{aligned} P\left((U_{n:k}, \beta(U_{n:k})) \in [a_k^{\alpha_1}, b_k^{\alpha_1}] \times T_{\sigma(k)}^{\alpha_2}\right) &= P\left(U_{n:k} \in [a_k^{\alpha_1}, b_k^{\alpha_1}] \text{ and } \beta(U_{n:k}) \geq T_{\sigma(k)}^{\alpha_2}\right) \\ &= P(A \cap B) \\ &= P(A) - P(A \setminus B) \\ &= P(A) - P(B). \end{aligned} \quad (\text{C.60})$$

Thus,

$$\begin{aligned} \lim_{T \rightarrow \infty} P\left((U_{n:k}, \beta(U_{n:k})) \in [a_k^{\alpha_1}, b_k^{\alpha_1}] \times T_{\sigma(k)}^{\alpha_2}\right) &= P(A) - \lim_{T \rightarrow \infty} P(B) \\ &= 1 - \alpha_1 - \alpha_2 \\ &= 1 - w\alpha - (1 - w)\alpha = 1 - \alpha. \end{aligned}$$

From Proposition 8, we know the true order of rank variable can be revealed by the order of the fitted outcome values with probability approaching 1, when n goes to infinity slower than T does with the rate relation $n = o\left(T^{\frac{\kappa}{3+\delta}}\right)$. In the fixed- n , large- T setting, it is automatically satisfied that as T goes to infinity, the fitted outcome values \hat{Y}_i become close enough to the true outcome values Y_i , smaller than all the increments $Y_{n:k} - Y_{n:k-1}$ that do not shrink as $T \rightarrow \infty$ when n is fixed. Therefore, the order of fitted outcome values can asymptotically represent the true order of rank variable, and the estimated permutation will equal to the true permutation with probability approaching one:

$$P(\hat{\sigma}(k) = \sigma(k), \forall k = 1, \dots, n) \rightarrow 1, \text{ as } T \rightarrow \infty, \text{ and } n \text{ fixed.} \quad (\text{C.61})$$

We can conservatively bound the coverage probability of the vertical interval based on the estimated permutation,

$$P\left(\beta(U_{n:k}) \geq T_{\hat{\sigma}(k)}^{\alpha_2}\right) = P\left(\beta(U_{n:k}) \geq T_{\sigma(k)}^{\alpha_2}\right) \underbrace{P(\hat{\sigma}(\cdot) \neq \sigma(\cdot))}_{\rightarrow 0, \text{ from (C.61)}} \quad (\text{C.62})$$

Therefore,

$$\begin{aligned} \lim_{T \rightarrow \infty} P\left((U_{n:k}, \beta(U_{n:k})) \in \mathcal{W}_k^{(\alpha_1, \alpha_2)}\right) &= \lim_{T \rightarrow \infty} P\left(U_{n:k} \in [a_k^{\alpha_1}, b_k^{\alpha_1}] \text{ and } \beta(U_{n:k}) \geq T_{\hat{\sigma}(k)}^{\alpha_2}\right) \\ &= \lim_{T \rightarrow \infty} P\left(U_{n:k} \in [a_k^{\alpha_1}, b_k^{\alpha_1}] \text{ and } \beta(U_{n:k}) \geq T_{\sigma(k)}^{\alpha_2}\right) + o(1) \\ &= 1 - \alpha. \end{aligned} \quad (\text{C.63})$$

That is, the confidence set $\mathcal{W}_k^{(\alpha_1, \alpha_2)}$ has at least $1 - \alpha$ asymptotic coverage probability for the individual coefficient. \square

C.1.10 Proof of Theorem 14: Dirichlet method in general model—
joint confidence sets for all n individual coefficients on the
generic coefficient function in fixed- n , large- T setting

Proof. For any $0 < \alpha_1 < 1$ and $0 < \alpha_2 < 1$, denote the event A_0, \dots, A_n as

$$A_0 = \left\{ U_{n:1} \geq [\underline{a}_1^{\alpha_1}, \underline{b}_1^{\alpha_1}] \text{ and } \dots \text{ and } U_{n:n} \geq [\underline{a}_n^{\alpha_1}, \underline{b}_n^{\alpha_1}] \right\},$$

$$A_k = \left\{ \beta(U_{n:k}) \geq T_{\sigma(k)}^{\alpha_2} \right\}, \quad k = 1, \dots, n.$$

From (3.26), $P(A_0) = 1 - \alpha_1$ and $P(A_k) = 1 - \alpha_2$. Define events B_i

$$B_i = \left\{ f\beta(U_i) \geq T_i^{\alpha_2} g, \quad i = 1, \dots, n. \right.$$

Assumption 24 implies

$$P(B_i) \rightarrow 1 - \alpha_2, \text{ as } T \rightarrow \infty. \quad (\text{C.64})$$

Denote the event

$$B = A_1 \cap \dots \cap A_n.$$

Then

$$\begin{aligned} P(B) &= P(A_1 \cap \dots \cap A_n) \\ &= P\left(\bigcap_{k=1}^n f\beta(U_{n:k}) \geq T_{\sigma(k)}^{\alpha_2} g\right) \\ &= P\left(\bigcap_{i=1}^n f\beta(U_i) \geq T_i^{\alpha_2} g\right) \\ &= P(B_1 \cap \dots \cap B_n) \\ &= P(B_1) \dots P(B_n) \text{ since the } B_i \text{ are independent across individuals} \\ &\rightarrow (1 - \alpha_2)^n, \text{ as } T \rightarrow \infty. \end{aligned}$$

Thus, $P(B) \rightarrow (1 - \alpha_2)^n$ as $T \rightarrow \infty$.

Altogether,

$$\begin{aligned}
& \mathbb{P}\left(\left\{U_{n:1} \geq [\mathfrak{a}_1^{\alpha_1}, \mathfrak{b}_1^{\alpha_1}] \text{ and } U_{n:n} \geq [\mathfrak{a}_n^{\alpha_1}, \mathfrak{b}_n^{\alpha_1}]\right\} \setminus \left\{\bigwedge_{k=1}^n \left\{\beta(U_{n:k}) \geq T_{\sigma(n)}^{\alpha_2}\right\}\right\}\right) \\
&= \mathbb{P}(A_0 \setminus A_1 \setminus \dots \setminus A_n). \\
&= \mathbb{P}(A_0 \setminus B) \\
&= 1 - \mathbb{P}(A_0) - \mathbb{P}(B).
\end{aligned}$$

Asymptotically,

$$\begin{aligned}
& \lim_{T \uparrow 1} \mathbb{P}\left(\left\{U_{n:1} \geq [\mathfrak{a}_1^{\alpha_1}, \mathfrak{b}_1^{\alpha_1}] \text{ and } U_{n:n} \geq [\mathfrak{a}_n^{\alpha_1}, \mathfrak{b}_n^{\alpha_1}]\right\} \setminus \left\{\bigwedge_{k=1}^n \left\{\beta(U_{n:k}) \geq T_{\sigma(n)}^{\alpha_2}\right\}\right\}\right) \\
&= 1 - \mathbb{P}(A_0) - \lim_{T \uparrow 1} \mathbb{P}(B) \\
&= 1 - \alpha_1 - (1 - (1 - \alpha_2)^n) \\
&= 1 - \alpha/2 - \alpha/2 = 1 - \alpha.
\end{aligned}$$

