QUASI-MAXIMUM LIKELIHOOD ESTIMATION METHODS WITH A CONTROL FUNCTION APPROACH TO ENDOGENEITY

By

Doosoo Kim

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Economics – Doctor of Philosophy

2017

ABSTRACT

QUASI-MAXIMUM LIKELIHOOD ESTIMATION METHODS WITH A CONTROL FUNCTION APPROACH TO ENDOGENEITY

By

Doosoo Kim

One of the fundamental problems in econometrics is the potential endogeneity in non-experimental data. This work focuses on econometric methods taking a control function approach to endogeneity. The agenda consists of two parts. In the first part, I study a general class of conditional mean regression methods with a control function, and their relative asymptotic efficiency relationship. Unlike previous results in the literature, the likelihood for the response variables can be incorrect up to the regression functions. My results provide more practical and general guidance on the choice of an estimator. In the second part, I propose a generalized Chamberlain device as a control function approach to time-invariant endogeneity in linear panel data quantile regression models with a finite time dimension. The new correlated effect (CE) estimator has substantial advantages compared to existing methods: (i) it is free of an incidental parameters problem, (ii) the correlated effect is not restricted to a linear functional form, and (iii) an arbitrary within-group dependence of regression errors is allowed. Due to the high-dimensionality of the control function, a nonconvex penalized estimator is adopted for sparse model selection.

In the first chapter, I study the asymptotic relative efficiency relationship among estimators based on a quasi-limited information likelihood (QLIL). First, I show that there exists a generalized method of moments estimator (GMM-QLIML) based on all the available quasi-scores. Second, the quasi-limited information maximum likelihood estimator (QLIML) is shown to be as efficient as GMM-QLIML under a set of generalized information matrix equalities. Third, I show that in a fully robust estimation of correctly specified conditional mean functions, QLIML is efficient relative to a two-step control function approach when the generalized linear model variance assumptions hold with a scaling restriction.

When a limited information structure is over-identified, the classical minimum distance (MD) estimator is often proposed as an estimation method. The purpose of the second chapter is to study its relative asymptotic efficiency relationship with respect to QLIML and two-step control function (CF) approach. First, I show that the MD estimator is asymptotically efficient relative to two other estimators. Second, I proved that the concentration of reduced form equation estimates does not affect the asymptotic efficiency of the structural parameter estimates in the MD estimation. Third, in a class of models, an if-and-only-if condition is derived for MD and other estimators to be asymptotically equivalent under the null hypothesis of exogeneity.

In the third chapter, I propose a point-identifying restriction and estimation procedure for a linear panel data quantile regression model with a fixed time dimension. The proposed model restriction reasonably accounts for the τ-quantile-specific time-invariant heterogeneity, and allows arbitrary within-group dependence of regression errors. The generalized Chamberlain device is taken analogously as a control function to capture τ-quantile-specific time-invariant endogenous variations. Since the sieve-approximated control function has high-dimensionality, the estimation procedure adopts penalization techniques under the sparsity assumption. Transformation of the sieve elements into a generalized Mundlak form is considered to make the sparsity assumption more plausible in some cases. The empirical application to birth weight analysis demonstrates a convincing case where the proposed estimator works as intended in real data.

Dedicated to my parents, and Kyuseon.

ACKNOWLEDGEMENTS

First of all, I'd like to thank my family members for their endless support. My wonderful parents and my great wife, Kyuseon (Kristy) always believed in me and encouraged me even when I was in deep trouble. All the support they have provided me over the years has made this work possible.

I would like to thank my advisor Jeffrey M. Wooldridge for his valuable advice and support. In addition to his insightful feedback on details of my research, his strong encouragement and optimistic view were essential ingredients for my work. He always led me to push my limits and achieve beyond my imagination. I deeply appreciate him giving me such great inspiration.

I am grateful to my dissertation committee members, Peter Schmidt and Kyoo il Kim for their rigorous and helpful feedback. Thanks to their brilliant comments, I could significantly improve the quality of my work. They were also a huge inspiration to me at all of the time.

I appreciate the wonderful support from the Department of Economics at Michigan State University. The comfortable work environment the department provided was crucial. Special thanks to the Chairs of the Department and Graduate Directors: Carl Davidson, Tim Vogelsang, Leslie E. Papke, and Todd Elder. I am also grateful to the following university staff: Belen Feight, Jay Feight, Margaret Lynch, Lori Jean Nichols, and Dean Olson for their unfailing support and assistance.

Among the great Ph.D. students in the Department of Economics, very special gratitude goes to my study group members: Muzna Alvi, Patrick Burke, Annie Chou, Po-Chun Huang, Riju Joshi, and Danielle Kaminski. It was fantastic to have the opportunity to work and interact with them. The moments we shared together greatly enriched my life in East Lansing.

Thank you all!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

2SLS Two-stage Least Square

AGLS Amemiya's Generalized Least Square

AIC Akaike Information Criterion

BIC Bayesian information criterion

CAN Consistent and Asymptotic Normal

CE Correlated Effect

CF Control Function

DGP Data Generating Process

GMM Generalized Method of Moments

LEF Linear Exponential Family

LIL Limited Information Likelihood

LIML Limited Information Maximum Likelihood

MCP Minimax Concave Penalty

MD Minimum Distance

QLIML Quasi-Limited Information Maximum Likelihood

QMLE Quasi-Maximum Likelihood Estimator

RMSE Root-Mean-Square Error

SCAD Smoothly Clipped Absolute Deviation

SD Standard Deviation

CHAPTER 1

RELATIVE EFFICIENCY OF QUASI-LIMITED INFORMATION MAXIMUM LIKELIHOOD ESTIMATOR

1.1 Introduction

Limited information likelihood (LIL)-based estimators have been widely used in instrumental variable estimation. The limited information maximum likelihood (LIML) estimator (Anderson and Rubin, 1949) and two-stage least square (2SLS) estimator (Theil, 1953; Basmann, 1957; Sargan, 1958) for linear models are workhorses of many empirical studies. In simultaneous probit models, analogously proposed LIML and two-stage conditional maximum likelihood estimator (Rivers and Vuong, 1988) are useful extensions of LIML and 2SLS to a nonlinear model. While correct specification of likelihoods has been assumed in the LIL literature, it is known that a certain class of maximum likelihood estimators have nice robustness against misspecification: quasi-maximum likelihood estimator (QMLE) is fully robust for correctly specified conditional mean if and only if the likelihood is in linear exponential familiy (LEF) under mild regularity conditions (Gouriéroux, Monfort and Trognon, 1984; White, 1994). Based on the result, Wooldridge (2014) reinterprets LIL as a quasi-limited information likelihood (QLIL) and expands its applicability noting that correctly specified regression functions are key assumptions for consistency in LEF.

Apart from robustness of QLIL-based estimators, their relative efficiency relationship is another important issue. Relative efficiency analysis of LIML or equivalent estimator in previous works assume away potentially misspecified likelihoods for both structural and reduced form equations. When the likelihood is allowed to be misspecified, relative efficiency comparison based on the correct specification of likelihood is no longer valid. Analysis accounting for potential misspecification of likelihood is more useful to empirical researchers because economic theories usually do not imply full characterization of distributions, and there is no solid reason to believe QMLE achieves the same asymptotic efficiency as the maximum likelihood estimator.

The purpose of this chapter is to study asymptotic relative efficiency relationship among estimators based on QLIL. Considering a research question raised by Wooldridge (2014), I focus on sufficient condtions for relative efficiency of QLIL maximizer with repect to two-step conditional quasi-likelihood maximizer, which will be called QLIML estimator and control function (CF) approach, respectively. The CF estimator is naturally defined once we take the conventional decomposition of QLIL into structural and reduced form components. The model restriction imposed on QLIL is general enough to include nonlinear models and, in particular, misspecification of likelihoods is allowed up to correctly specified regression functions when fully robust estimation is considered.

The main contributions of this chapter are followings. First, I show there exists a generalized method of moments estimator (GMM-QLIML) based on the all available quasi-scores. The asymptotic variance of GMM-QLIML estimator constitutes a lower bound for those of QLIML and CF in matrix positive semidefinite sense. Second, the QLIML estimator is proved to be as efficient as GMM-QLIML estimator under a set of generalized information matrix equalities. Third, the asymptotic equivalence of LIML and 2SLS is established via linearity of regression functions and L_2 loss function incorporated in normal density. This new proof clearly shows why the equivalence holds without normality or conditional homoskedasticity which is often assumed in the assertion. Sufficient conditions for general equivalence of QLIML and CF are also found. Fourth, in fully robust estimation of correctly specified conditional mean functions, QLIML estimator is shown to be efficient relative to CF estimator if generalized linear model variance assumptions hold with a scaling restriction. In particular, correctly specified conditional moments up to second order are sufficient.

The rest of this chapter is organized as follows. In Section 1.2, basic model restrictions are given with GMM interpretation of QLIML and CF estimators. In Section 1.3, GMM-QLIML estimator is defined and relative efficiency results for QLIML and CF estimator are presented. Section 1.4 contains concluding remarks.

1.2 Model Restrictions

Assume random sampling from a population. For a random draw i, consider the system of equations

$$y_{i1} = \mathbf{f}\left(\mathbf{y}_{i2}, \mathbf{z}_{i1}, u_{i1}; \theta_1, \theta_2\right) \tag{1.1}$$

$$\mathbf{y}_{i2} = \mathbf{g}\left(\mathbf{z}_i, \mathbf{v}_{i2};\theta_2\right) \tag{1.2}$$

where function **f** and **g** are known up to $(p_1 + p_2) \times 1$ vector of parameter $\theta = \left(\theta_1', \theta_2'\right)'$, y_{i1} is a scalar response variable, $\mathbf{y}_{i2}$ is a $1 \times r$ vector of potentially endogenous variables, $\mathbf{z}_i = (\mathbf{z}_{i1}, \mathbf{z}_{i2})$ is $1 \times k$ vector of included/excluded exogenous instruments with $k = k_1 + k_2$, and $(u_{i1}, \mathbf{v}_{i2})$ is a $(1 + r) \times 1$ vector of unobservables.

Under potentially incorrect distributional assumptions for u_1 and $\mathbf{v}_2$, taking log operator on the decomposed quasi-joint likelihood $l\left(y_{i1}|\mathbf{y}_{i2}, \mathbf{z}_i; \theta_1, \theta_2\right) l\left(\mathbf{y}_{i2}|\mathbf{z}_i; \theta_2\right)$ delivers QLIL

$$QLIL = q_1\left(y_{i1}, \mathbf{y}_{i2}, \mathbf{z}_i, \theta_1, \theta_2\right) + q_2\left(\mathbf{y}_{i2}, \mathbf{z}_i, \theta_2\right) \tag{1.3}$$

which offers flexible model specifications. The decomposition is more of ‘composition’ in the sense that q_1 and q_2 do not need to be derived from a single joint quasi-likelihood. For example, Poisson log-likelihood q_1 and normal log-likelihood q_2 can be used as long as quasi-likelihood-driven regression functions are correct: Wooldridge (2014) showed that, in LEF, the key model restrictions for consistent estimation of conditional mean $E_o\left[y_{i1}|\mathbf{y}_{i2}, \mathbf{z}_i\right]$ are

$$E_o\left[y_{i1}|\mathbf{y}_{i2}, \mathbf{z}_i\right] = E_q\left[\mathbf{f}\left(\mathbf{y}_{i2}, \mathbf{z}_{i1}, u_{i1}; \theta_{o1}, \theta_{o2}\right)|\mathbf{y}_{i2}, \mathbf{z}_i\right] \tag{1.4}$$

$$E_o\left[\mathbf{y}_{i2}|\mathbf{z}_i\right] = E_q\left[\mathbf{g}\left(\mathbf{z}_i, \mathbf{v}_{i2};\theta_{o2}\right)|\mathbf{z}_i\right] \tag{1.5}$$

where subscripts ‘o’ and ‘q’ denotes (expectation) operators based on the true and quasi-likelihood, respectively. As long as (1.4) and (1.5) hold, even failure of (1.1) is allowed in consistent estimation as shown by Example 1.2.1 below. In the linear model with quasi-normality (Anderson and Rubin, 1949), linear projection operators $L_o[\cdot|\cdot]$ with appropriate regressors can replace the expectation operators $E_o[\cdot|\cdot]$ when the objects of interest are linear projections rather than conditional mean functions. The decomposed nature of QLIL and correctly specified regression functions typically involve with the existence of a control function. See Wooldridge (2014) for details. Following example demonstrates derivation of QLIL in linear and Probit models, and discusses their robustness properties.

Example 1.2.1 (models with quasi-normality of $(u_{i1}, \mathbf{v}_{i2})$) Consider the following simultaneous equation systems

$$\text{Linear Model:} y_{i1} = \mathbf{y}_{i2}\boldsymbol{\alpha} + \mathbf{z}_{i1}\delta_1 + u_{i1},\ \mathbf{y}_{i2} = \mathbf{z}_i\boldsymbol{\delta}_2 + \mathbf{v}_{i2} \tag{1.6}$$

$$\text{Probit Model:} y_{i1} = 1\left[\mathbf{y}_{i2}\boldsymbol{\alpha} + \mathbf{z}_{i1}\delta_1 + u_{i1} > 0\right],\ \mathbf{y}_{i2} = \mathbf{z}_i\boldsymbol{\delta}_2 + \mathbf{v}_{i2} \tag{1.7}$$

Assume

$$(u_{i1}, \mathbf{v}_{i2})\,|\mathbf{z}_i \sim_q N\,(0, \Sigma)$$

where $V_q\,(u_{i1}|\mathbf{z}_i) = \Sigma_{11}$, $cov_q\,(u_{i1}, \mathbf{v}_{i2}|\mathbf{z}_i) = \Sigma_{12} = \Sigma_{21}^t$, $V_q\,(\mathbf{v}_{i2}|\mathbf{z}_i) = \Sigma_{22}$, $\boldsymbol{\delta}_2 = \left(\boldsymbol{\delta}_{21}', \boldsymbol{\delta}_{22}'\right)'$ is a $k \times r$ matrix and other parameters are defined comformably. In the notation '$X \sim_q \Psi$', the subscript q indicates that the distributional assumption '$X \sim \Psi$' is allowed to be incorrect and is used only for deriving the quasi-likelihood of X or its transformation. The decomposed quasi-likelihoods are easily derived noting that

$$e_{i1}|\,(\mathbf{y}_{i2}, \mathbf{z}_i) \sim_q N\left(0, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)$$

where $e_{i1} \equiv u_{i1} - \mathbf{v}_{i2}\Sigma_{22}^{-1}\Sigma_{21}$. In Probit model, it is assumed that $V_q\,(e_{i1}|\mathbf{y}_{i2}, \mathbf{z}_i) = 1$ for normalization. The quasi-likelihoods for linear and Probit model are given explicitly in Example 1.3.5 and 1.3.15, respectively. Concerning robustness property, there are three things to be mentioned: first, since q_{i1} and q_{i2} in both models belong to LEF, correctly specified conditional mean can be consistently estimated by QLIL-based estimators regardless of the true distribution. Second, in linear model, the conditional mean functions derived from the quasi-likelihood have an interpretation of true linear projections. In particular, the quasi-likelihood-based conditional mean of y_{i1} conditioned on $(\mathbf{y}_{i2}, \mathbf{z}_i)$ can be regarded as the linear projection of y_{i1} on $(\mathbf{y}_{i2}, \mathbf{z}_{i1}, \mathbf{v}_{i2})$

$$E_q\left[y_{i1}|\mathbf{y}_{i2}, \mathbf{z}_i\right] = \mathbf{y}_{i2}\boldsymbol{\alpha} + \mathbf{z}_{i1}\delta_1 + \mathbf{v}_{i2}\Sigma_{22}^{-1}\Sigma_{21} = L_o\left[y_{i1}|\mathbf{y}_{i2}, \mathbf{z}_{i1}, \mathbf{v}_{i2}\right]$$

where $\Sigma_{22}^{-1}\Sigma_{21}$ can be reparameterized to be η for convenience. Since this interpretation is definitional through quasi-scores, even when conditional mean functions are incorrectly specified, $(\boldsymbol{\alpha}, \delta_1)$ is consistently estimated as linear projection coefficients under regularity conditions. Third, the y_{i1} equation in (1.7) of Probit model is not a restrictive condition for consistency. When y_{i1} is a fractional response taking values in [0, 1], the equation may not hold for some observations. Such failure of y_{i1} equation does not necessarily harm consistent estimation of the conditional mean function if Probit response function is correct.

$$E_q\left[y_{i1}|\mathbf{y}_{i2}, \mathbf{z}_i\right] = \Phi\left(\mathbf{y}_{i2}\boldsymbol{\alpha} + \mathbf{z}_{i1}\delta_1 + \mathbf{v}_{i2}\Sigma_{22}^{-1}\Sigma_{21}\right) = E_o\left[y_{i1}|\mathbf{y}_{i2}, \mathbf{z}_i\right]$$

However, Probit response function does not have the robust interpretation of a linear projection as in linear model when it is incorrectly specified. □

Given QLIL, the QLIML and CF estimators are defined as

$$\hat{\theta}_{QLIML} = \arg\max_{\theta} \sum_{i=1}^{N} \left[q_{i1}\left(\theta_1, \theta_2\right) + q_{i2}\left(\theta_2\right)\right] \quad \text{and} \quad \begin{cases} \hat{\theta}_{2,CF} & = \arg\max_{\theta_2} \sum_{i=1}^{N} q_{i2}\left(\theta_2\right) \\ \hat{\theta}_{1,CF} & = \arg\max_{\theta_1} \sum_{i=1}^{N} q_{i1}\left(\theta_1, \hat{\theta}_{2,CF}\right) \end{cases}$$

respectively. Focusing on relative efficiency comparison of these two, it is assumed that both QLIML and CF estimators are consistent and asymptotic normal (CAN) for the true parameter values. Also, we assume that expected quasi-scores uniquely determines true parameters so that GMM interpretation of QLIML and CF estimator is valid. This is a mild assumption since the necessity of LEF for fully robust estimation is shown under enough differentiability of likelihood and interiority of a population maximizer (White, 1994, Theorem 5.6). Consequently, QLIML and CF estimators can be defined as GMM estimators based on quasi-score moment conditions

$$E_o\begin{bmatrix} \frac{\partial}{\partial\theta_1}q_{i1}\left(\theta_{o1}, \theta_{o2}\right) \\ \frac{\partial}{\partial\theta_2}q_{i1}\left(\theta_{o1}, \theta_{o2}\right) + \frac{\partial}{\partial\theta_2}q_{i2}\left(\theta_{o2}\right) \end{bmatrix} = 0 \quad \text{and} \quad E_o\begin{bmatrix} \frac{\partial}{\partial\theta_1}q_{i1}\left(\theta_{o1}, \theta_{o2}\right) \\ \frac{\partial}{\partial\theta_2}q_{i2}\left(\theta_{o2}\right) \end{bmatrix} = 0$$

respectively. Appendix A.1 contains relevant standard regularity conditions (Assumption 1-12). These assumptions are maintained for simplicity. They can be relaxed, for example, to allow non-smooth q_{i1} or q_{i2} via smoothness in the limit and stochastic differentiability (Pollard, 1985).

1.3 Relative Efficency Comparison

The key idea that enables intuitive analysis of relative efficiency relationship is to acknowledge the existence of an estimator whose asymptotic variance constitutes a lower bound for those of QLIML and CF. The estimator is discovered, and called as GMM-QLIML in this chapter. It is defined to be efficient GMM estimator based on a maximal linearly independent set of all quasi-scores available in QLIL. Its construction and potential relative efficiency over QLIML and CF can be easily shown by elementary linear algebra.

Recall the definition of linear independence in the context of moment function space along with its well-known relationship with variance matrix in the following remark.

Definition 1.3.1 A set of scalar moment functions $\{h_l(\mathbf{w}_i, \theta)\}_{l=1}^{L}$ is linearly independent at θ^*if $P\left(\sum_{l=1}^{L} \alpha_l(\theta^*) h_l(\mathbf{w}_i, \theta^*) = 0\right) = 1$ implies $\alpha_l(\theta^*) = 0$ for all l where $\alpha_l(\theta)$ is arbitrary real-valued function of θ. □

Remark 1.3.2 $\{h_l(\mathbf{w}_i, \theta)\}_{l=1}^{L}$ is linearly independent at θ^* if and only if the variance matrix of $\{h_l(\mathbf{w}_i, \theta^*)\}_{l=1}^{L}$ is invertible, assuming that second moments are finite.

Now, consider stacking all available quasi-scores in (2.1):

$$\begin{bmatrix} \frac{\partial}{\partial \theta_1} q_{i1}(\theta_1, \theta_2) \\ \frac{\partial}{\partial \theta_2} q_{i1}(\theta_1, \theta_2) \\ \frac{\partial}{\partial \theta_2} q_{i2}(\theta_2) \end{bmatrix} \tag{1.8}$$

The vector of moment functions (1.8) constitutes, when taken summation or integral, all available first order conditions from factor-by-factor QLIL maximization problem. We might hope conducting efficient GMM on these moment functions yields an estimator efficient relative to QLIML and CF. However, it turns out that (1.8) typically has a singluar variance matrix since the set of moment functions in $\frac{\partial}{\partial \theta} q_{i1}(\theta_1, \theta_2)$ is linearly dependent. The singularity is closely related to the fundamental reason why we need simultaneous equation system: the quasi-likelihood function q_{i1} alone cannot identify θ_{o1} and θ_{o2} in general. To avoid such linear dependence, a maximal linearly idependent set in (1.8) can be used instead. Since moment functions in $\frac{\partial}{\partial\theta_1}q_{i1}\left(\theta_1,\theta_2\right)$ and $\frac{\partial}{\partial\theta_2}q_2\left(\theta_2\right)$ are assumed to be linearly independent by rank condition of CF (Assumption 12), a maximal linearly idependent set can be found by extending the set of CF moment functions.

Definition 1.3.3 GMM-QLIML is an efficient GMM estimator based on a maximal linearly independent set of moment functions at $(\theta_{o1},\theta_{o2})$ in $\left[\begin{array}{ccc}\frac{\partial q_1(\theta_1,\theta_2)}{\partial\theta_1} & \frac{\partial q_1(\theta_1,\theta_2)}{\partial\theta_2} & \frac{\partial q_2(\theta_2)}{\partial\theta_2}\end{array}\right]$:

$$\left[\begin{array}{c}\frac{\partial}{\partial\theta_1}q_1\left(\theta_1,\theta_2\right)\\ \frac{\partial}{\partial\theta_{22}}q_1\left(\theta_1,\theta_2\right)\\ \frac{\partial}{\partial\theta_2}q_2\left(\theta_2\right)\end{array}\right] \tag{1.9}$$

where $\theta_2=\left(\theta_{21}',\theta_{22}'\right)'$, $\theta_{21}\in\mathbb{R}^{p21}$, $\theta_{22}\in\mathbb{R}^{p22}$ with $p_2=p_{21}+p_{22}$ and θ_{22} can be empty. □

The following proposition shows that the GMM-QLIML estimator is asymptotically normal without additional model restrictions other than those of CF and QLIML.

Proposition 1.3.4 Under regularity conditions and identification conditions for CF and QLIML (Assumption 1–12), the GMM-QLIML estimator is asymptotically normal.

$$\sqrt{N}\left(\hat{\theta}_{GMM-QLIML}-\theta_o\right)\overset{d}{\to}N\left(0,\left(A'B^{-1}A\right)^{-1}\right)$$

where

$$A=E\left[\begin{array}{c}\frac{\partial q_{i1}(\theta_o,)}{\partial\theta_1\partial\theta'}\\ \frac{\partial q_{i1}(\theta_o,)}{\partial\theta_{22}\partial\theta'}\\ \frac{\partial q_{i2}(\theta_{o2})}{\partial\theta_2\partial\theta'}\end{array}\right]\text{ and }B=V\left(\begin{array}{c}\frac{\partial q_{i1}(\theta_o,)}{\partial\theta_1}\\ \frac{\partial q_{i1}(\theta_o,)}{\partial\theta_{22}}\\ \frac{\partial q_{i2}(\theta_{o2})}{\partial\theta_2}\end{array}\right)$$

□

Typically, the extra moment functions $\frac{\partial}{\partial\theta_{22}}q_1\left(\theta_1,\theta_2\right)$ are orthogonality conditions between exogeneous part of structural error and overidentifying (k_2-r) instruments. Below example shows determination of $\frac{\partial}{\partial\theta_{22}}q_1\left(\theta_1,\theta_2\right)$ in linear model setting.