Proposition 8 implies that as $T \uparrow 1$ and n staying fixed, the estimated permutation equals the true permutation with probability approaching one. Thus, the probability of the n individual coefficients inside the area $R^{(\alpha_1, \alpha_2)}$ can be bounded below,

$$\begin{aligned}
& \mathbb{P}\left(f\left[\bigwedge_{i=1}^n (U_i, \beta(U_i))\right]g \mid R^{(\alpha_1, \alpha_2)}\right) \\
&= \mathbb{P}\left(f\left[\bigwedge_{k=1}^n (U_{n:k}, \beta(U_{n:k}))\right]g \mid \left\{\bigwedge_{k=1}^n R_k^{(\alpha_1, \alpha_2)}\right\}\right) \\
&= \mathbb{P}\left(f(U_{n:1}, \beta(U_{n:1})), \dots, (U_{n:n}, \beta(U_{n:n}))g \mid \left\{\bigwedge_{k=1}^n \left\{[\mathfrak{a}_k^{\alpha_1}, \mathfrak{b}_k^{\alpha_1}] \leq T_{\hat{\sigma}(k)}^{\alpha_2}\right\}\right\}\right) \\
&= \mathbb{P}\left(U_{n:1} \geq [\mathfrak{a}_1^{\alpha_1}, \mathfrak{b}_1^{\alpha_1}] \text{ and } \dots, U_{n:n} \geq [\mathfrak{a}_n^{\alpha_1}, \mathfrak{b}_n^{\alpha_1}] \text{ and } \beta(U_{n:1}) \geq T_{\hat{\sigma}(1)}^{\alpha_2} \text{ and } \dots, \beta(U_{n:n}) \geq T_{\hat{\sigma}(n)}^{\alpha_2}\right) \\
&= \mathbb{P}\left(\left\{U_{n:1} \geq [\mathfrak{a}_1^{\alpha_1}, \mathfrak{b}_1^{\alpha_1}] \text{ and } \dots, U_{n:n} \geq [\mathfrak{a}_n^{\alpha_1}, \mathfrak{b}_n^{\alpha_1}]\right\} \setminus \left\{\bigwedge_{k=1}^n \left\{\beta(U_{n:k}) \geq T_{\hat{\sigma}(n)}^{\alpha_2}\right\}\right\}\right) \\
&= \mathbb{P}\left(\left\{U_{n:1} \geq [\mathfrak{a}_1^{\alpha_1}, \mathfrak{b}_1^{\alpha_1}] \text{ and } \dots, U_{n:n} \geq [\mathfrak{a}_n^{\alpha_1}, \mathfrak{b}_n^{\alpha_1}]\right\} \setminus \left\{\bigwedge_{k=1}^n \left\{\beta(U_{n:k}) \geq T_{\sigma(n)}^{\alpha_2}\right\}\right\}\right) \\
&= \underbrace{\mathbb{P}(\hat{\sigma}(\cdot) \neq \sigma(\cdot))}_{! 0, \text{ as } T \uparrow 1, \text{ and } n \text{ fixed}}. \tag{C.65}
\end{aligned}$$

Thus,

$$\lim_{T \rightarrow \infty} \mathbb{P}(f[\frac{1}{n} \sum_{i=1}^n (U_i, \beta(U_i))] \in R^{(\alpha_1, \alpha_2)}) \\ \lim_{T \rightarrow \infty} \mathbb{P}\left(\left\{U_{n:1} \geq [\mathbf{a}_1^{\alpha_1}, \mathbf{b}_1^{\alpha_1}] \text{ and } \dots, U_{n:n} \geq [\mathbf{a}_n^{\alpha_1}, \mathbf{b}_n^{\alpha_1}]\right\} \setminus \left\{\bigcap_{k=1}^n \left\{\beta(U_{n:k}) \geq T_{\sigma(n)}^{\alpha_2}\right\}\right\}\right) \\ 1 - \alpha. \quad (\text{C.66})$$

Thus, the confidence set $R^{(\alpha_1, \alpha_2)}$ with $\alpha_1 = \alpha/2$ and $\alpha_2 = 1 - (\alpha/2)^{1/n}$ has asymptotic coverage probability bounded below by $1 - \alpha$. \square

C.1.11 Lemma 23: asymptotic normality of empirical process

Assume a general version of Assumption 25 about the differentiability of the coefficient function. Under this assumption, we build the uniform asymptotic normality of the empirical process as the coefficient function composited with any general distribution function. This provide the intermediate results for Theorem 15, which is the special case that applies to standard uniform distribution.

Assumption 26. Consider the K -vector coefficient function $\beta : (s, t) \in \mathbb{R} \rightarrow \mathbb{R}^K$. Assume each of its element function $\beta_k : (s, t) \in \mathbb{R} \rightarrow \mathbb{R}$, $k = 1, \dots, K$, is differentiable with uniform continuous and bounded derivative.

Lemma 23 (Asymptotic normality of empirical process). *Let F be a distribution function that have compact support $[s, t]$ and be continuously differentiable on its support with strictly positive derivative f . Let F_n be the empirical distribution of an iid sample from F with size n . Suppose Assumption 26 holds. Then the sequence of empirical processes $\sqrt{n}(\beta_k(F_n) - \beta_k(F))$ converges in distribution in the space $\ell^1(0, 1)$ to a limit Gaussian process*

$$\sqrt{n}(\beta_k(F_n) - \beta_k(F)) \xrightarrow{D} \beta_k^0(F) \frac{G}{f(F)}, \text{ in } \ell^1(0, 1), \text{ as } n \rightarrow \infty, \quad (\text{C.67})$$

where G is the standard Brownian bridge. The limit process is Gaussian with mean zero and covariance function

$$\beta_k^\theta(F^{-1}(u))\beta_k^\theta(F^{-1}(v))\frac{u \wedge v - uv}{f(F^{-1}(u))f(F^{-1}(v))}, \quad u, v \in \mathbb{R}.$$

Proof. Let $D[s, t]$ denote the Skorohod space of all cadlag functions on the interval $[s, t] \subset \mathbb{R}$ with uniform norm. Let $C[s, t]$ denote the space of all continuous functions on $[s, t]$. Let $\ell^1(T)$ denote the collection of all bounded functions $f : T \rightarrow \mathbb{R}$ with the norm $\|f\|_1 = \sup_{t \in T} |f(t)|$.

Let D_ϕ be the set of all distribution functions with measures concentrated on $[s, t]$. Van der Vaart and Wellner (1996, Lemma 3.9.23(ii)) show that the inverse map $G \mapsto G^{-1}$ as a map $\phi : D_\phi \subset D[s, t] \rightarrow \ell^1(0, 1)$ is Hadamard-differentiable at F tangentially to $C[s, t]$. The derivative is

$$\phi_F^\theta(\alpha) = (\alpha/f) \circ F^{-1} = \frac{\alpha \circ F^{-1}}{f \circ F^{-1}},$$

for any function $\alpha \in D[s, t]$.

Consider the map $\psi : \ell^1(0, 1) \rightarrow \ell^1(0, 1)$ given by $\psi(A)(x) = \beta_k(A(x)) = \beta_k \circ A(x)$. The domain of this map is the set of elements of $\ell^1(0, 1)$ that take their values in the domain of β_k . Let $D_\psi = \{A : A \in \ell^1(0, 1) : s < A < tg\}$. By van der Vaart and Wellner (1996, Lemma 3.9.25), we know the map $A \mapsto \beta_k \circ A$ as a map $D_\psi \subset \ell^1(0, 1) \rightarrow \ell^1(0, 1)$ is Hadamard-differentiable at every $A \in D_\psi$. The derivative is $\psi_A^\theta(\alpha) = \beta_k^\theta(A(x))\alpha(x)$.

Therefore, the map $F^{-1} \mapsto \beta_k \circ F^{-1}$ as a map $\psi : D_\psi \subset \ell^1(0, 1) \rightarrow \ell^1(0, 1)$ is Hadamard-differentiable at every $F^{-1} \in D_\psi$. The derivative is $\psi_{F^{-1}}^\theta(\alpha) = \beta_k^\theta(F^{-1}(x))\alpha(x)$, for function $\alpha \in \ell^1(0, 1)$.