Example 1.3.5 (Linear Model) Consider linear model in Example 1.2.1. For notational convenience, define the following

$$\begin{aligned}
\mathbf{v}_{i2}\left(\boldsymbol{\delta}_2\right) &\equiv \mathbf{y}_{i2} - \mathbf{z}_i \boldsymbol{\delta}_2 \\
e_i\left(\theta\right) &\equiv y_{i1} - \mathbf{y}_{i2}\boldsymbol{\alpha} - \mathbf{z}_{i1}\delta_1 - \mathbf{v}_{i2}\left(\boldsymbol{\delta}_2\right)\Sigma_{22}^{-1}\Sigma_{21} \\
\sigma_{11|2}\left(\theta\right) &\equiv \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \\
h_i\left(\theta\right) &\equiv e_i\left(\theta\right)^2 - \sigma_{11|2}\left(\theta\right)
\end{aligned}$$

where parameters are defined comformably. $e_i\left(\theta\right)$ can be interpreted as remaining part of structural error after endogeneous variation $\mathbf{v}_{i2}$ is projected out. The quasi-log-likelihoods are

$$\begin{aligned}
q_{i1}\left(\theta_1, \theta_2\right) &= -\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln \sigma_{11|2}\left(\theta\right) - \frac{1}{2}e_i\left(\theta\right)^2 \sigma_{11|2}\left(\theta\right)^{-1} \\
q_{i2}\left(\theta_2\right) &= -\frac{k}{2}\ln 2\pi - \frac{1}{2}\ln\left|\Sigma_{22}\right| - \frac{1}{2}\mathbf{v}_{i2}\left(\boldsymbol{\delta}_2\right)\Sigma_{22}^{-1}\mathbf{v}_{i2}\left(\boldsymbol{\delta}_2\right)'
\end{aligned}$$

and quasi-scores can be expressed as follows

$$\begin{aligned}
\frac{\partial q_1}{\partial \theta_1} &= \begin{bmatrix} \sigma_{11|2}\left(\theta\right)^{-1} e_i\left(\theta\right)\mathbf{y}_2' \\ \sigma_{11|2}\left(\theta\right)^{-1} e_i\left(\theta\right)\mathbf{z}_1' \\ -\sigma_{11|2}\left(\theta\right)^{-2} h_i\left(\theta\right)\Sigma_{22}^{-1}\Sigma_{21} + \sigma_{11|2}\left(\theta\right)^{-1} e_i\left(\theta\right)\Sigma_{22}^{-1}\mathbf{v}_{i2}\left(\boldsymbol{\delta}_2\right)' \\ \frac{1}{2}\sigma_{11|2}\left(\theta\right)^{-2} h_i\left(\theta\right) \end{bmatrix} \\
\frac{\partial q_1}{\partial \theta_2} &= \begin{bmatrix} -\sigma_{11|2}\left(\theta\right)^{-1} e_i\left(\theta\right)\left[\Sigma_{22}^{-1}\Sigma_{21} \otimes \mathbf{z}'\right] \\ L_r\left[\Sigma_{22}^{-1} \otimes \Sigma_{22}^{-1}\right] D\left(\theta\right) \end{bmatrix} \\
\frac{\partial q_2}{\partial \theta_2} &= \begin{bmatrix} \left[I_r \otimes \mathbf{z}'\right]\Sigma_{22}^{-1}\mathbf{v}_2\left(\boldsymbol{\delta}_2\right)' \\ \frac{1}{2}L_r vec\left(\Sigma_{22}^{-1}\mathbf{v}_{i2}\left(\boldsymbol{\delta}_2\right)'\mathbf{v}_{i2}\left(\boldsymbol{\delta}_2\right)\Sigma_{22}^{-1} - \Sigma_{22}^{-1}\right) \end{bmatrix}
\end{aligned}$$

where $D\left(\theta\right) = \left[\Sigma_{21} \otimes \Sigma_{21}\right]\frac{1}{2}\sigma_{11|2}\left(\theta\right)^{-2} h_i\left(\theta\right) - \left[\Sigma_{21} \otimes \mathbf{v}_{i2}\left(\boldsymbol{\delta}_2\right)'\right]\sigma_{11|2}\left(\theta\right)^{-1} e_i\left(\theta\right),\ \theta_1 = \left(\alpha', \delta_1', \Sigma_{21}', \Sigma_{11}\right)'$, $\theta_2 = \left(vec\left(\boldsymbol{\delta}_2\right)', vech\left(\Sigma_{22}\right)'\right)'$ and L_r is a $\frac{r(r+1)}{2} \times r^2$ elemination matrix (Section 5.7.3, Turkington, 2014) . To determine $\frac{\partial q_1(\theta)}{\partial \theta_{22}}$, we should find a set of moment functions in $\frac{\partial q_1}{\partial \theta_2}$ that cannot be expressed by a linear combination of moment functions in $\frac{\partial q_1}{\partial \theta_1}$ and $\frac{\partial q_2}{\partial \theta_2}$ at true parameter values. If $\Sigma_{o21} = 0$, then $\frac{\partial q_1}{\partial \theta_2} = 0$ and θ_{22} is empty. For now, assume $\Sigma_{o21} \neq 0$

and, at least one element of $\Sigma_{o22}^{-1}\Sigma_{o21}$, say i_o th component, is nonzero. Then, by some tedious algebra[1], the extra moments can be shown to be at most

$$\frac{\partial q_1(\theta)}{\partial \theta_{22}} = -\sigma_{1|2}(\theta)^{-1} e_i(\theta) \left(\Sigma_{22}^{-1}\Sigma_{21}\right)_{i_o} \mathbf{z}'_{2,-r} \tag{1.10}$$

where '$-r$' denotes 'leaving r instruments out' in $\mathbf{z}_2$. Suppose there exists enough variation in $\mathbf{z}_2$ so that (1.10) is indeed $\frac{\partial q_1(\theta)}{\partial \theta_{22}}$. Since GMM-QLIML moment functions constitute a basis of linear vector space spanned by (1.8), it can be shown that any choice of extra moments yields asymptotically equivalent estimator. If the model is just-identified ($k_2 = r$), or if there exists no endogeneity ($\Sigma_{o21} = 0$), then θ_{22} is empty, and GMM-QLIML, CF and QLIML are asymptotically equivalent to each other. □

Example 1.3.5 illustrates the following general proposition.

Proposition 1.3.6 Under regularitiy conditions and identification conditions (Assumption 1–12), (a) Each choice of $\frac{\partial q_1(\theta)}{\partial \theta_{22}}$ yields asymptotically equivalent GMM-QLIML estimator. (b) If θ_{22} is empty, then GMM-QLIML, QLIML and CF are asymptotically equivalent. □

Example 1.3.5 shows that a preliminary step is required for GMM-QLIML to be used in practice. In the example, it is necessary to test whether there exists a component of $\Sigma_{22}^{-1}\Sigma_{21}$ significantly different from zero. Then, the extra moment functions will be chosen correspondingly. This preliminary procedure probably is not very appealing to practitioners. A practical approach in general would be to employ a generalized inverse matrix for optimal weighting and resolve singularity issue. Also, in this specific example, we can consider including moment condition

[1] It is easy to see that $\frac{\partial q_1}{\partial \mathbf{\Sigma}_{22}}$ (the second part of $\frac{\partial q_1}{\partial \theta_2}$) is a linear combination of $\frac{\partial q_1}{\partial \Sigma_{21}}$ and $\frac{\partial q_1}{\partial \Sigma_{11}}$ (the third and fourth part of $\frac{\partial q_1}{\partial \theta_1}$).In $\frac{\partial q_1}{\partial \boldsymbol{\delta}_2}$ (the first part of $\frac{\partial q_1}{\partial \theta_2}$), all moments with $\mathbf{z}_1$ can be generated by $\frac{\partial q_1}{\partial \boldsymbol{\delta}_1}$ (the second part of $\frac{\partial q_1}{\partial \theta_1}$). So, we are left with moments with $\mathbf{z}_2$. Among these, due to explicit linear relationship $\mathbf{y}_2 = \mathbf{Z}_i \boldsymbol{\delta}_2 + \mathbf{v}_{i2}(\boldsymbol{\delta}_2)$, only $(k_2 - r)$ moments at most can be included in $\frac{\partial q_1(\theta)}{\partial \theta_{22}}$. For all $(k_2 - r)$ moments to be included, we need enough variation in instruments.

without $\left(\Sigma_{22}^{-1}\Sigma_{21}\right)_{io}$ term

$$\frac{\partial q_1(\theta)}{\partial \theta_{22}} = -\sigma_{11|2}(\theta)^{-1} e_i(\theta) \mathbf{z}'_{2,-r} \tag{1.11}$$

and the resulting optimal GMM estimator is more efficient relative to GMM-QLIML though it may require additional model restrictions in general.

GMM-QLIML has an important role in relative efficiency study while GMM with (1.11) has more practical usage. As a basis of linear space spanned by QLIML and CF moment functions, the asymptotic variance of GMM-QLIML forms a sharper lower bound for those of QLIML and CF. Since eliminating $\Sigma_{22}^{-1}\Sigma_{21}$ from (1.10) is equivalent to adding extra information that is not used either by QLIML or CF when $\Sigma_{o21} = 0$, the asymptotic variance of optimal GMM estimator using (1.11) can strictly smaller than that of GMM-QLIML in matrix positive definite sense. This delicate distinction offers a convenient general framework of relative efficiency comparison.

Potential relative efficiency gain of GMM-QLIML with respect to QLIML and CF is clear from its definition. It is worth noting that such potential improvement is not based on additional model restrictions as shown in Proposition 1.3.4. When efficiency gain is present, it is implied that QLIML and CF make use of only a part of information that GMM-QLIML uses. In such a case, relative efficiency comparison of QLIML and CF is not obvious in general. When GMM-QLIML is equivalent to either QLIML or CF, one can conclude that the one equivalent to GMM-QLIML is superior than the other one. Conditions under which GMM-QLIML is equivalent to each estimator can be derived by applying moment redundancy conditions (Breuch, Qian, Schmidt and Wyhowski, 1999; BQSW). In the following propostions, denote $V_{est^r} \equiv Avar\left(\sqrt{N}\left(\hat{\theta}_{est^r} - \theta_o\right)\right)$, $V_{est^r}^{\theta_S} \equiv Avar\left(\sqrt{N}\left(\hat{\theta}_{S,est^r} - \theta_{oS}\right)\right)$ for partition $\theta = (\theta_S, \theta_{-S})$, and $\partial q_{il}^o = \partial q_{il}(\theta_o)$ for $l = 1, 2$.

Proposition 1.3.7 Assume that Assumptions 1-12 hold and that θ_{22} is nonempty. Then

(a) $V_{GMM-QLIML} \preceq V_{QLIML}, V_{CF}$

(b) $V_{GMM-QLIML} = V_{CF}$ if and only if

$$E_o\left[\frac{\partial q_{i1}^o}{\partial \theta_{22} \partial \theta'}\right] = cov_o\left(\frac{\partial q_{i1}^o}{\partial \theta_{22}}, \begin{bmatrix} \frac{\partial q_{i1}^o}{\partial \theta_1} \\ \frac{\partial q_{i2}^o}{\partial \theta_2} \end{bmatrix}\right) V_o\begin{pmatrix} \frac{\partial q_{i1}^o}{\partial \theta_1} \\ \frac{\partial q_{i2}^o}{\partial \theta_2} \end{pmatrix}^{-1} E_o\begin{bmatrix} \frac{\partial q_{i1}^o}{\partial \theta_1 \partial \theta'} \\ \frac{\partial q_{i2}^o}{\partial \theta_2 \partial \theta'} \end{bmatrix}$$

(c) $V_{GMM-QLIML} = V_{QLIML}$ if and only if

$$E_o\left[\frac{\partial q_{i1}^o}{\partial \theta_2^* \partial \theta'}\right] = cov_o\left(\frac{\partial q_{i1}^o}{\partial \theta_2^*}, \frac{\partial \left(q_{i1}^o + q_{i2}^o\right)}{\partial \theta}\right) V_o\left(\frac{\partial \left(q_{i1}^o + q_{i2}^o\right)}{\partial \theta}\right)^{-1} E_o\left[\frac{\partial \left(q_{i1}^o + q_{i2}^o\right)}{\partial \theta \partial \theta'}\right]$$

where θ_2^* is a subvector of θ_2 such that $\frac{\partial q_1^o}{\partial \theta_2^*}, \frac{\partial q_{i1}^o}{\partial \theta_1}$ and $\frac{\partial q_{i1}^o}{\partial \theta_2} + \frac{\partial q_{i2}^o}{\partial \theta_2}$ are maximal linearly independent.

□

Remark 1.3.8 (b) and (c) can be derived for arbitrary subvector θ_S of $\theta = (\theta_S, \theta_{-S})$. Corresponding results are given in the appendix.

The equivalence conditions (b) and (c) characterize when the extra moments in GMM-QLIML contain no useful information about parameters. Rigorously put, they describe cases where the orthogonal complement of QLIML or CF moment functions in the linear span of (1.8) does not contain additional information about parameters. One interesting implication of (c) is that a set of generalized information matrix equalities (GIME; Wooldridge, 2010) for q_1, q_2 and $q_1 + q_2$ with some common scaling factor $\tau > 0$ is sufficient for QLIML to be efficient relative to CF:

$$V_o\left(\frac{\partial q_1^o}{\partial \theta}\right) = \tau E_o\left[-\frac{\partial q_1^o}{\partial \theta \partial \theta'}\right]$$

$$V_o\left(\frac{\partial q_2^o}{\partial \theta_2}\right) = \tau E_o\left[-\frac{\partial q_2^o}{\partial \theta_2 \partial \theta_2'}\right]$$

$$V_o\left(\frac{\partial \left(q_1^o + q_2^o\right)}{\partial \theta}\right) = \tau E_o\left[-\frac{\partial \left(q_1^o + q_2^o\right)}{\partial \theta \partial \theta'}\right]$$

Note that this result is stronger than one in previous studies under correctly specified likelihoods. Even if QLIML is not a maximum likelihood estimator, it is efficient relative to CF whenever a finite number of moment conditions in GIMEs are met.

Following corollary contains relevant implications of Proposition 1.3.7 regarding GIMEs. In particular, it claims an if and only if condition for CF and QLIML estimator of θ_{11} to be asymptotically equivalent under GIMEs where $\theta_1 = (\theta_{11}, \theta_{12})$.