By the chain rule, (See van der Vaart and Wellner (1996, Lemma 3.9.3).) the map $\psi \circ \phi : D_\phi \rightarrow \ell^1(0, 1)$ is Hadamard-differentiable at F tangentially to $C[s, t]$, with

derivative

$$(\psi_{\phi(F)}^\theta - \phi_F^\theta)(\alpha) = (\psi_{F^{-1}}^\theta - \phi_F^\theta)(\alpha) = \beta_k^\theta(F^{-1})\phi_F^\theta(\alpha) = \beta_k^\theta(F^{-1})(\alpha/f - F^{-1}),$$

for any function $\alpha \in D[s, t]$. That is, at any $p \in (0, 1)$, the derivative is

$$(\psi_{\phi(F)}^\theta - \phi_F^\theta)(\alpha)(p) = \beta_k^\theta(F^{-1}(p)) \left(\frac{\alpha(F^{-1}(p))}{f(F^{-1}(p))} \right).$$

Applying the functional delta-methods (See van der Vaart and Wellner (1996, Theorem 3.9.4)) to the Donsker's theorem and the above result that the map $\psi - \phi : D_\phi \mathcal{V} \rightarrow \ell^1(0, 1)$ is Hadamard-differentiable at F tangentially to $C[s, t]$,

$$\begin{aligned} \rho_n^-(\beta_k - F_n^{-1} - \beta_k - F^{-1}) &= \rho_n^-(\psi - \phi)(F_n) - (\psi - \phi)(F) \\ &= (\psi - \phi)_F^\theta(G - F) \\ &= \beta_k^\theta(F^{-1}) \frac{(G - F) - F^{-1}}{f(F^{-1})} \\ &= \beta_k^\theta(F^{-1}) \frac{G}{f(F^{-1})} \\ &= \beta_k^\theta(F^{-1}) \frac{G}{f(F^{-1})}, \text{ in } \ell^1(0, 1), \end{aligned} \quad (\text{C.68})$$

where G is the standard Brownian bridge. The last equality is due to the symmetry of the Gaussian process G . The limit process is Gaussian with mean zero and covariance function

$$\beta_k^\theta(F^{-1}(u))\beta_k^\theta(F^{-1}(v)) \frac{u \wedge v - uv}{f(F^{-1}(u))f(F^{-1}(v))}, \quad u, v \in \mathbb{R}. \quad \square$$

C.1.12 Proof of Theorem 15: uniform asymptotic normality of coefficient function estimator in simplified model

Proof. Consider the standard uniform distribution F_U with domain $(s, t) = (0, 1)$. It is continuously differentiable on $(0, 1)$ with a strictly positive constant derivative $f_U = 1$

on $(0, 1)$. Therefore, under Assumption 25, from Lemma 23,

$$\rho_{\bar{n}}(\beta_k \ F_{n,U}^{-1} \ \beta_k \ F_U^{-1}) \ \beta_k^\theta(F_U^{-1})G = \beta_k^\theta G, \text{ in } \ell^1(0, 1). \quad (\text{C.69})$$

The last equality is due to F_U is an identity function on $(0, 1)$. The limit is Gaussian process with mean zero and covariance function

$$\beta_k^\theta(F_U^{-1}(u))\beta_k^\theta(F_U^{-1}(v))(u \wedge v - uv) = \beta_k^\theta(u)\beta_k^\theta(v)(u \wedge v - uv), \ u, v \geq \mathbb{R}.$$

Under Assumptions 14–18 in the simplified model, we get the perfect coefficient estimation $\hat{\beta}_k(\tau) = \beta_k(U_{n:d\tau ne})$ for all τ on $(0, 1)$. Since $U_{n:d\tau ne} = F_{n,U}^{-1}(\tau)$ we have $\hat{\beta}_k = \beta_k \ F_{n,U}^{-1}$ on $(0, 1)$. Also, $F_U^{-1}(\tau) = \tau$. Therefore,

$$\begin{aligned} \rho_{\bar{n}}(\hat{\beta}_k(\cdot) \ \beta_k(\cdot)) &= \rho_{\bar{n}}(\beta_k \ F_{n,U}^{-1} \ \beta_k \ F_U^{-1}) \\ \beta_k^\theta(F_U^{-1})G &= \beta_k^\theta G, \text{ in } \ell^1(0, 1), \end{aligned} \quad (\text{C.70})$$

where the limit is Gaussian process with mean zero and covariance function $\beta_k^\theta(u)\beta_k^\theta(v)(u \wedge v - uv)$, $u, v \geq \mathbb{R}$. □

C.1.13 Proof of Corollary 16: pointwise asymptotic normality of coefficient estimator in simplified model

Proof. It is straightforward to apply Theorem 15 the specific point $\tau \geq (0, 1)$. □

C.1.14 Proof of Theorem 17: uniform asymptotic normality of coefficient function estimator in general model

Proof. Similar as shown in the proof in Theorem 15, when applying Lemma 23 to the standard uniform distribution F_U , under Assumption 25,

$$\rho_{\bar{n}}(\beta_k \ F_{n,U}^{-1} \ \beta_k \ F_U^{-1}) \ \beta_k^\theta(F_U^{-1})G = \beta_k^\theta G, \text{ in } \ell^1(0, 1). \quad (\text{C.71})$$

The limit process is Gaussian with mean zero and covariance function

$$\beta_k^\theta(F_U^{-1}(u))\beta_k^\theta(F_U^{-1}(v))(u \wedge v - uv) = \beta_k^\theta(u)\beta_k^\theta(v)(u \wedge v - uv), \quad u, v \geq \mathbb{R}.$$

Under Assumptions 14–17 and 19–23 and the n/T rate relation $n = o\left(T^{\frac{\kappa}{3+\delta}}\right)$, we know the ordering of rank variable can be correctly discovered, as shown in the proof of Theorem 9; The individual with the k th ordering in the true outcome values $Y = \mathbf{x}^\theta(U_{n:k})$ at $X = x$ is also the same individual who has the k th ordering in the fitted outcome values $\hat{Y} = \mathbf{x}^{\theta^\wedge}(U_{n:k})$ at $X = x$. Specifically Assumption 21 (ii) assumes the gap between the estimated individual coefficient and the true coefficient in terms of T order.

$$\hat{\beta}_k(U_i) - \beta_k(U_i) = O_p(T^{-\kappa}), \quad \text{as } T \rightarrow \infty, \text{ for } \kappa > 0. \quad (\text{C.72})$$

Therefore, the coefficient estimator can be written as

$$\begin{aligned} \hat{\beta}_k(\tau) &= \hat{\beta}_k(U_{n:(\tau n e)}) = \hat{\beta}_k(U_{\sigma(\tau n e)}) \\ &= \beta_k(U_{\sigma(\tau n e)}) + O_p(T^{-\kappa}) \\ &= \beta_k(U_{n:(\tau n e)}) + O_p(T^{-\kappa}). \end{aligned} \quad (\text{C.73})$$

Then,

$$\begin{aligned} \rho_n^-(\hat{\beta}_k(\cdot) - \beta_k(\cdot)) &= \rho_n^-(\beta_k - F_{n,U}^{-1} + O_p(T^{-\kappa}) - \beta_k - F_U^{-1}) \\ &= \rho_n^-(\beta_k - F_{n,U}^{-1} - \beta_k - F_U^{-1}) + n^{1/2}O_p(T^{-\kappa}) \\ &= \beta_k^\theta(F_U^{-1})G = \beta_k^\theta G, \quad \text{in } \ell^1(0, 1), \quad \text{as } n \rightarrow \infty, T \rightarrow \infty, \end{aligned} \quad (\text{C.74})$$

since $n^{1/2}O_p(T^{-\kappa}) \xrightarrow{p} 0$ under the rate relation $n = o\left(T^{\frac{\kappa}{3+\delta}}\right)$. That is, $0 < \frac{n}{T^{2\kappa}} \rightarrow 0$, as $n \rightarrow \infty, T \rightarrow \infty$, for $\kappa > 0, \delta > 0$.¹ \square

¹The δ here is the same as it is introduced in the proof of Theorem 9.