Corollary 1.3.9 *Assume that Assumptions 1-12 hold and that θ_{22} is nonempty. If generalized information matrix equalities hold for each factor of likelihood and joint likelihood with the same scaling factor, we have*

$$(a)\ V_{GMM-QLIML} = V_{QLIML} \preceq V_{CF} \ \textit{(in particular, } V_{QLIML} \neq V_{CF}\textit{)}$$

$$(b)\ V^{\theta_{11}}_{GMM-QLIML} = V^{\theta_{11}}_{QLIML} = V^{\theta_{11}}_{CF} \ \textit{if and only if}$$

$$\begin{aligned} 0_{p_{22}\times p_{11}} = & \left[E_o\left[\frac{\partial q^o_{i1}}{\partial\theta_{22}\partial\theta_2'}\right] - E_o\left[\frac{\partial q^o_{i1}}{\partial\theta_{22}\partial\theta_1'}\right] V_o\left(\frac{\partial q^o_{i1}}{\partial\theta_1}\right)^{-1} E_o\left[\frac{\partial q^o_{i1}}{\partial\theta_1\partial\theta_2'}\right]\right] \\ & \times \left[R_{21}E_o\left[\frac{\partial q^o_{i1}}{\partial\theta_{12}\partial\theta_{11}'}\right] + R_{22}E_o\left[\frac{\partial q^o_{i1}}{\partial\theta_2\partial\theta_{11}'}\right]\right] \end{aligned}$$

where R_{21} and R_{22} are defined in the proof. □

These results are useful to study asymptotic equivalence of QLIML and CF since, if there exists a case where QLIML and CF are asymptotically equivalent in general, then it must also be the case under GIMEs. The result (a) of Corollary 1.3.9 shows that, when θ_{22} is nonempty, QLIML and CF are never asymptotically equivalent for all element of θ. But this does not rule out the case where QLIML and CF are asymptotically equivalent for strict subvector of θ. The formula in Corollary 1.3.9 (b) (and another one given in Proposition 1.3.13 later) informs us about key conditions for QLIML and CF to be asymptotically equivalent for subvector θ_{11} of θ_1: It seems that some part of the expected cross partials $E_o\left[\frac{\partial q^o_{i1}}{\partial\theta_{12}\partial\theta_{11}'}\right]$ and $E_o\left[\frac{\partial q^o_{i1}}{\partial\theta_2\partial\theta_{11}'}\right]$ should vanish to have general equivalence. Based on this observation, following proposition explicitly claims a condition under which QLIML and CF are asymptotic equivalent for a subvector of θ. The well-known result of asymptotic equivalence of LIML and 2SLS is an implication.

Proposition 1.3.10 Assume that Assumptions 1-12 hold. Let (ζ_1, ζ_2) be a partition of θ. If there exists $p \times p$ invertible matrices $T_1(\theta)$ and $T_2(\theta)$ such that

$$T_1(\theta)\begin{bmatrix} \frac{\partial}{\partial\theta_1} q_1(\theta_1, \theta_2) \\ \frac{\partial}{\partial\theta_2} q_1(\theta_1, \theta_2) + \frac{\partial}{\partial\theta_2} q_{i2}(\theta_2) \end{bmatrix} = \begin{bmatrix} m_1(\zeta_1, \zeta_2) \\ m_2(\zeta_1, \zeta_2) \end{bmatrix}$$

$$T_2(\theta)\begin{bmatrix} \frac{\partial}{\partial\theta_1} q_1(\theta_1, \theta_2) \\ \frac{\partial}{\partial\theta_2} q_2(\theta_2) \end{bmatrix} = \begin{bmatrix} m_1(\zeta_1, \zeta_2) \\ m_3(\zeta_1, \zeta_2) \end{bmatrix}$$

where $m_1(\zeta_1, \zeta_2)$ identifies ζ_{o1} given ζ_{o2}, $E_o\left[\frac{\partial m_1(\zeta_{o1}, \zeta_{o2})}{\partial \zeta_2}\right] = 0$ and $E_o\left[\frac{\partial m_{ig}(\zeta_{o1}, \zeta_{o2})}{\partial \zeta_2'}\right]$ is invertible for $g = 2, 3$, then QLIML and CF estimator for ζ_1 are asymptotically equivalent. □

Corollary 1.3.11 *LIML and 2SLS are asymptotically equivalent for* (α, δ_1) . □

The asymptotic equivalence of LIML and 2SLS is mainly due to linearity of regression functions and L_2 loss function endowed in normal density. The intuition behind the proof is that $\mathbf{v}_2$ does not need to be controlled as regressors to estimate (α, δ_1): the orthogonality conditions $\frac{\partial q_1}{\partial \alpha}$ and $\frac{\partial q_1}{\partial \delta_1}$ in Example 1.3.5 can be transformed into

$$\begin{bmatrix} (y_{i1} - \mathbf{y}_{i2}\boldsymbol{\alpha} - \mathbf{z}_{i1}\delta_1)\,\delta_{22}'\mathbf{z}_2' \\ (y_{i1} - \mathbf{y}_{i2}\boldsymbol{\alpha} - \mathbf{z}_{i1}\delta_1)\,\mathbf{z}_1' \end{bmatrix} \tag{1.12}$$

by an invertible linear map. Treating $\left(\alpha', \delta_1'\right)'$ as ζ_1 in Proposition 1.3.10, the equivalence follows. Clearly, neither normality nor conditional homoskedasticity is needed for the result, which is not very well recognized in the literature. Amemiya (1984) proves the equivalence under conditional homoskedastic non-normal errors and non-random instruments but his argument is, in fact, valid without assuming conditional homoskedasticity. In nonlinear models such as probit, the regression function does not allow the control function part to vanish as linear model does in (1.12). Also, when the loss function is other than L_2, for example, L_1 as in median regression with tick-exponential family (Komunjer, 2009), then, even if the regression function is linear, again there exists no invertible linear transformation of quasi-scores that eliminates the control function part in general. Thus equivalence of QLIML and CF does not seem to hold for nonlinear regression models.

Apart from linearity of regression function and L_2 loss function, another condition for general asymptotic equivalence of QLIML and CF for θ_1 is

$$E_o\left[-\frac{\partial q_1^o}{\partial\theta_1\partial\theta_2}\right]=cov_o\left(\frac{\partial q_1^o}{\partial\theta_1},\frac{\partial q_2^o}{\partial\theta_2}\right) \tag{1.13}$$

together with $cov_o\left(\frac{\partial q_1}{\partial\theta_1},\frac{\partial q_2}{\partial\theta_2}\right)=0$. The equivalence is easily proved by taking $T_1(\theta)=T_2(\theta)=I_p$, $\zeta_1=\theta_1$ and $\zeta_2=\theta_2$ in Proposition 1.3.10. A set of sufficient conditions for (1.13) is well-known to be (a) q_2 is correctly specified log-likelihood for $\mathbf{w}_{i2}$ and (b) $w_{i1}\perp\!\!\!\perp\mathbf{w}_{i2}|\,\mathbf{z}_i$ where $q_1=q_1(w_{i1},\mathbf{w}_{i2},\mathbf{z}_i,\theta_1,\theta_2)$ and $q_2=q_2(\mathbf{w}_{i2},\mathbf{z}_i,\theta_2)$ for some random variable $(w_{i1},\mathbf{w}_{i2})$. This is a fairly general condition applicable to numerous models. However, it should be noted that w_{i1} cannot be a latent error term such as u_{i1} or $u_{i1}-\mathbf{v}_{i2}\eta$ in Probit model of Example 1.2.1 since q_1 is required to be a quasi-log-likelihood of w_{i1} given $(\mathbf{w}_{i2},\mathbf{z}_i)$.

The next two propositions refine GIMEs to derive weaker conditions for relative efficiency of QLIML. Proposition 1.3.12 helps reducing the number of conditions in GIMEs by treating nuisance parameters as known. Multivariate normal log-likelihood becomes a member of LEF when this result is applicable to its variance parameters. Proposition 1.3.13 relaxes common scaling factors in GIMEs. When different scaling factors for q_1 and q_2 are allowed, $\tau_1\le\tau_2$ is shown to be sufficient for relative efficiency of QLIML for θ_1. Note that, with different scaling factors, having the GIME hold in both models does not necessarily imply asymptotic equivalence of QLIML and GMM-QLIML. Following Zhang (2005), the Schur complement of B in A is denoted as A/B for notational convenience.

Proposition 1.3.12 Assume that Assumptions 1-12 hold. Suppose there exists (l_1+l_2) nuisance parameters $\lambda=(\lambda_1,\lambda_2)$ such that

$$E_o\left[\frac{\partial q_{i1}(\theta_{o1},\theta_{o2},\lambda_{o1},\lambda_{o2})}{\partial\theta\partial\left(\lambda_1',\lambda_2'\right)}\right]=0_{p\times(l_1+l_2)}$$

$$E_o\left[\frac{\partial q_{i2}(\theta_{o2},\lambda_{o2})}{\partial\theta_2\partial\lambda_2'}\right]=0_{p_2\times l_2}$$

Then, V_{QLIML}^{θ} and V_{CF}^{θ} are not affected by treating λ as known and redefining $\tilde{q}_{i1}(\theta)=q_{i1}(\theta,\lambda_o)$ and $\tilde{q}_{i2}(\theta_2)=q_{i2}(\theta_2,\lambda_{o2})$. Moreover, if GMM-QLIML moment function (1.9) contains exactly $(l_1 + l_2)$ scores regarding λ, then $V^{\theta}_{GMM-QLIML}$ is also not affected by the redefinition. □

Proposition 1.3.13 Assume that Assumptions 1-12 hold. Suppose GIMEs with scaling factors τ_1 and τ_2 for quasi-log-likelihood q_1 and q_2, respectively. Also, assume $cov_o\left(\frac{\partial q_1^o}{\partial \theta}, \frac{\partial q_2^o}{\partial \theta_2}\right) = 0$. Then, $V^{\theta_1}_{CF} - V^{\theta_1}_{QLIML}$ is equal to

$$E_o\left[-\frac{\partial q_1^o}{\partial \theta_1 \partial \theta_1'}\right]^{-1} E_o\left[-\frac{\partial q_1^o}{\partial \theta_1 \partial \theta_2'}\right]\left[\tau_2 W_1 + (\tau_1 - \tau_2) W_2\right] E_o\left[-\frac{\partial q_1^o}{\partial \theta_2 \partial \theta_1'}\right] E_o\left[-\frac{\partial q_1^o}{\partial \theta_1 \partial \theta_1'}\right]^{-1}$$

where

$$\begin{aligned} W_1 &= E_o\left[-\frac{\partial q_2^o}{\partial \theta_2 \partial \theta_2'}\right]^{-1} - \left[E_o\left[-\frac{\partial\left(q_1^o + q_2^o\right)}{\partial \theta \partial \theta'}\right] \Big/ E_o\left[-\frac{\partial q_1^o}{\partial \theta_1 \partial \theta_1'}\right]\right]^{-1} \\ W_2 &= -\left[E_o\left[-\frac{\partial\left(q_1^o + q_2^o\right)}{\partial \theta \partial \theta'}\right] \Big/ E_o\left[-\frac{\partial q_1^o}{\partial \theta_1 \partial \theta_1'}\right]\right]^{-1}\left[E_o\left[-\frac{\partial q_1^o}{\partial \theta \partial \theta'}\right] \Big/ E_o\left[-\frac{\partial q_1^o}{\partial \theta_1 \partial \theta_1'}\right]\right] \\ &\quad \times \left[E_o\left[-\frac{\partial\left(q_1^o + q_2^o\right)}{\partial \theta \partial \theta'}\right] \Big/ E_o\left[-\frac{\partial q_1^o}{\partial \theta_1 \partial \theta_1'}\right]\right]^{-1} \end{aligned}$$

In particular, $\tau_2 \geq \tau_1$ implies $V^{\theta_1}_{CF} \succeq V^{\theta_1}_{QLIML}$. □

When $\tau_1 \neq \tau_2$, GIME for $(q_1 + q_2)$ is not met, and QLIML is not optimally weighting $\frac{\partial q_1}{\partial \theta_2}$ and $\frac{\partial q_2}{\partial \theta_2}$ as GMM-QLIML does. In this sense, Proposition 1.3.13 helps us to understand the situation where complete GIMEs start to break down. Contrary to the unambiguous case of $\tau_1 \leq \tau_2$ in the proposition, when $\tau_1 > \tau_2$, the expression $\tau_2 W_1 + (\tau_1 - \tau_2) W_2$ is indefinite in general. This observation explains why general efficiency ordering of QLIML and CF is not obvious without any form of GIMEs.

The next proposition shows how the general theory applies in a class of fully robust models specified with multivariate normal q_2. It is one of the most frequently used specification that attains fully robust estimation but not the only class of models that results can be applied to. It is shown that correct specification of conditional means and GLM variance assumptions with a restriction on scaling factors are sufficient for relative efficiency of QLIML for the structural parameters. In particular, correctly specified conditional moments up to second order are sufficient.

Proposition 1.3.14 Assume that Assumptions 1-12 hold. Suppose that q_1 is a member of LEF with conditional mean $G(\mathbf{y}_2, \mathbf{z}_1, \mathbf{v}_2, \theta_1)$, and that q_2 is a multivariate normal density for linear reduced form equations. In other words,

$$\begin{aligned} q_{i1}(\theta_1, \theta_2) &= a(G(\mathbf{y}_2, \mathbf{z}_1, \mathbf{v}_2, \theta_1)) + b(y_{i1}) + y_{i1} c(G(\mathbf{y}_2, \mathbf{z}_1, \mathbf{v}_2, \theta_1)) \\ q_{i2}(\theta_2) &= -\frac{k}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_{22}| - \frac{1}{2}\mathbf{v}_{i2}\Sigma_{22}^{-1}\mathbf{v}_{i2}' \end{aligned}$$

where a, b, c and G are smooth enough functions, $\mathbf{v}_2 = \mathbf{y}_2 - \mathbf{z}\boldsymbol{\delta}_2$ and $\theta_2 = \left(vec(\boldsymbol{\delta}_2)', vech(\Sigma_{22})'\right)'$. Assume that $E_q(y_1|\mathbf{y}_2, \mathbf{z})$ and $E_q(\mathbf{y}_2|\mathbf{z})$ are correctly specified. Then, $V_o(y_1|\mathbf{y}_2, \mathbf{z}) = \tau_1 V_q(y_1|\mathbf{y}_2, \mathbf{z})$ and $V_o(\mathbf{y}_2, |\mathbf{z}) = \tau_2 V_q(\mathbf{y}_2|\mathbf{z})$ with $0 < \tau_1 \leq \tau_2$ is sufficient for QLIML to be efficient relative to CF for θ_1. □

As a special case of Proposition 1.3.14, the next example considers a probit response function with endogeneous explanatory variables. Specifically, Proposition 1.3.14 implies that relative efficiency of QLIML holds under a much weaker condition than correct specification of likelihood given in Rivers-Vuong (1988), and this result is new in the literature.

Example 1.3.15 (Rivers-Vuong, 1988) Consider probit model in Example 2.1. Note that y_1 is not restricted to be binary response as long as the probit response function is correct. Assume regularity and identification conditions (Assumption 1–12). For computational convienience, impose following reparametrization

$$\eta \equiv \Sigma_{22}^{-1}\Sigma_{21}$$

along with normalization of $e_1 = u_1 - \mathbf{v}_2\eta$. Then, quasi-likelihood can be simplified as following

$$\begin{aligned} q_1(\theta_1, \theta_2) &= (1 - y_1)\log[1 - \Phi(\mathbf{w}(\theta))] + y_1 \log\Phi(\mathbf{w}(\theta)) \\ q_{i2}(\theta_2) &= -\frac{k}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_{22}| - \frac{1}{2}\mathbf{v}_{i2}(\boldsymbol{\delta}_2)\Sigma_{22}^{-1}\mathbf{v}_{i2}(\boldsymbol{\delta}_2)' \end{aligned}$$

where $\theta_1 = \left(\alpha', \delta_1', \eta'\right)'$, $\theta_2 = \left(vec(\boldsymbol{\delta}_2)', vech(\Sigma_{22})'\right)'$, $\mathbf{x} = \begin{bmatrix} \mathbf{y}_2 & \mathbf{z}_1 & \mathbf{v}_2 \end{bmatrix}$ and $\mathbf{w}(\theta) = \mathbf{x}\theta_1$.

Taking derivatives, quasi-scores can be expressed as

$$\begin{aligned}
\frac{\partial q_1}{\partial \theta_1} &= \frac{y_1 - \Phi\left(\mathbf{w}\left(\theta\right)\right)}{\left[1 - \Phi\left(\mathbf{w}\left(\theta\right)\right)\right]\Phi\left(\mathbf{w}\left(\theta\right)\right)}\phi\left(\mathbf{w}\left(\theta\right)\right)\begin{bmatrix} \mathbf{y}_2' \\ \mathbf{z}_1' \\ \mathbf{v}_2' \end{bmatrix} \\
\frac{\partial q_1}{\partial \theta_2} &= -\frac{y_1 - \Phi\left(\mathbf{w}\left(\theta\right)\right)}{\left[1 - \Phi\left(\mathbf{w}\left(\theta\right)\right)\right]\Phi\left(\mathbf{w}\left(\theta\right)\right)}\phi\left(\mathbf{w}\left(\theta\right)\right)\begin{bmatrix} -\eta \otimes \mathbf{z}' \\ 0_{\frac{r(r+1)}{2}\times 1} \end{bmatrix} \\
\frac{\partial q_2}{\partial \theta_2} &= \begin{bmatrix} \left[I_r \otimes \mathbf{z}'\right]\Sigma_{22}^{-1}\mathbf{v}_2\left(\boldsymbol{\delta}_2\right)' \\ \frac{1}{2}L_r vec\left(\Sigma_{22}^{-1}\mathbf{v}_{i2}\left(\boldsymbol{\delta}_2\right)'\mathbf{v}_{i2}\left(\boldsymbol{\delta}_2\right)\Sigma_{22}^{-1} - \Sigma_{22}^{-1}\right) \end{bmatrix}
\end{aligned}$$

Assume GMM-QLIML extra moment functions are

$$\frac{\partial q_1}{\partial \theta_{22}} = -\frac{y_1 - \Phi\left(\mathbf{w}\left(\theta\right)\right)}{1 - \Phi\left(\mathbf{w}\left(\theta\right)\right)\Phi\left(\mathbf{w}\left(\theta\right)\right)}\phi\left(\mathbf{w}\left(\theta\right)\right)\left[\eta_i \mathbf{z}_{2,-r}'\right]$$

where $\eta_{oi} \neq 0$. To derive conditions under which QLIML is efficient relative to CF, first note that Σ_{22} is nuisance parameter, that is, under correctly specifed regression functions,

$$\begin{aligned}
E_o\left[\frac{\partial q_1^o}{\partial \theta_1 \partial vech\left(\Sigma_{22}\right)'}\right] &= 0 \\
E_o\left[-\frac{\partial q_2^o}{\partial vec\left(\boldsymbol{\delta}_2\right)\partial vech\left(\Sigma_{22}\right)'}\right] &= E\left[\left[I_r \otimes \mathbf{z}'\right]\left[\mathbf{v}_2\Sigma_{o22}^{-1} \otimes \Sigma_{o22}^{-1}\right]\right]L_r' = 0
\end{aligned}$$

Therefore, we can assume Σ_{22} is known, that is, redefine $\theta_2 = \delta_2$. Then, expected Hessian and score outer product matrices are

$$\begin{aligned}
E_o\left[-\frac{\partial q_1^o}{\partial \theta_1 \partial \theta_1'}\right] &= E_o\left[\frac{\left[\phi\left(\mathbf{w}\left(\theta_o\right)\right)\right]^2}{\left[\Phi\left(\mathbf{w}\left(\theta_o\right)\right)\right]\left[1 - \Phi\left(\mathbf{w}\left(\theta_o\right)\right)\right]}\mathbf{x}_i^t\mathbf{x}_i\right] \\
E_o\left[-\frac{\partial q_1^o}{\partial \theta_1 \partial \theta_2'}\right] &= E_o\left[\frac{\left[\phi\left(\mathbf{w}\left(\theta_o\right)\right)\right]^2}{\left[\Phi\left(\mathbf{w}\left(\theta_o\right)\right)\right]\left[1 - \Phi\left(\mathbf{w}\left(\theta_o\right)\right)\right]}\mathbf{x}^t\left[-\eta_o' \otimes \mathbf{z}\right]\right] \\
E_o\left[-\frac{\partial q_1^o}{\partial \theta_2 \partial \theta_2'}\right] &= E_o\left[\frac{\left[\phi\left(\mathbf{w}\left(\theta_o\right)\right)\right]^2}{\left[\Phi\left(\mathbf{w}\left(\theta_o\right)\right)\right]\left[1 - \Phi\left(\mathbf{w}\left(\theta_o\right)\right)\right]}\left[\eta_o\eta_o' \otimes \mathbf{z}'\mathbf{z}\right]\right] \\
E_o\left[-\frac{\partial q_2^o}{\partial vec\left(\boldsymbol{\delta}_2\right)\partial vec\left(\boldsymbol{\delta}_2\right)'}\right] &= E_o\left[\Sigma_{o22}^{-1} \otimes \mathbf{z}'\mathbf{z}\right]
\end{aligned}$$

and

$$\begin{aligned}
V_o\left(\frac{\partial q_1^o}{\partial \theta_1}\right) &= E_o\left[\left(\frac{y_1-\Phi\left(\mathbf{w}\left(\theta_o\right)\right)}{\Phi\left(\mathbf{w}\left(\theta_o\right)\right)\left[1-\Phi\left(\mathbf{w}\left(\theta_o\right)\right)\right]}\right)^2 \phi\left(\mathbf{w}\left(\theta_o\right)\right)^2 \mathbf{x}^t \mathbf{x}\right] \\
cov_o\left(\frac{\partial q_1^o}{\partial \theta_1}, \frac{\partial q_1^o}{\partial \theta_2}\right) &= E_o\left[\left(\frac{y_1-\Phi\left(\mathbf{w}\left(\theta_o\right)\right)}{\left[1-\Phi\left(\mathbf{w}\left(\theta_o\right)\right)\right]\Phi\left(\mathbf{w}\left(\theta_o\right)\right)}\right)^2 \left[\phi\left(\mathbf{w}\left(\theta_o\right)\right)\right]^2 \mathbf{x}^t \left[-\eta_o' \otimes \mathbf{z}\right]\right] \\
V_o\left(\frac{\partial q_1^o}{\partial \theta_2}\right) &= E_o\left[\left(\frac{y_1-\Phi\left(\mathbf{w}\left(\theta_o\right)\right)}{\left[1-\Phi\left(\mathbf{w}\left(\theta_o\right)\right)\right]\Phi\left(\mathbf{w}\left(\theta_o\right)\right)} \phi\left(\mathbf{w}\left(\theta_o\right)\right)\right)^2 \left[-\eta_o \otimes \mathbf{z}'\right]\left[-\eta_o' \otimes \mathbf{z}\right]\right] \\
V_o\left(\frac{\partial q_2^o}{\partial vec\left(\boldsymbol{\delta}_2\right)}\right) &= E_o\left[\left[I_r \otimes \mathbf{z}'\right] \Sigma_{o22}^{-1} \mathbf{v}_2' \mathbf{v}_2 \Sigma_{o22}^{-1} \left[I_r \otimes \mathbf{z}\right]\right]
\end{aligned}$$

The orthogonality between scores holds if conditional means are correct since

$$\begin{aligned}
E_o\left[\frac{\partial q_1^o}{\partial \theta} \frac{\partial q_2^o}{\partial \theta_2'}\right] &= E_o\left[E_o\left[\frac{\partial q_1^o}{\partial \theta}\middle| \mathbf{y}_2, \mathbf{z}_1, \mathbf{v}_2\right] \frac{\partial q_2^o}{\partial \theta_2'}\right] \\
&= 0
\end{aligned}$$

Then, it is implied by Proposition 1.3.14 that followings are sufficient for QLIML to be efficient relative to CF for θ_1

$$\begin{aligned}
E_o\left[y_1 | \mathbf{y}_2, \mathbf{z}\right] &= \Phi\left(\mathbf{w}\left(\theta_o\right)\right) \\
V_o\left[y_1 | \mathbf{y}_2, \mathbf{z}\right] &= \tau_1 \Phi\left(\mathbf{w}\left(\theta_o\right)\right)\left[1-\Phi\left(\mathbf{w}\left(\theta_o\right)\right)\right] \\
E_o\left[\mathbf{y}_2 | \mathbf{z}\right] &= \mathbf{z}\boldsymbol{\delta}_{o2} \\
V_o\left[\mathbf{y}_2 | \mathbf{z}\right] &= \tau_2 \Sigma_{o22}
\end{aligned}$$

with

$$\tau_1 \leq \tau_2$$

□

The restriction $\tau_1 \leq \tau_2$ is especially plausible for application of Probit model to fractional response $y_1 \in [0, 1]$. Note that, with correctly specified conditional mean function, the conditional variance of y_1 is bounded above by $V_q\left[y_1|\mathbf{y}_2,\mathbf{z}\right]$:

$$\begin{aligned} V_o\left[y_1|\mathbf{y}_2,\mathbf{z}\right] &= E\left[y_1^2|\mathbf{y}_2,\mathbf{z}\right] - \left[\Phi\left(\mathbf{w}\left(\theta_o\right)\right)\right]^2 \\ &\leq E\left[y_1|\mathbf{y}_2,\mathbf{z}\right] - \left[\Phi\left(\mathbf{w}\left(\theta_o\right)\right)\right]^2 \\ &= \Phi\left(\mathbf{w}\left(\theta_o\right)\right)\left[1-\Phi\left(\mathbf{w}\left(\theta_o\right)\right)\right] \end{aligned}$$

And τ_1 often appears to be very small in practice when τ_2 is normalized to 1.

Another example where the relative efficiency conditions are applicable is Poisson regression model for positive response (such as count data) with endogenous explanatory variable.