C.1.15 Proof of Lemma 19: consistency of bootstrap empirical process

Proof. Let F be a distribution function that have compact support $[s, t]$ and be continuously differentiable on its support with strictly positive derivative f . Let \mathbb{P} denote its probability measure. Let F_n be the empirical distribution of an iid sample from F with size n . Let \mathbb{P}_n denote its probability measure.

Let \hat{F}_n be the bootstrap empirical distribution of an iid sample from F_n with size n . Let $\hat{\mathbb{P}}_n$ denote its probability measure.

Let D_ϕ be the set of all distribution functions with measures concentrated on $[s, t]$. Let map $\psi : \ell^1(0, 1) \rightarrow \ell^1(0, 1)$ be $\psi(A)(x) = \beta_k(A(x)) = \beta_k \cdot A(x)$. Let $\phi : D_\phi \rightarrow D[s, t] \rightarrow \ell^1(0, 1)$ be the inverse map $\phi(G) = G^{-1}$. In the Lemma 23 proof, we have shown by chain rule the map $\psi \circ \phi : D_\phi \rightarrow \ell^1(0, 1)$ is Hadamard-differentiable at \mathbb{P} tangentially to $C[s, t]$ under Assumption 25.

As Lemma 18 stated, which is proved in van der Vaart and Wellner (1996, Theorem (3.6.13)), conditional on almost all the random dataset X_1, \dots, X_n , the sequence of bootstrap empirical process converges in distribution to some limit process H

$$\rho_{\hat{\mathbb{P}}_n - \mathbb{P}_n}^-(X_1, \dots, X_n) \rightarrow H. \quad (\text{C.75})$$

The limit process H is the same as the limit of the empirical process $\rho_{\mathbb{P}_n - \mathbb{P}}^-(G - F)$, which is $G - F$ previously with G being the standard Brownian bridge. The Lemma 18 is the same as the equation (3.9.9) in van der Vaart and Wellner (1996), which is a condition used to apply the delta method. That is,

$$\sup_{h \in BL_1(D)} \left| \mathbb{E}_M h \left(\rho_{\hat{\mathbb{P}}_n - \mathbb{P}_n}^- \right) - \mathbb{E} h(H) \right| \neq 0, \quad (\text{C.76})$$

where $BL_1(D)$ is the set of all functions $h : \ell^1(F) \rightarrow [0, 1]$ on D with Lipschitz norm bounded by 1. The \mathbb{E}_M denotes the expectation is with respect to the random vector

$\mathbf{M} = (M_{n1}, \dots, M_{nn})$. The random vector \mathbf{M} follows a multinomial distribution with parameter $(n; \frac{1}{n}, \dots, \frac{1}{n})$.² The left-hand side is conditional on (almost all) the (random) dataset X_1, \dots, X_n , since $P_n = P_n(X_1, \dots, X_n)$ and $\hat{P}_n = \hat{P}_n(X_1, \dots, X_n, \mathbf{M})$.

By van der Vaart and Wellner (1996, Theorem 3.9.11), applying the delta-method to the bootstrap empirical process with the Hadamard-differentiable function $\psi \circ \phi$,

$$\sup_{h \in 2BL_1(D)} \left| E_M h \left(\rho_n^- \left(\psi \circ \phi(\hat{P}_n) - \psi \circ \phi(P_n) \right) \right) - E h \left((\psi \circ \phi)_P^\theta(H) \right) \right| \rightarrow 0. \quad (\text{C.77})$$

That is, conditional on almost all the random dataset, the bootstrap empirical process of the $\psi \circ \phi$ transformed functions converges in distribution to a limit process

$$\rho_n^- \left((\psi \circ \phi)(\hat{P}_n) - (\psi \circ \phi)(P_n) \right) \rightarrow (\psi \circ \phi)_P^\theta(H). \quad (\text{C.78})$$

Denote the inverse map ϕ of a probability measure as the inverse of the distribution function. That is, $\phi(P) = F^{-1}$. Thus, $\phi(P_n) = F_n^{-1}$ and $\phi(\hat{P}_n) = \hat{F}_n^{-1}$. The equation (C.78) can be written as

$$\rho_n^- \left(\beta_k \hat{F}_n^{-1} - \beta_k F_n^{-1} \right) \rightarrow (\beta_k F^{-1})_F^\theta(G \circ F), \quad (\text{C.79})$$

where G is the standard Brownian bridge. The limit process has been derived in Lemma 23. That is, conditional on almost all the random draw of data,

$$\rho_n^- \left(\beta_k \hat{F}_n^{-1} - \beta_k F_n^{-1} \right) \rightarrow \beta_k^\theta(F^{-1}) \frac{G}{f(F^{-1})}, \text{ in } \ell^1(0, 1). \quad \square$$

C.1.16 Proof of Theorem 20: consistency of bootstrap estimation function in simplified model

Proof. Under Assumptions 14–18 in the simplified model, we have perfect coefficient estimation $\hat{\beta}_k(\tau) = \beta_k(U_{n:d\tau ne})$ for all τ on $(0, 1)$. Since $U_{n:d\tau ne} = F_{n,U}^{-1}(\tau)$, the coefficient

²Here we consider the standard bootstrap. It can be viewed as a special case of the weighted bootstrap, where \mathbf{M} is exchangeable nonnegative random vector. The generalization to weighted bootstrap is showed in Appendix C.2.2.

function estimator can be written as a composition $\hat{\beta}_k = \beta_k \circ F_{n,U}^{-1}$.

Let $(Y_{it}^b, X_{it}^b), i = 1, \dots, n, t = 1, \dots, T$ denote the bootstrap sample. Denote it in a matrix form $\mathbf{Y}_i^b = (Y_1, \dots, Y_T)$ and $\mathbf{X}_i^b = (\mathbf{X}_{i1}, \dots, \mathbf{X}_{iT})$. For each bootstrap individual, we can perfectly estimate its coefficient parameter $\beta^b(U_i) = (\mathbf{X}_i^b)^{-1} \mathbf{Y}_i^b$.

Let (U_1^b, \dots, U_n^b) denote the unobserved rank variable associated with each individual for the bootstrap sample. The (U_1^b, \dots, U_n^b) is implicitly drawn from (U_1, \dots, U_n) with probability measure \mathbb{P}_n . Each unobserved rank variable in the original sample follows a standard uniform distribution $U_i \sim \text{Unif}(0, 1)$. Let $U_{n:\mathcal{O}n\tau e}^b$ denote the $\mathcal{O}n\tau e$ th order statistic of the bootstrap sample (U_1^b, \dots, U_n^b) . We have $U_{n:\mathcal{O}n\tau e}^b = \hat{F}_{n,U}^{-1}(\tau)$, where $\hat{F}_{n,U}$ is the distribution function for the bootstrap sample (U_1^b, \dots, U_n^b) .

Given the data $(Y_{it}, \mathbf{X}_{it})$ for $i = 1, \dots, n, t = 1, \dots, T$, implementing the estimation process in the simplified model, we obtain the n individuals coefficients parameters in the bootstrap sample $(U_1^b), \dots, (U_n^b)$, since $(U_i^b) = \beta^b(U_i)$. Based on Assumption 16, the order of $Y^b = \mathbf{x}^\theta \circ \beta^b$ represents the order of U^b . Thus, by the definition of coefficient estimator in the simplified model, the coefficient associated with the $\mathcal{O}n\tau e$ th ordering of Y^b is the the coefficient at the $\mathcal{O}n\tau e$ th order of the rank variable values (U_1^b, \dots, U_n^b) . That is, $\hat{\beta}^b(\tau) = (U_{n:\mathcal{O}n\tau e}^b)$. Therefore, $\hat{\beta}^b(\tau) = (U_{n:\mathcal{O}n\tau e}^b) = (\hat{F}_{n,U}^{-1}(\tau))$. The bootstrap coefficient function estimator can be written as the composition of two functions. For each element of the bootstrap coefficient function $\hat{\beta}_k(\cdot)$, we have $\hat{\beta}_k(\cdot) = \beta_k \circ \hat{F}_{n,U}^{-1}$.

Therefore, conditional on almost all the random data, the bootstrap empirical process converges to a limit process

$$\begin{aligned} \mathcal{P}_n^-(\hat{\beta}_k(\cdot) - \beta_k(\cdot)) &= \mathcal{P}_n^-(\beta_k \circ \hat{F}_{n,U}^{-1} - \beta_k \circ F_{n,U}^{-1}) \\ &= \beta_k^\theta(F_U^{-1}) \frac{G}{f_U(F_U^{-1})}, \text{ by Lemma 19} \\ &= \beta_k^\theta G, \text{ in } \ell^1(0, 1). \quad \square \end{aligned}$$

C.2 Supplemental details and proofs

C.2.1 Background about empirical process

Let $D[a, b]$ denote the Skorohod space of all cadlag functions on the interval $[a, b] \subset \mathbb{R}$ with uniform norm. Let $C[a, b]$ denote the space of all continuous function on $[a, b]$. Let $\ell^1(T)$ denote the collection of all bounded functions $f : T \rightarrow \mathbb{R}$ with the norm $\|f\|_1 = \sup_{t \in T} |f(t)|$.

Donsker Theorem

Let F_n be the sequence of the empirical distribution function of an iid sampling X_1, X_2, \dots with size n from distribution function F . The Donsker's theorem is

$$\rho_{\bar{n}}(F_n - F) \Rightarrow G - F, \text{ in } D[-1, +1] \quad (\text{C.80})$$

where G is the standard Brownian bridge with mean zero and variance function $\text{Cov}[G(s), G(t)] = s \wedge t - st$.³

For the standard uniform distribution, the proceeding Donsker's theorem results become

$$\rho_{\bar{n}}(F_{n,U} - F_U) \Rightarrow G - F_U = G, \text{ in } D[0, 1] \quad (\text{C.81})$$

since F_U is the identity function on $[0, 1]$.

Inverse map

Let $0 < p < 1$. Let D_ϕ be the collection of all nondecreasing functions. Let the distribution function $F \in D_\phi$ be differentiable at $\xi_p \in (a, b)$ such that $F(\xi_p) = p$ with strictly positive derivative. By the Lemma 3.9.20 in van der Vaart and Wellner (1996), the inverse map $G \mapsto G^{-1}$ as a map $\phi : D_\phi \rightarrow D[a, b] \rightarrow \mathbb{R}$ is Hadamard-differentiable

³The right-hand side has mean zero and variance function $\text{Cov}[G_F(s), G_F(t)] = F(s) \wedge F(t) - F(s)F(t)$.

at F tangentially to the set of D_0 that are the set of functions $\alpha \in D[a, b]$ that are continuous at $\xi_p = F^{-1}(p)$. The derivative is $\phi_F(\alpha) = \alpha(\xi_p)/F'(\xi_p) = (\alpha/f)(\xi_p)$.

Empirical quantile: pointwise central limit theorem

Apply both delta-method (See van der Vaart and Wellner (1996, Theorem 3.9.4)) and the inverse map results (van der Vaart and Wellner (1996, Lemma 3.9.20)) to the Donsker's theorem, we have the pointwise central limit theorem for the empirical quantile function. (See van der Vaart and Wellner (1996, Example 3.9.21).) That is, for any $q \in (0, 1)$,

$$\mathcal{D}_n^-(F_n^{-1}(p) - F^{-1}(p)) \stackrel{\mathcal{D}}{\rightarrow} \phi_F^0(G - F) = \frac{G(p)}{f(F^{-1}(p))}. \quad (\text{C.82})$$

For the standard uniform distribution function F_U and its sequence of empirical distribution function $F_{n,U}$, the pointwise empirical quantile convergence is, for any $0 < \tau < 1$,

$$\mathcal{D}_n^-(F_{n,U}^{-1}(\tau) - F_U^{-1}(\tau)) = \mathcal{D}_n^-(U_{n:\lfloor n\tau \rfloor} - \tau) \stackrel{\mathcal{D}}{\rightarrow} G(\tau), \quad (\text{C.83})$$

since its density function f_U is uniformly 1 on $(0, 1)$ and the standard Brownian bridge is symmetric around zero, $G(\tau) = -G(\tau)$.

Empirical quantile processes: uniform central limit theorem

More general uniform central limit theorem of the sequence of empirical quantile processes can be obtained through the Hadamard-differentiability of the inverse map (See van der Vaart and Wellner (1996, Lemma 3.9.23 (i))) and the delta-method.

Let $0 < p < q < 1$. Let F be continuously differentiable on the interval $[a, b] = [F^{-1}(p) - \epsilon, F^{-1}(q) + \epsilon]$ for some $\epsilon > 0$, with positive derivative f . Then the inverse map $G \mapsto G^{-1}$ as $\phi : D_1 \subset D[a, b] \rightarrow \ell^1[p, q]$ is Hadamard-differentiable at F tangentially to $C[a, b]$, by van der Vaart and Wellner (1996, Lemma 3.9.23 (i)). Applying the delta-method to the inverse map, together with the Donsker's theorem, it yields the empirical

quantile process convergence. (See van der Vaart and Wellner (1996, Example 3.9.24))

$$\rho_{\bar{n}}(F_n^{-1} - F^{-1}) = \frac{G - F(F^{-1})}{f(F^{-1})} = \frac{G}{f(F^{-1})} \text{ in } \ell^1[p, q]. \quad (\text{C.84})$$

For the uniform distribution function, the proceeding results yields a uniform convergence of the empirical quantile process.

$$\rho_{\bar{n}}(F_{n,U}^{-1} - F_U^{-1}) = G - G \text{ in } \ell^1(0, 1). \quad (\text{C.85})$$

That is the same as the equation (61) in Shorack and Wellner (1986, Section 3.1 Theorem 1).

C.2.2 Generalization to the weighted empirical bootstrap process

The standard bootstrap in this paper can be generalized to the weighted bootstrap.

Given a sample X_1, X_2, \dots, X_n , the bootstrap sample $\hat{X}_1, \hat{X}_2, \dots, \hat{X}_m$ is an iid sampling drawn from the original sample X_1, X_2, \dots, X_n with replacement, It treats the empirical measure P_n as the true measure for bootstrap sampling. Let \hat{P}_m denote the bootstrap empirical measure,

$$\hat{P}_m = \frac{1}{m} \sum_{i=1}^m M_{ni} \delta_{X_i} \quad (\text{C.86})$$

where M_{ni} denote the number of times X_i is redrawn in the size m bootstrap sample. It is the standard bootstrap, if $m = n$. It is the m out of n bootstrap, if $m < n$. The bootstrap empirical process is

$$\hat{G}_{n,m} = \rho_{\bar{m}}(\hat{P}_m - P_n) = \rho_{\bar{m}} \left[\frac{1}{m} \sum_{i=1}^n M_{mi} \delta_{X_i} - \frac{1}{n} \sum_{i=1}^n \delta_{X_i} \right] = \rho_{\bar{m}} \sum_{i=1}^n \left(M_{mi} - \frac{m}{n} \right) \delta_{X_i}, \quad (\text{C.87})$$

with $\sum_{i=1}^n M_{mi} = m$ and $M_{mi} \in \{0, 1, \dots, m\}$.