Example 1.3.16 (exponential model) In the following simulataneous equation system

$$\begin{aligned} y_1 &= exp\left(\mathbf{y}_2\alpha + \mathbf{z}_1\delta_1 + \mathbf{v}_2\eta\right)u_1 \\ \mathbf{y}_2 &= \mathbf{z}\boldsymbol{\delta}_2 + \mathbf{v}_2 \end{aligned}$$

assume $y_1|\mathbf{z},\mathbf{y}_2 \sim_q Poisson\left(exp\left(\mathbf{y}_2\alpha + \mathbf{z}_1\delta_1 + \mathbf{v}_2\eta\right)\right)$ and $\mathbf{y}_2|\mathbf{z} \sim_q Normal\left(\mathbf{z}\boldsymbol{\delta}_2, \Sigma_{22}\right)$. Then, quasi-log-likelihood is

$$\begin{aligned} q_1\left(\theta_1,\theta_2\right) &= -\log\left(y_1!\right) - exp\left(\mathbf{y}_2\alpha + \mathbf{z}_1\delta_1 + \mathbf{v}_2\eta\right) + y_1\left(\mathbf{y}_2\alpha + \mathbf{z}_1\delta_1 + \mathbf{v}_2\eta\right) \\ q_2\left(\theta_2\right) &= -\frac{k}{2}\ln 2\pi - \frac{1}{2}\ln\left|\Sigma_{22}\right| - \frac{1}{2}\mathbf{v}_{i2}\left(\boldsymbol{\delta}_2\right)\Sigma_{22}^{-1}\mathbf{v}_{i2}\left(\boldsymbol{\delta}_2\right)' \end{aligned}$$

Since Poisson likelihood also belongs to LEF, by Proposition 1.3.14, a set of sufficient conditions for relative efficiency of QLIML for θ_1 is

$$\begin{aligned} E_o\left[y_1|\mathbf{y}_2,\mathbf{z}\right] &= exp\left(\mathbf{y}_2\alpha_o + \mathbf{z}_1\delta_{o1} + \mathbf{v}_2\eta_o\right) \\ V_o\left[y_1|\mathbf{y}_2,\mathbf{z}\right] &= \tau_1 exp\left(\mathbf{y}_2\alpha_o + \mathbf{z}_1\delta_{o1} + \mathbf{v}_2\eta_o\right) \\ E_o\left[\mathbf{y}_2|\mathbf{z}\right] &= \mathbf{z}\boldsymbol{\delta}_{o2} \\ V_o\left[\mathbf{y}_2|\mathbf{z}\right] &= \tau_2\Sigma_{o22} \end{aligned}$$

with

$$\tau_1 \leq \tau_2$$

□

1.4 Concluding Remarks

I show that, when both QLIML and CF estimators are consistent and asymptotic normal, there exists an efficient GMM estimator called GMM-QLIML whose asymptotic variance constitutes a lower bound of those of QLIML and CF estimators. In particular, a set of generalized information matrix equalities is shown to be sufficient for QLIML estimator to be as efficient as GMM-QLIML. In fully robust estimation of correctly specified conditional means, the condition is further weakened to GLM variance assumption with a scaling restriction. As Example 1.3.15 demonstrates, this condition is especially appealing for Probit model applied to fractional response.

Still, there are remaining questions to be answered. Regarding Proposition 1.3.13, can we derive a refined condition for relative efficiency of QLIML estimator in $\tau_1 > \tau_2$ case? As discussed for Poisson regression model in Example 1.3.16, there are models that often exibits large τ_1, and this refinement (if possible) will be useful. Also, we cannot rule out the possibility of even weaker condition than GIMEs with scaling restriction given in Proposition 1.3.13. Direct comparison of asymptotic variances does not seem to work well in that direction of research.

Moreover, relative efficiency relationship with other QLIL-based estimators can be studied. For example, when reduced form model for y_1 is available, the minimum distance estimator suggested by Amemiya (1978, 1979) can be used. Newey (1987) showed its asymptotic efficiency in limited information structure with normal errors. It would be interesting to study its relative efficiency relationship when it is based on QLIL.

CHAPTER 2

EFFICIENT MINIMUM DISTANCE ESTIMATOR BASED ON QUASI-LIMITED INFORMATION LIKELIHOOD

2.1 Introduction

When a model is over-identified in limited information simulataneous system, the classical minimum distance estimator is often proposed as an estimation method. Amemiya (1978,1979) first introduced its application to Probit and Tobit model with endogenous explanatory variables and gave an interpretation of 'generalized least square'. Newey (1987) called this estimator as 'Amemiya's GLS (AGLS)' and showed its asymptotic efficiency under correct specification of likelihood in a general class of limited information structures. Recent work by Wooldridge (2014) implies that, in linear exponential family (LEF), correct specification of regression functions of reduced form model guarantees robustness of AGLS. Still, its relative efficiency relationship has not been clarified for the case of potentially misspecified likelihood.

The purpose of this chapter is to study asymptotic behavior of minimum distance estimator based on quasi-limited information likelihood. The primary focus is on its relative efficiency relationship with respect to quasi-limited information maximum likelihood (QLIML) estimator and two-step control function (CF) approach. this chapter takes the quasi-limited information framework from Wooldridge (2014) and relies on results from Chapter 1.

The main contributions of this chapter are followings. First, AGLS is interpreted as a concentrated estimator (cMD-QLIML) and 'full' minimum distance estimator (MD-QLIML) is proposed. Based on an analogous result of Crepon, Kramarz, Trognon (1997), cMD-QLIML is shown to be asymptotically equivalent to MD-QLIML for structural parameters. Second, given quasi-limited information likelihood, cMD-QLIML is proved to be asymptotically efficient relative to QLIML and CF. In particular, cMD-QLIML can be strictly more efficient than QLIML in Newey's framework if enough degree of misspecification is present in likelihood. Third, if and only if condition for cMD-QLIML and other estimators to be asymptotically equivalent under the null hypothesis of exogeneity is derived. Immediate implication shows that a set of generalized information matrix equalities for reduced form model is sufficient. Fourth, an explicit formula of cMD-QLIML estimator for linear model is derived. It is the same as GMM but with a different weighting matrix derived from the reduced form parameter estimates.

The rest of this chapter is organized as follows. In Section 2.2, basic model restrictions are given. In Section 2.3, MD-QLIML and cMD-QLIML are defined, and relative efficiency relationship is presented. Section 2.4 contains application to linear model and quantile regression model with endogeneous explanatory variables.

2.2 Model Restrictions

Assume random sampling from a population. Model restrictions start from a decomposed quasi-limited information log-likelihood framework in Wooldridge (2014)

$$QLL = q_1\left(y_{i1}, \mathbf{y}_{i2}, \mathbf{z}_i, \theta_1, \theta_2\right) + q_2\left(\mathbf{y}_{i2}, \mathbf{z}_i, \theta_2\right) \tag{2.1}$$

where $\theta = \left(\theta_1', \theta_2'\right)'$ is $(p_1 + p_2)$-dimensional vector of parameter, y_{i1} is the i th observation of a scalar response variable, $\mathbf{y}_{i2}$ is a $1 \times r$ vector of potentially endogenous variables, $\mathbf{z}_i = (\mathbf{z}_{i1}, \mathbf{z}_{i2})$ is $1 \times k$ vector of included/excluded exogenous instruments with $k = k_1 + k_2$. For details, see Wooldridge (2014).

QLIML and CF estimators are initially given as

$$\hat{\theta}_{QLIML} = \arg\max_{\theta} \sum_{i=1}^{N} \left[q_{i1}(\theta) + q_{i2}(\theta_2)\right] \quad \text{and} \quad \begin{cases} \hat{\theta}_{2,CF} = \arg\max\limits_{\theta_2} \sum_{i=1}^{N} q_{i2}(\theta_2) \\ \hat{\theta}_{1,CF} = \arg\max\limits_{\theta_1} \sum_{i=1}^{N} q_{i1}\left(\theta_1, \hat{\theta}_{2,CF}\right) \end{cases}$$

and redefined as GMM estimators based upon first order conditions

$$\sum_{i=1}^{N} \begin{bmatrix} \frac{\partial}{\partial\theta_1} q_{i1}\left(\hat{\theta}_{QLIML}\right) \\ \frac{\partial}{\partial\theta_2} q_{i1}\left(\hat{\theta}_{QLIML}\right) + \frac{\partial}{\partial\theta_2} q_{i2}\left(\hat{\theta}_{2,QLIML}\right) \end{bmatrix} = 0 \quad \text{and} \quad \sum_{i=1}^{N} \begin{bmatrix} \frac{\partial}{\partial\theta_1} q_{i1}\left(\hat{\theta}_{CF}\right) \\ \frac{\partial}{\partial\theta_2} q_{i2}\left(\hat{\theta}_{2,CF}\right) \end{bmatrix} = 0$$

respectively. Finite sample estimates of above extremum estimator and GMM-interpreted estimator may not coincide. Such numerical discrepancy does not harm our asymptotic analysis since they are asymptotically equivalent under regularity conditions.

Both QLIML and CF estimators are assumed to be $\sqrt{N}-$consistent for the true parameter values and asymptotic normal. The essential model restriction for validity of the relative efficiency results in this chapter is that the asymptotic variance of each estimator is in the standard sandwich form. To explicitly account for some cases of non-smooth objective functions, the Jacobian matrix of expected score is used rather than the expected Jacobian matrix of score. Note that redundancy conditions in Breusch, Qian, Schmidt and Wyhowski (1999) are compatible with such modification. Generalized information matrix equalities (GIMEs) are also defined accordingly. A set of standard regularity conditions for GMM interpreted estimators (Assumptions 1–13) are given in Appendix B.1. It is easy to show that, under these conditions, Proposition 1.3.4 in Chapter 1 holds with Jacobian matrices of expected score in the sandwich form.

To consider a distance minimization problem, a reduced form model should be defined. The existence of link function $\gamma : \Theta \to \Gamma \subset \mathbb{R}^g$ is essential in the following characterization of a reduced form model.

Definition 2.2.1 A reduced form model is a pair $\left(q_{i1}^R(\gamma, \theta_2), \gamma(\theta)\right)$ such that

$$q_{i1}^R(\gamma(\theta), \theta_2) = q_{i1}(\theta_1, \theta_2) \text{ a.s. for } \forall \theta \in \Theta \tag{2.2}$$

$$\frac{\partial q_{i1}^R}{\partial \theta_2}(\gamma, \theta_2) = C(\theta) \frac{\partial q_{i1}^R}{\partial \gamma}(\gamma, \theta_2) \text{ a.s.} \tag{2.3}$$

for some $p_2 \times g$ matrix $C(\theta)$ whose elements are real-valued function of θ. □

The link function $\gamma(\theta)$ represents the functional relationship between structural parameters and reduced form parameters. It relates likelihoods of structural and reduced form model as in the first condition (2.2). Note that the decomposed likelihood

$$q_{i1}^R(\gamma, \theta_2) + q_2(\theta_2)$$

for a reduced form model still belongs to QLIML framework, and q_1 alone cannot identify γ without help of q_2. Based on relative efficiency results in Chapter 1, this model is 'reduced' in the sense that GMM-QLIML for this model has no additional moment functions from $\frac{\partial q_{i1}^R}{\partial \theta_2}$. In other words, all nonredundant effects of $(\mathbf{z}_i, \mathbf{y}_{i2})$ on y_1 are captured in γ. This is chracterized by the second condition (2.3). In turn, QLIML and CF estimator are asymptotically equivalent for reduced form model parameters (γ, θ_2) . A set of standard regularity conditions for a reduced form model and a link function (Assumptions 14–17) are given in Appendix B.1.

2.3 Minimum Distance Estimators: MD/cMD-QLIML

Given reduced form estimates $(\hat{\gamma}, \hat{\theta}_2)$, MD-QLIML is defined as minimum distance estimator of θ minimizing optimally weighted sum of distance $\hat{\gamma} - \gamma(\theta)$ and $\hat{\theta}_2 - \theta_2$.

Definition 2.3.1 Let $(\hat{\gamma}, \hat{\theta}_2)$ be reduced form parameter estimates and suppose

$$\sqrt{N}\left(\left(\begin{array}{c}\widehat{\gamma}\\ \hat{\theta}_2\end{array}\right)-\left(\begin{array}{c}\gamma_o\\ \theta_{o2}\end{array}\right)\right)\xrightarrow{d} N\left(0,\Omega_R\right)$$

Then MD-QLIML estimator $\hat{\theta}_{MD-QLIML}$ is a solution to

$$\min_{\theta}\left[\left(\begin{array}{c}\widehat{\gamma}\\ \hat{\theta}_2\end{array}\right)-h\left(\theta\right)\right]'\hat{\Omega}_R^{-1}\left[\left(\begin{array}{c}\widehat{\gamma}\\ \hat{\theta}_2\end{array}\right)-h\left(\theta\right)\right] \tag{2.4}$$

where $h(\theta) = (\gamma(\theta)', \theta_2')'$ and $\hat{\Omega}_R \xrightarrow{p} \Omega_R$. □

Hence, MD-QLIML is a two-step procedure:

1. estimate $(\hat{\gamma}, \hat{\theta}_2)$ by solving just-identified reduced form model. $(\hat{\gamma}, \hat{\theta}_2)$ mainly represents estimated mean responsiveness of y_{i1} and $\mathbf{y}_{i2}$ with respect to all available exogenous variation of instuments.

2. compress information contained in $(\hat{\gamma}, \hat{\theta}_2)$ into structural parameter estimates by solving the distance minimization problem (2.4).

Later, we will see this two-step estimation procedure can enhance finite sample performance remarkably compared to asymptotically equivalent optimal GMM estimator when model is over-identified. The next proposition states the asymptotic distribution of MD-QLIML.