More generally, consider the interchangeable weighted bootstrap. For each n , let

(W_{n1}, \dots, W_{nn}) be an exchangeable, nonnegative random vector.⁴ Let \hat{P}_n denote the weighted bootstrap empirical measure,

$$\hat{P}_n = \frac{1}{n} \sum_{i=1}^n W_{ni} \delta_{X_i}. \quad (\text{C.88})$$

It includes both the standard bootstrap and m out of n bootstrap.⁵ With the exchangeable weighted bootstrap measure, we can write the bootstrap empirical distribution as \hat{F}_n .

$$\hat{F}_n(x) = \frac{1}{n_B} \sum_{i=1}^{n_B} \mathbf{1}\{\hat{X}_i \leq x\} \quad (\text{C.89})$$

where $n_B = \sum_{i=1}^n W_{ni}$ is the size of bootstrap sample and \hat{X}_i denotes the bootstrap sample. The empirical bootstrap process is

$$\hat{G}_n = \rho_{\hat{P}_n}^{-1}(\hat{P}_n, W_n P_n), \quad (\text{C.90})$$

where $W_n = 1/n \sum_{i=1}^n W_{ni}$. In the standard bootstrap that n variables are resampled, $W_n = 1$ and the empirical process is $\hat{G}_n = \rho_{\hat{P}_n}^{-1}(\hat{P}_n, P_n)$.

Assumption 27 (Exchangeable weighted bootstrap weights from van der Vaart and Wellner (1996) (Section 3.6.2)). For each n , the exchangeable, nonnegative random weights vector (W_{n1}, \dots, W_{nn}) satisfies

$$\sup_n \|W_{n1} - W_n\|_{2,1} < 1, \quad (\text{i})$$

$$\frac{1}{n} E \max_{1 \leq i \leq n} |W_{ni} - W_n| \leq 0, \quad (\text{ii})$$

$$\frac{1}{n} \sum_{i=1}^n (W_{ni} - W_n)^2 \leq c^2 > 0, \quad (\text{iii})$$

where $W_n = 1/n \sum_{i=1}^n W_{ni}$; $\|\cdot\|_{2,1}$ means the $L_{2,1}$ norm of a random variable η , $\|\eta\|_{2,1} =$

⁴ W_{ni} can be non-integer valued.

⁵By van der Vaart and Wellner (1996, Example 3.6.10 and Example 3.6.11), the standard bootstrap is the case $(W_{n1}, \dots, W_{nn}) \sim \text{Multinomial}(n; 1/n, \dots, 1/n)$ with $c = 1$; the m out of n bootstrap is the case $(W_{n1}, \dots, W_{nn}) \sim \sqrt{\frac{n}{m}} \text{Multinomial}(m; 1/n, \dots, 1/n)$ with $c = 1$.

$\int_0^1 \sqrt{P(|\eta_j| > x)} dx$; c is some constant.

We can rescale the weights (W_{n1}, \dots, W_{nn}) by dividing c . Then the rescaled weights satisfy Assumption 27 with $c=1$. Under Assumption 27, we have the uniform convergence of the bootstrap empirical process.

Lemma 24 (The weighted bootstrap empirical process convergence, same as van der Vaart and Wellner (1996) Theorem 3.6.13). *Under Assumption 27, the weighted bootstrap empirical process converges to a limit process. $\hat{G}_n \Rightarrow H$, as $n \rightarrow \infty$.*

C.2.3 45 countries in empirical example and their democracy level

See Table C.1.

Table C.1: The 45 countries in empirical example and their average democracy level indexed between 0 and 1.

Country	Dem	Country	Dem	Country	Dem	Country	Dem	Country	Dem
DZA	0.20	CAN	1.00	HUN	0.43	NLD	1.00	SYR	0.21
AGO	0.28	TCD	0.24	IND	0.93	NGA	0.45	THA	0.47
AZE	0.21	CHL	0.63	IDN	0.35	NOR	1.00	ARE	0.10
ARG	0.46	CHN	0.14	IRN	0.21	PAK	0.54	TUN	0.17
AUS	1.00	COL	0.77	ITA	0.82	PER	0.59	TUR	0.69
BHR	0.05	DNK	1.00	KWT	0.09	ROM	0.33	EGY	0.27
BOL	0.52	ECU	0.63	MYS	0.75	RUS	0.72	GRB	1.00
BRA	0.54	FXX	0.87	MEX	0.37	SAU	0.00	USA	1.00
CMR	0.19	DEU	0.59	OMN	0.08	SDN	0.31	VEN	0.66

Use 0.5 as the cutoff to divide the democratic and non-democratic countries, as Cotet and Tsui (2013) did, there are 21 democratic countries and 24 countries as non-democratic countries in the empirical example.

C.2.4 Bootstrap confidence interval at 95% level and bootstrap critical value in empirical example

See Table C.2.

Table C.2: The estimates of the effect oil wealth per capita of on military defense spending as ratio of GDP, its standard error, and 95% CI at ranks 0.1 to 0.9.

τ	Slope Estimates	Bootstrap S.E.	Bootstrap Critical Value	95% Bootstrap CI	
				LowerBound	UpperBound
0.1	0.18	0.13	4.41	0.38	0.75
0.2	0.43	0.32	3.63	1.60	0.75
0.3	0.06	0.23	5.20	1.11	1.23
0.4	0.14	0.20	4.39	1.03	0.75
0.5	0.18	0.25	2.47	0.79	0.43
0.6	0.12	0.22	3.27	0.58	0.83
0.7	0.19	0.22	2.76	0.81	0.42
0.8	0.01	0.24	1.98	0.47	0.46
0.9	1.01	0.40	3.68	2.47	0.46
OLS	0.04	0.04	1.96	0.03	0.12

700 bootstrap times. The bootstrap critical values are the converged limit values as bootstrap times increases.

BIBLIOGRAPHY

- Abadie, Alberto, Joshua Angrist, and Guido Imbens. 2002a. “Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings.” *Econometrica* 70 (1):91–117. URL <https://doi.org/10.1111/1468-2627.02282>. [2]
- . 2002b. “Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings.” *Econometrica* 70 (1):91–117. [55]
- Abrevaya, Jason and Christian M. Dahl. 2008. “The Effects of Birth Inputs on Birthweight.” *Journal of Business & Economic Statistics* 26 (4):379–397. URL <https://doi.org/10.1198/0735-1725.0269>. [57]
- Andrews, Donald W. K. 1991. “Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation.” *Econometrica* 59 (3):817–858. URL <http://www.jstor.org/stable/2938229>. [7]
- Arellano, Manuel and Stéphane Bonhomme. 2012. “Identifying Distributional Characteristics in Random Coefficients Panel Data Models.” *The Review of Economic Studies* 79 (3):987–1020. URL <https://doi.org/10.1193/restud/rdr45>. [59]
- . 2016. “Nonlinear panel data estimation via quantile regressions.” *Econometrics Journal* 19 (3):C61–C94. URL <https://doi.org/10.1111/ectj.12062>. [57]
- Armstrong, Timothy B. and Michal Kolesár. 2021. “Sensitivity Analysis using Approximate Moment Condition Models.” *Quantitative Economics* 12 (1):77–108. [49]
- Baker, Matthew, Travis Smith, and David Powell. 2016. “QREGPD: Stata module to perform Quantile Regression for Panel Data.” Statistical Software Components, Boston College Department of Economics. URL <https://ideas.repec.org/c/boc/bocode/s458157.html>. [58]
- Besley, Timothy and Torsten Persson. 2011. “The Logic of Political Violence.” *Quarterly Journal of Economics* 126 (3):1411–1445. URL <https://doi.org/10.1193/qje/12631411>.

qj r 25. [54]

Bohn, Henning and Robert T. Deacon. 2000. "Ownership Risk, Investment, and the Use of Natural Resources." *American Economic Review* 90 (3):526–549. URL <https://doi.org/10.1257/aer.90.3.526>. [54]

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2008. "Bootstrap-Based Improvements for Inference with Clustered Errors." *Review of Economics and Statistics* 90 (3):414–427. URL <https://doi.org/10.1162/rest.90.3.414>. [89]

Campbell, John Y. 2003. "Consumption-Based Asset Pricing." In *Handbook of the Economics of Finance: Financial Markets and Asset Pricing*, vol. 1, Part B, edited by George M. Constantinides, Milton Harris, and René M. Stulz, chap. 13. North Holland, 803–887. URL [https://doi.org/10.116/S1574-1222\(03\)122-7](https://doi.org/10.116/S1574-1222(03)122-7). [16]

Canay, Ivan A. 2011. "A simple approach to quantile regression for panel data." *Econometrics Journal* 14 (3):368–386. URL <https://www.jstor.org/stable/23116937>. [56]

Chamberlain, Gary. 1982. "Multivariate regression models for panel data." *Journal of Econometrics* 18 (1):5–46. URL [https://doi.org/10.1016/0344-7682\(82\)90094-X](https://doi.org/10.1016/0344-7682(82)90094-X). [57]

Chen, Le-Yu and Sokbae Lee. 2018. "Exact computation of GMM estimators for instrumental variable quantile regression models." *Journal of Applied Econometrics* 33 (4):553–567. URL <https://doi.org/10.1111/jae.2619>. [23]

Cheng, Xu, Zhipeng Liao, and Ruoyao Shi. 2016. "An Averaging GMM Estimator Robust to Misspecification." Working paper, available at http://www.econ.ucla.edu/liao/papers_pdf/ChengLiaoShi_AvgGMM.pdf. [18]

———. 2019. "On Uniform Asymptotic Risk of Averaging GMM Estimators." *Quantitative Economics* 10 (3):931–979. URL <https://onlinelibrary.wiley.com/doi/abs/10.3982/QE711>. [ix, 21, 22, 23, 24, 28, 30, 31, 34, 35, 36, 37, 48, 113, 116, 117]

- Chernozhukov, Victor, Iván Fernández-Val, Jinyong Hahn, and Whitney Newey. 2013. “Average and Quantile Effects in Nonseparable Panel Models.” *Econometrica* 81 (2):535–580. [58]
- Chernozhukov, Victor, Iván Fernández-Val, and Blaise Melly. 2013. “Inference on Counterfactual Distributions.” *Econometrica* 81 (6):2205–2268. URL <https://www.jstor.org/stable/23524318>. [88]
- Chernozhukov, Victor and Christian Hansen. 2005a. “An IV Model of Quantile Treatment Effects.” *Econometrica* 73 (1):245–261. URL <http://www.jstor.org/stable/3598944>. [1]
- . 2005b. “An IV Model of Quantile Treatment Effects.” *Econometrica* 73 (1):245–261. URL <https://www.jstor.org/stable/3598944>. [20, 23, 25, 42, 51, 53, 55, 58, 61, 62]
- . 2006a. “Instrumental Quantile Regression Inference for Structural and Treatment Effect Models.” *Journal of Econometrics* 132 (2):491–525. URL <https://doi.org/10.116/j.jeconom.2005.12.009>. [2, 10]
- . 2006b. “Instrumental quantile regression inference for structural and treatment effect models.” *Journal of Econometrics* 132 (2):491–525. [20, 23, 55]
- . 2008. “Instrumental Variable Quantile Regression: A Robust Inference Approach.” *Journal of Econometrics* 142 (1):379–398. [55]
- Chernozhukov, Victor, Christian Hansen, and Kaspar Wüthrich. 2017. “Instrumental Variable Quantile Regression.” In *Handbook of Quantile Regression*, edited by R. Koenker, V. Chernozhukov, X. He, and L. Peng. CRC/Chapman-Hall, forthcoming. [13]
- Chernozhukov, Victor and Han Hong. 2003. “An MCMC Approach to Classical Estimation.” *Journal of Econometrics* 115 (2):293–346. URL [https://doi.org/10.116/S0344-7663\(03\)1113-3](https://doi.org/10.116/S0344-7663(03)1113-3). [5]

- Chesher, Andrew. 2003. "Identification in nonseparable models." *Econometrica* 71 (5):1405–1441. [55]
- Collier, Paul and Anke Hoeffler. 1998. "On economic causes of civil war." *Oxford Economic Papers* 50 (4):563–573. URL <https://doi.org/10.1093/oep/50.4.563>. [54]
- . 2004. "Greed and grievance in civil war." *Oxford Economic Papers* 56 (4):563–595. URL <https://doi.org/10.1093/oep/gpf064>. [54]
- Cotet, Anca M. and Kevin K. Tsui. 2013. "Oil and Conflict: What Does the Cross Country Evidence Really Show?" *American Economic Journal: Macroeconomics* 5 (1):49–80. URL <https://doi.org/10.1257/mac.5.1.49>. [54, 90, 91, 92, 181]
- Davison, Anthony Christopher and D V Hinkley. 1997. *Bootstrap methods and their application*. Cambridge University Press. [90]
- de Castro, Luciano and Antonio F. Galvao. 2017. "Dynamic Quantile Models of Rational Behavior." University of Iowa, mimeo, available at <http://dx.doi.org/10.2139/ssrn.2932382>. [3, 14]
- de Castro, Luciano, Antonio F. Galvao, and David M. Kaplan. 2017. "Smoothed instrumental variables quantile regression, with estimation of quantile Euler equations." Working paper, available at <https://faculty.missouri.edu/~kaplndm>. [2, 3, 6, 7, 8, 10, 11, 12, 96, 100, 101, 102, 105]
- de Castro, Luciano, Antonio F. Galvao, David M. Kaplan, and Xin Liu. 2019. "Smoothed GMM for quantile models." *Journal of Econometrics* 213 (1):121–144. URL <https://doi.org/10.1016/j.jeconom.2019.04.008>. [24, 28, 30, 31, 34, 113, 114, 118]
- Fearon, James D. and David D. Laitin. 2003. "Ethnicity, Insurgency, and Civil War." *American Political Science Review* 97 (1):75–90. URL <http://www.jstor.org/stable/3118222>. [54]
- Fernández-Val, Iván and Joonhwhah Lee. 2013. "Panel data models with nonadditive unobserved heterogeneity: Estimation and inference." *Quantitative Economics* 4 (3):453–

481. URL <https://doi.org/10.3982/QE75>. [59]
- Fru mento, Paolo, Matteo Bottai, and Iván Fernández-Val. 2020. “Parametric Modeling of Quantile Regression Coefficient Functions with Longitudinal Data.” Working paper, available at <https://arxiv.org/abs/2006.16156>. [56]
- Galvao, Antonio F. 2011. “Quantile regression for dynamic panel data with fixed effects.” *Journal of Econometrics* 164 (1):142–157. [57]
- Galvao, Antonio F. and Kengo Kato. 2016. “Smoothed quantile regression for panel data.” 193 (1):92–112. URL <https://doi.org/10.1162/j.jeconom.2016.11.8>. [57]
- Galvao, Antonio F. and Liang Wang. 2015. “Efficient minimum distance estimator for quantile regression fixed effects panel data.” *Journal of Multivariate Analysis* 133:1–26. URL <https://doi.org/10.1162/j.jmva.2014.8.7>. [56]
- Goldman, Matt and David M. Kaplan. 2018. “Comparing distributions by multiple testing across quantiles or CDF values.” *Journal of Econometrics* 206 (1):143–166. URL <https://doi.org/10.1162/j.jeconom.2018.4.3>. [72, 73]
- Graham, Bryan S., Jinyong Hahn, Alexandre Poirier, and James L. Powell. 2018. “A quantile correlated random coefficients panel data model.” 206 (2):305–335. URL <https://doi.org/10.1162/j.jeconom.2018.6.4>. [57]
- Green, Edwin J. and William E. Strawderman. 1991. “A James-Stein Type Estimator for Combining Unbiased and Possibly Biased Estimators.” *Journal of the American Statistical Association* 86 (416):1001–1006. [117]
- Hamilton, James D. 1994. *Time Series Analysis*. Princeton, NJ: Princeton University Press. [68]
- Hansen, Bruce E. 2017a. “A Stein-Like 2SLS Estimator.” *Econometric Reviews* 36:840–852. [18]
- . 2017b. “A Stein-Like 2SLS Estimator.” *Econometric Reviews* 36 (6–9):840–852.