Proposition 2.3.2 Assume Assumption 1–17. MD-QLIML is asymptotically normal

$$\sqrt{N}\left(\hat{\theta}_{MD-QLIML}-\theta_o\right)\overset{d}{\to}N\left(0,\left(H_o'\Omega_R^{-1}H_o\right)^{-1}\right)$$

where $H_o=\frac{\partial}{\partial\theta'}h\left(\theta_o\right)$ □

One significant distinction of MD-QLIML from Amemiya's GLS estimator is that the distance minimization problem of MD-QLIML considers almost redundant looking distance $\hat{\theta}_2-\theta_2$ while that of AGLS imposes a constraint $\theta_2=\hat{\theta}_2$ with corresponding adjustment of the weighting matrix. In fact, AGLS can be interpreted as a 'concentrated MD-QLIML' where 'concentration' means θ_2 is regarded as a implicit function of θ_1 in the solution space of minimization problem. This interpretation is based on the following general result which is analogous to Proposition 1 in Crepon, Kramarz and Trognon (1997).

Proposition 2.3.3 (Concentrated Minimum Distance Estimator) Assume (1) $h\left(\theta\right)=\left(h_1',h_2'\right)'$ is continuously differentiable in $\theta=\left(\theta_1,\theta_2\right)$ where $\theta_1\in\mathbb{R}^{p_1}$, $\theta_2\in\mathbb{R}^{p_2}$, $h_1\in\mathbb{R}^g$, $h_2\in\mathbb{R}^{p_2}$ with $g\geq p_1$ (2) $\frac{\partial h(\theta_o)}{\partial\theta}$ has full column rank (3) $\gamma_o-h\left(\theta\right)\neq0$ if $\theta\neq\theta_o$ where $\gamma_o=\left(\gamma_{o1},\gamma_{o2}\right)$, $\gamma_{o1}\in\mathbb{R}^g$ and $\gamma_{o2}\in\mathbb{R}^{p_2}$ (4) $\sqrt{N}\left(\hat{\gamma}-\gamma_o\right)\overset{d}{\to}N\left(0,\Omega_o\right)$ (5) $\det\left(\frac{\partial h_2(\theta)}{\partial\theta_2}\right)\neq0$ for $\forall\theta\in\Theta$. Then, $\varphi_N\left(\theta_1\right)$ is well-defined by $\hat{\gamma}_2-h_2\left(\theta_1,\varphi_N\left(\theta_1\right)\right)=0$ for each $\left(\theta_1,N\right)$, and an estimator $\hat{\theta}_{c,1}$ derived from

$$\min_{\theta_1}\left(\hat{\gamma}_1-h_1\left(\theta_1,\varphi_N\left(\theta_1\right)\right)\right)'\hat{W}_1\left(\hat{\gamma}_1-h_1\left(\theta_1,\varphi_N\left(\theta_1\right)\right)\right)$$

is asymptotically equivalent to a minimum distance estimator of θ_1 which solves

$$\min_{\theta}\left(\hat{\gamma}-h\left(\theta\right)\right)'\hat{W}\left(\hat{\gamma}-h\left(\theta\right)\right)$$

where $\hat{W}_1 \xrightarrow{p} \left(S_o \Omega_o S_o'\right)^{-1}$, $S_o = \left[\begin{array}{cc} I_{g_1} & -\frac{\partial h_1(\theta_o)}{\partial \theta_2'}\left[\frac{\partial h_2(\theta_o)}{\partial \theta_2'}\right]^{-1}\end{array}\right]$ and $\hat{W} \xrightarrow{p} \Omega_o^{-1}$. Moreover, the asymptotic distribution of $\hat{\theta}_{c,1}$ is

$$\sqrt{N}\left(\hat{\theta}_{c,1} - \theta_{o1}\right) \xrightarrow{d} N\left(0, \left[H_c'\left(S_o \Omega_o S_o'\right)^{-1} H_c\right]^{-1}\right)$$

where $H_c = \frac{\partial}{\partial \theta_1'} h_1(\theta_o) - \frac{\partial}{\partial \theta_2'} h_1(\theta_o)\left[\frac{\partial h_2(\theta_o)}{\partial \theta_2'}\right]^{-1} \frac{\partial h_2(\theta_o)}{\partial \theta_1'}$. □

Proposition 2.3.3 provides a method to construct an asymptotically equivalent minimum distance estimator for θ_1 by concentrating θ_2 out. The key condition is that the dimension of concentrated parameter θ_2 is the same as that of concentrating equation $\hat{\gamma}_2 - h_2(\theta_1, \theta_2) = 0$. This condition is presicely satisfied for θ_1 and θ_2 in QLIML framework. Applying Proposition 2.3.3 to QLIML framework, we have $h_2(\theta_1, \theta_2) = \theta_2$ and the implict function in the proposition is merely $\varphi_N(\theta_1) = \hat{\theta}_2$, a constant function of θ_1. In turn, AGLS can be defined as concentrated MD-QLIML (cMD-QLIML) and its asymptotic distribution and asymptotic equivalence to MD-QLIML follows as a corollary.

Definition 2.3.4 cMD-QLIML (=AGLS) is defined to be a solution $\hat{\theta}_{1,cMD-QLIML}$ to

$$\min_{\theta_1}\left[\widehat{\gamma} - \gamma\left(\theta_1, \hat{\theta}_2\right)\right]' \hat{W}_1 \left[\widehat{\gamma} - \gamma\left(\theta_1, \hat{\theta}_2\right)\right] \tag{2.5}$$

where $\hat{W}_1 \xrightarrow{p} \left(S_o \Omega_R S_o'\right)^{-1}$ and $S_o = \left[\begin{array}{cc} I_g & -\frac{\partial \gamma(\theta_o)}{\partial \theta_2'}\end{array}\right]$ □

Corollary 2.3.5 *Assume Assumption 1–17. Then,*

$$V^1_{MD-QLIML} = V^1_{cMD-QLIML}$$

and

$$\sqrt{N}\left(\theta_{1,cMD-QLIML} - \theta_{o1}\right) \xrightarrow{d} N\left(0, \left[H_c'\left(S_o \Omega_R S_o'\right)^{-1} H_c\right]^{-1}\right)$$

where $H_c = \frac{\partial \gamma(\theta_o)}{\partial \theta_1'}$ *and* $S_o = \left[\begin{array}{cc} I_g & -\frac{\partial \gamma(\theta_o)}{\partial \theta_2'}\end{array}\right]$ □

To study relative efficiency relationship between GMM-QLIML and AGLS(cMD-QLIML), it is useful to consider following GMM counterpart for MD-QLIML.

Definition 2.3.6 (MD-QLIML equivalent GMM) mGMM-QLIML is defined to be an optimal GMM estimator based on

$$\begin{bmatrix} \frac{\partial}{\partial \gamma} q_1^R\left(\gamma\left(\theta\right), \theta_2\right) \\ \frac{\partial}{\partial \theta_2} q_2\left(\theta_2\right) \end{bmatrix} \tag{2.6}$$

□

The moments in (2.6) are the same first order conditions used in the first stage of MD-QLIML except that γ is being treated as a function of θ. This estimator can be understood as combining two-step procedure of MD-QLIML into one-step: accounting for mean responsiveness of y_1 and $\mathbf{y}_2$ with respect to all exogeneous variation of instruments, choose θ optimally. Under regularity and identification conditions for MD-QLIML (Assumption 1–17), this estimator is well-defined and asymptotically equivalent to MD-QLIML.

Proposition 2.3.7 Assume Assumption 1–17. Then MD-QLIML and mGMM-QLIML are asymptotically equivalent.

mGMM-QLIML is efficient relative to GMM-QLIML in matrix positive semi-definite sense. Proposition 2.3.8 formalizes the results along with a condition under which mGMM-QLIML and GMM-QLIML are asymptotically equivalent.

Proposition 2.3.8 Assume Assumption 1–17. Then

$$V_{mGMM-QLIML} \preceq V_{GMM-QLIML}$$

where the inequality becomes equality if $p_1 + p_{22} = g$. □

The following proposition summarizes the results in the framework of linear index model which is most frequently used but not the only class of models that results can be applied to.

Proposition 2.3.9 Assume Assumption 1–18. Suppose

$$q_{i1}\left(\theta_1, \theta_2\right) = l\left(y_{i1}, \mathbf{y}_2\boldsymbol{\alpha} + \mathbf{z}_1\delta_1 + \mathbf{v}_2\eta, \lambda\right)$$
$$q_{i2}\left(\theta_2\right) = -\frac{k}{2}\ln 2\pi - \frac{1}{2}\ln\left|\Sigma_{22}\right| - \frac{1}{2}\mathbf{v}_{i2}\left(\boldsymbol{\delta}_2\right)\Sigma_{22}^{-1}\mathbf{v}_{i2}\left(\boldsymbol{\delta}_2\right)'$$

where $\theta_1 = \left(\boldsymbol{\alpha}', \delta_1', \eta', \lambda'\right)'$, $\theta_2 = \left(vec\left(\boldsymbol{\delta}_2\right)', vec\left(\Sigma_{22}\right)'\right)'$ and $\mathbf{v}_{i2}\left(\boldsymbol{\delta}_2\right) \equiv \mathbf{y}_{i2} - \mathbf{z}_i\boldsymbol{\delta}_2$. Then following results hold:

(a) $V_{MD-QLIML} \preceq V_{GMM-QLIML} \preceq V_{QLIML}, V_{CF}$

(b) If $k_2 = r$, then $V_{MD-QLIML} = V_{GMM-QLIML} = V_{QLIML} = V_{CF}$

(c) If $\eta_o \neq 0$, then $V_{MD-QLIML} = V_{GMM-QLIML} \preceq V_{QLIML}, V_{CF}$

(d) If $\eta_o = 0$, then $V_{MD-QLIML} \preceq V_{GMM-QLIML} = V_{QLIML} = V_{CF}$

(e) If $\eta_o = 0$ and $k_2 > r$, then $V_{MD-QLIML} = V_{GMM-QLIML}$ if and only if

$$\begin{aligned}
&\frac{\partial}{\partial\theta'}E\left[\frac{\partial q_{i1}}{\partial\gamma^*}\right]\Bigg|_{\left(\gamma',\theta_2'\right)=\left(\gamma_o',\theta_{o2}'\right)} \\
&= cov\left(\frac{\partial q_{i1}^o}{\partial\gamma^*}, \begin{array}{c}\frac{\partial q_{i1}^o}{\partial\theta_1} \\ \frac{\partial q_{i1}^o}{\partial\theta_2} + \frac{\partial q_{i2}^o}{\partial\theta_2}\end{array}\right) V\left(\begin{array}{c}\frac{\partial q_{i1}^o}{\partial\theta_1} \\ \frac{\partial q_{i1}^o}{\partial\theta_2} + \frac{\partial q_{i2}^o}{\partial\theta_2}\end{array}\right)^{-1} \frac{\partial}{\partial\theta'}E\left[\begin{array}{c}\frac{\partial q_{i1}}{\partial\theta_1} \\ \frac{\partial q_{i1}}{\partial\theta_2} + \frac{\partial q_{i2}}{\partial\theta_2}\end{array}\right]\Bigg|_{\theta=\theta_o}
\end{aligned}$$

where γ^* is such that optimal GMM on $\frac{\partial q_{i1}^o}{\partial\gamma^*}$ together with QLIML moment functions is asymptotically equivalent to mGMM-QLIML. A sufficient condition is GIME's for q_1, q_2, $q_1 + q_2$ with same scaling factor for reduced form model. □

(a) holds also for the cases where the index contains higher order terms of $\mathbf{y}_2$. However, asymptotic equivalence of MD-QLIML and GMM-QLIML in (b) and (c) doesn't hold in general when higher order terms of $\mathbf{y}_2$ are present. (b) is well-known property and, in fact, estimators are numerically equivalent. (c) is a typical relative efficiency relationship when endogeneity is present. If a set of GIME's with common scaling factor holds, QLIML will also be asymptotically equivalent to MD-QLIML. (d) and (e) shows relative efficiency of MD-QLIML when there exists no endogeneity. With over-identification, there exists potential efficiency gain which vanishes under a set of GIME's for reduced form model.