- URL <https://doi.org/10.1080/01674544.2017.137579>. [22, 26, 117]
- . 2020. “Econometrics.” URL <https://www.ssc.wisc.edu/~bhansen/econometrics>. Textbook draft. [89]
- Hansen, Lars Peter. 1982. “Large Sample Properties of Generalized Method of Moments Estimators.” *Econometrica* 50 (4):1029–1054. URL <http://www.jstor.org/stable/1912775>. [13]
- Harding, Matthew and Carlos Lamarche. 2009. “A quantile regression approach for estimating panel data models using instrumental variables.” *Economics Letters* 104 (3):133–135. URL <https://doi.org/10.1016/j.econlet.2009.04.025>. [57]
- Imbens, Guido W. and Whitney K. Newey. 2009. “Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity.” 77 (5):1481–1512. URL <https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA718>. [55, 62]
- Judge, George G and Ron C Mittelhammer. 2004. “A Semiparametric Basis for Combining Estimation Problems Under Quadratic Loss.” *Journal of the American Statistical Association* 99 (466):479–487. [117]
- Kaido, Hiroaki and Kaspar Wüthrich. 2021. “Decentralization Estimators for Instrumental Variable Quantile Regression Models.” *Quantitative Economics* :forthcoming. [23]
- Kaplan, David M. and Yixiao Sun. 2017a. “Smoothed Estimating Equations for Instrumental Variables Quantile Regression.” *Econometric Theory* 33 (1):105–157. URL <https://doi.org/10.1017/S266466615000047>. [2, 3, 5, 8, 10]
- . 2017b. “Smoothed estimating equations for instrumental variables quantile regression.” *Econometric Theory* 33 (1):105–157. URL <https://doi.org/10.1017/S266466615000047>. [23, 27, 31, 40, 112, 113]
- Kato, Kengo. 2012. “Asymptotic normality of Powell’s kernel estimator.” *Annals of the Institute of Statistical Mathematics* 64 (2):255–273. [31, 48, 108, 109, 112]

- Kato, Kengo, Antonio F. Galvao, and Gabriel V. Montes-Rojas. 2012. "Asymptotics for panel quantile regression models with individual effects." 170 (1):76–91. URL <https://doi.org/10.1162/j.jeconom.2012.2.7>. [56]
- Kim, Tae-Hwan and Halbert White. 2001. "James-Stein-Type Estimators in Large Samples With Application to the Least Absolute Deviations Estimator." *Journal of the American Statistical Association* 96 (454):697–705. [117]
- Kinal, Terrence W. 1980. "The existence of moments of k-class estimators." *Econometrica* 48 (1):241–249. URL <https://www.jstor.org/stable/191227>. [21]
- Koenker, R. 2005. *Quantile regression, Econometric Society Monographs*, vol. 38. Cambridge University Press. URL <https://doi.org/10.1177/CB0978051175498>. [51, 57]
- Koenker, Roger. 2004. "Quantile regression for longitudinal data." *Journal of Multivariate Analysis* 91 (1):74–89. [56]
- Koenker, Roger and Gilbert Bassett, Jr. 1978a. "Regression Quantiles." *Econometrica* 46 (1):33–50. URL <http://www.jstor.org/stable/1913643>. [1]
- . 1978b. "Regression Quantiles." *Econometrica* 46 (1):33–50. URL <https://www.jstor.org/stable/1913643>. [20, 55]
- Koenker, Roger and Stephen Portnoy. 1987. "L-Estimation for Linear Models." *Journal of the American Statistical Association* 82 (399):851–857. URL <https://doi.org/10.1187/1621459.1987.147858>. [30, 49]
- Lee, Sokbae. 2007. "Endogeneity in quantile regression models: A control function approach." *Journal of Econometrics* 141 (2):1131–1158. URL <https://doi.org/10.1162/j.jeconom.2007.1.14>. [55]
- Maasoumi, E. 1978. "A modified Stein-like estimator for the reduced form coefficients of simultaneous equations." *Econometrica* 46 (3):695–703. [22]
- Melly, Blaise and Kaspar Wüthrich. 2017. "Local Quantile Treatment Effects." In

- Handbook of Quantile Regression*, edited by Roger Koenker, Victor Chernozhukov, Xuming He, and Limin Peng. CRC/Chapman-Hall, 145–164. URL <https://www.routledgehandbooks.com/doi/10.1215/978131512256>. [55]
- Mittelhammer, Ron C. and George G. Judge. 2005. “Combining estimators to improve structural model estimation and inference under quadratic loss.” *Journal of Econometrics* 128 (1):1–29. URL <https://www.sciencedirect.com/science/article/pii/S03447641484>. [117]
- Newey, Whitney K. and Daniel McFadden. 1994. “Large Sample Estimation and Hypothesis Testing.” In *Handbook of Econometrics*, vol. 4, edited by Robert F. Engle and Daniel L. McFadden, chap. 36. Elsevier, 2111–2245. URL [https://doi.org/10.16/S1573-4412\(95\)8055-4](https://doi.org/10.16/S1573-4412(95)8055-4). [6, 11, 13, 100, 103, 105]
- Newey, Whitney K. and Kenneth D. West. 1987. “A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix.” *Econometrica* 55 (3):703–708. URL <http://www.jstor.org/stable/191361>. [7]
- Powell, David. 2020. “Quantile Regression with Nonadditive Fixed Effects.” Working paper available at <https://sites.google.com/site/davidmatthewpowell/quantile-regression-with-nonadditive-fixed-effects>. [58]
- Rosen, Adam M. 2012. “Set identification via quantile restrictions in short panels.” *Journal of Econometrics* 166 (1):127–137. URL <https://doi.org/10.1016/j.jeconom.2011.06.011>. [57]
- Shorack, Galen R. and Jon A. Wellner. 1986. *Empirical Processes with Applications to Statistics*. New York: John Wiley & Sons. [67, 147, 179]
- Stein, Charles and W. James. 1961. “Estimation with quadratic loss.” *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* 1:361–380. [22]
- van der Vaart, Aad W. 1998a. *Asymptotic Statistics*. Cambridge: Cambridge University

- Press. [11, 96, 98, 99, 101]
- . 1998b. *Asymptotic Statistics*. Cambridge: Cambridge University Press. URL <https://books.google.com/books?id=UEuQEM5RjWgC>. [83]
- van der Vaart, Aad W. and Jon A. Wellner. 1996. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. New York: Springer. [86, 170, 171, 174, 175, 177, 178, 179, 180, 181]
- Wilks, Samuel Stanley. 1962. *Mathematical Statistics*. New York: Wiley. [65]
- Wüthrich, Kaspar. 2016. “A comparison of two quantile models with endogeneity.” Working paper. [2]
- Xiang, Yang, Sylvain Gubian, Brian Suomela, and Julia Hoeng. 2013. “Generalized Simulated Annealing for Global Optimization: The GenSA Package.” *The R Journal* 5 (1):13–28. URL <https://journal.r-project.org/archive/2013/RJ-2013-02/index.html>. [7]
- Yogo, Motohiro. 2004. “Estimating the Elasticity of Intertemporal Substitution When Instruments are Weak.” *Review of Economics and Statistics* 86 (3):797–810. URL <https://doi.org/10.1162/089464604181177>. [3, 16]
- Zhu, Yinchu. 2019. “Learning non-smooth models: instrumental variable quantile regressions and related problems.” Available at <https://arxiv.org/abs/1805.06855>. [23]

VITA

Xin Liu was born in China. She obtained B.S. in Finance from a top tier university in China. She was admitted into the Ph.D. program at the Department of Economics, University of Missouri in 2016. She received her doctoral degree in May 2021.

Xin Liu specialized in econometrics, working with Professor Kaplan as her doctoral advisor. As a teaching assistant, Xin helped undergraduate and master's students learn econometrics, both in-person and online. As a PhD student, she also presented three different research papers at five conferences including the European Winter Meeting of the Econometric Society; Xin was awarded funding for conference travel by our department, MU, and one of the conferences.