2.4 Example 1: Linear Regression Model with Endogeneous Explanatory Variables

From equation $\mathbf{y}_{i2} = \mathbf{z}_i \boldsymbol{\delta}_2 + \mathbf{v}_{i2}$ and uniqueness of linear projection, it is clear that $L\left(y_{i1}|\mathbf{y}_{i2}, \mathbf{z}_i\right) = L\left(y_{i1}|\mathbf{v}_{i2}, \mathbf{z}_i\right)$. Substituting $\mathbf{y}_2$ into q_1, or equivalently, substituting into regression equation y_{i1} yields

$$\begin{aligned} y_{i1} &= \mathbf{y}_{i2}\boldsymbol{\alpha} + \mathbf{z}_{i1}\delta_1 + \mathbf{v}_{i2}\Sigma_{22}^{-1}\Sigma_{21} + e_{i1} \\ &= \mathbf{z}_{1i}\left(\boldsymbol{\delta}_{21}\boldsymbol{\alpha} + \delta_1\right) + \mathbf{z}_{2i}\boldsymbol{\delta}_{22}\boldsymbol{\alpha} + \mathbf{v}_{i2}\left(\boldsymbol{\alpha} + \Sigma_{22}^{-1}\Sigma_{21}\right) + e_{i1} \\ &\equiv \mathbf{z}_{1i}\gamma_1 + \mathbf{z}_{2i}\gamma_2 + \mathbf{v}_{i2}\gamma_3 + e_{i1} \end{aligned}$$

Along with $\gamma_4 \equiv \sigma_{11|2}(\theta)$, $\gamma(\theta)$ is naturally defined as

$$\gamma(\theta) = \left(\left(\boldsymbol{\delta}_{21}\boldsymbol{\alpha} + \delta_1\right)', \left(\boldsymbol{\delta}_{22}\boldsymbol{\alpha}\right)', \left(\boldsymbol{\alpha} + \Sigma_{22}^{-1}\Sigma_{21}\right)', \sigma_{11|2}(\theta)\right)'$$

while dependence of q_{i1} on θ_2 is through the control function $\mathbf{v}_{i2}(\boldsymbol{\delta}_2)$. Consistency and asymptotic normality of estimator of (γ, θ_2) is implied by invertibility of $E\left[\mathbf{z}'\mathbf{z}\right]$ and other regularity conditions assumed for QLIML and CF estimators. However, additional assumptions are needed in nonlinear models in general. For example, when we allow higher order terms of $\mathbf{y}_2$ in regression function as in

$$y_{i1} = y_{i2}^2\alpha + \mathbf{z}_{i1}\delta_1 + \mathbf{v}_{i2}\Sigma_{22}^{-1}\Sigma_{21} + e_{i1}$$

it is required to impose additional orthogonality and rank condition for structural and reduced form model. i.e. (a) regression function specification is done with conditional mean operator rather than linear projection operator. (b) linear independence of higher order terms of instruments is assumed. Following demonstrates a computationally useful reparameterization of classical LIML under which explicit expressions for H_o, H_c, S_o and the closed form of cMD-QLIML estimator are given.

Consider reparametrization

$$\begin{aligned} \eta &\equiv \Sigma_{22}^{-1}\Sigma_{21} \\ \sigma_{11|2} &\equiv \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \end{aligned}$$

It can be easily shown that this reparameterization does not alter other parameter estimates of any estimation method discussed in this chapter. Taking derivative of $h(\theta)$ at θ_o, we have

$$H_o = \begin{bmatrix} H_{11}(\theta_{o2}) & H_{12}(\theta_{o1}) \\ 0 & I_{p_2} \end{bmatrix}$$

where

$$H_{11}(\theta_2) = \begin{bmatrix} \delta_{21_{k_1 \times r}} & I_{k_1} & 0 & 0 \\ \delta_{22_{k_2 \times r}} & 0 & 0 & 0 \\ I_r & 0 & I_r & 0 \\ 0 & 0 & 0 & 1_{1\times 1} \end{bmatrix} \text{ and } H_{12}(\theta_1) = \begin{bmatrix} \alpha' \otimes I_k & \\ 0_{r\times rk} & 0_{g\times\frac{r(r+1)}{2}} \\ 0_{1\times rk} & \end{bmatrix}$$

Hence, we have $H_c = H_{11}(\theta_{o2})$ and $S_o = \begin{bmatrix} I_g & -H_{12}(\theta_{o1}) \end{bmatrix}$. Preliminary estimates for α can be calculated by QLIML or CF when constructing weighting maxtrix for cMD-QLIML. Moreover, since $\gamma(\theta) = H_{11}(\theta_2)\theta_1$, taking first order condition of (2.5) yields

$$\theta_{1,cMD-QLIML} = \left[H_{11}\left(\hat{\theta}_2\right)' \hat{W}_1 H_{11}\left(\hat{\theta}_2\right)\right]^{-1}\left[H_{11}\left(\hat{\theta}_2\right)' \hat{W}_1 \widehat{\gamma}\right]$$

Note that $e_i(\theta) = e_i(\gamma(\theta))$ where

$$e_i(\theta) \equiv y_{i1} - \mathbf{y}_{i2}\boldsymbol{\alpha} - \mathbf{z}_{i1}\delta_1 - \mathbf{v}_{i2}\Sigma_{22}^{-1}\Sigma_{21}$$

$$e_i(\gamma(\theta)) \equiv y_{i1} - \mathbf{z}_{1i}\gamma_1(\theta) - \mathbf{z}_{2i}\gamma_2(\theta) - \mathbf{v}_{i2}\gamma_3(\theta)$$

By replacing $e_i(\theta)$ with $e_i(\gamma)$ in q_{i1} along with reparameterizing $\sigma_{11|2}(\theta)$, a reduced form model likelihood $q_1(\gamma, \theta_2)$ is derived:

$$q_{i1}(\gamma, \theta_2) = -\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln\gamma_4 - \frac{1}{2}e_i(\gamma)^2\gamma_4^{-1}$$

Taking derivatives with respect to γ, we have

$$\frac{\partial q_{i1}(\gamma, \theta_2)}{\partial\gamma} = \begin{bmatrix} \gamma_4^{-1} e_i(\gamma)\mathbf{z}' \\ \gamma_4^{-1} e_i(\gamma)\mathbf{v}_2' \\ -\frac{1}{2}\gamma_4^{-1} + \frac{1}{2}e_i(\gamma)^2\gamma_4^{-2} \end{bmatrix}$$

Then, mGMM-QLIML moment functions are derived by noting $\gamma(\theta)$ as a function of θ

$$\frac{\partial q_{i1}(\gamma(\theta),\theta_2)}{\partial \gamma} = \left[\begin{array}{c} \sigma_{11|2}(\theta)^{-1} e_i(\theta)\mathbf{z}' \\ \sigma_{11|2}(\theta)^{-1} e_i(\theta)\mathbf{v}_2' \\ -\frac{1}{2}\sigma_{11|2}(\theta)^{-1} + \frac{1}{2}e_i(\theta)^2\sigma_{11|2}(\theta)^{-1} \end{array}\right]$$

It is not difficult to see that mGMM-QLIML is asymtotically equivalent to GMM-QLIML whose $\frac{\partial q_1}{\partial \theta_{22}}$ replaced with (1.11) as discussed in previous section.[1] To see existence of $C(\theta)$, assuming, without loss of generality, that first $k_2 - r$ instruments in $\mathbf{z}_2$ are chosen in $\frac{\partial q_1}{\partial \theta_{22}}$, it is implied that

$$C(\theta) = \left[\begin{array}{cccc} 0_{(k_2-r)\times k_1} & -\left(\Sigma_{22}^{-1}\Sigma_{21}\right)_{i_o} I_{k_2-r} & 0_{(k_2-r)\times(2r+1)} \end{array}\right]$$

when $\left(\Sigma_{22}^{-1}\Sigma_{21}\right)_{i_o}$ is nonzero for chosen i_o and $k_2 > r$.

2.5 Example 2: Probit with Endogeneous Explanatory Variables

Consider probit model with endogeneity.

$$\begin{aligned} y_{i1} &= 1\left[\mathbf{y}_{i2}\boldsymbol{\alpha} + \mathbf{z}_{i1}\delta_1 + u_{i1} > 0\right], \\ \mathbf{y}_{i2} &= \mathbf{z}_i\boldsymbol{\delta}_2 + \mathbf{v}_{i2} \end{aligned}$$

Assume regularity and identification conditions (Assumption 1–17). For computational convienience, similar reparametrization in Example 1 is done along with normalization of $e_1 = u_1 - \mathbf{v}_2\eta$. Also, Σ_{22} is taken out from q_2 since its exclusion does not affect other parameter estimates. Then, quasi-likelihood can be simplified as following

$$\begin{aligned} q_1(\theta_1,\theta_2) &= (1-y_1)\log\left[1-\Phi(\mathbf{w}(\theta))\right] + y_1\log\Phi(\mathbf{w}(\theta)) \\ q_{i2}(\theta_2) &= -\frac{1}{2}\mathbf{v}_{i2}(\boldsymbol{\delta}_2)\mathbf{v}_{i2}(\boldsymbol{\delta}_2)' \end{aligned}$$

[1] Moreover, by invertible transformation of mGMM-moment functions and by separability condition of GMM, it can be shown that mGMM-QLIML is asymptotically equivalent to optimal GMM on $E\left[z'u\right]$.

where $\theta_1' = \left(\alpha', \delta_1', \eta'\right)'$, $\theta_2 = vec\left(\boldsymbol{\delta}_2\right)$, and $\mathbf{w}(\theta) = \mathbf{y}_2\alpha + \mathbf{z}_1\delta_1 + \mathbf{v}_2\eta$. Taking derivatives, quasi-scores can be expressed as

$$\frac{\partial q_1}{\partial \theta_1} = \frac{y_1 - \Phi\left(\mathbf{w}(\theta)\right)}{1 - \Phi\left(\mathbf{w}(\theta)\right)\Phi\left(\mathbf{w}(\theta)\right)}\phi\left(\mathbf{w}(\theta)\right)\begin{bmatrix} \mathbf{y}_2' \\ \mathbf{z}_1' \\ \mathbf{v}_2' \end{bmatrix}$$

$$\frac{\partial q_1}{\partial \theta_2} = -\frac{y_1 - \Phi\left(\mathbf{w}(\theta)\right)}{1 - \Phi\left(\mathbf{w}(\theta)\right)\Phi\left(\mathbf{w}(\theta)\right)}\phi\left(\mathbf{w}(\theta)\right)\left[\eta \otimes \mathbf{z}'\right]$$

$$\frac{\partial q_2}{\partial \theta_2} = \left[\left[I_r \otimes \mathbf{z}'\right]\left(\mathbf{y}_2 - z\delta_2\right)'\right]$$

It is easy to see that, GMM-QLIML extra moment functions are

$$\frac{\partial q_1}{\partial \theta_{22}} = -\frac{y_1 - \Phi\left(\mathbf{w}(\theta)\right)}{1 - \Phi\left(\mathbf{w}(\theta)\right)\Phi\left(\mathbf{w}(\theta)\right)}\phi\left(\mathbf{w}(\theta)\right)\left[\eta_{i_o}\mathbf{z}_{2,-r}'\right]$$

where $\eta_{i_o} \neq 0$.

Components of H_o, H_c and S_o can be calculated similarily as in Example 1.

$$H_{11}(\theta_2) = \begin{bmatrix} \delta_{21_{k_1 \times r}} & I_{k_1} & 0 \\ \delta_{22_{k_2 \times r}} & 0 & 0 \\ I_r & 0 & I_r \end{bmatrix} \text{ and } H_{12}(\theta_1) = \begin{bmatrix} \alpha' \otimes I_k \\ 0_{r \times rk} \end{bmatrix}$$

To derive mGMM moment functions, note

$$\begin{aligned} \mathbf{w}(\theta) &= \mathbf{y}_2\alpha + \mathbf{z}_1\delta_1 + \mathbf{v}_2\eta \\ &= \mathbf{z}_1\gamma_1 + \mathbf{z}_2\gamma_2 + \mathbf{v}_2\gamma_3 \\ &= \mathbf{w}\left(\gamma(\theta)\right) \end{aligned}$$

Then, differentiating with respect to reduced form parameters, we have

$$\frac{\partial q_1\left(\gamma(\theta), \theta_2\right)}{\partial \gamma} = \frac{y_1 - \Phi\left(\mathbf{w}(\theta)\right)}{1 - \Phi\left(\mathbf{w}(\theta)\right)\Phi\left(\mathbf{w}(\theta)\right)}\phi\left(\mathbf{w}(\theta)\right)\begin{bmatrix} \mathbf{z}' \\ \mathbf{v}_2' \end{bmatrix}$$

and it shows that this model is LL-class as claimed by Proposition 2.3.9.

2.6 Monte Carlo Simulation on Probit Model with EEV

In this section, Monte Carlo simulations are conducted on probit model under several specifications. The purpose is to investigate effects of GIME's on finite sample performance of estimators when model is over-identified. Based on the relative asymptotic efficiency results in previous sections, it is expected that MD and equivalent estimators outperform CF and QLIML (in terms of standard deviation) at large enough sample size when enough misspecification is present in an overidentified model. Root mean squared error (RMSE) is used as main performance measure in this study to take account of bias as well as mean deviation.

The assumptions on data generation is as follows: all data points are independently and identically generated. The instruments $\{z_k\}_{k=1}^{10}$ are mutually independent and $z_k \sim SBin\left(10^4, \frac{1}{3}\right)$ for each $1 \leq k \leq 10$ where $SBin\,(n, p) \overset{d}{=} \frac{Bin(n,p)-np}{\sqrt{np(1-p)}}$. The regression equation for scalar y_2 is $y_2 = z_1 + \cdots + z_{10} + v_2$. For specification of GIME's of q_1 and q_2, following restrictions were imposed in each case

Probit	GIME holds	GIME doesn't hold
$y_1 =$	$1_{[z_1+y_2+\eta v_2+e_2>0]}$	$\frac{1}{4}1_{[z_1+y_2+\eta v_2+e_2>0]} + \frac{3}{4}1_{\left[z_1+y_2+\eta v_2+(e_2+e_3)/\sqrt{2}>0\right]}$
$v_2 \sim$	$SBin\left(10^4, 1/3\right)$	$z_1 z_2 \cdot SBin\left(10^4, 1/3\right)$

where e_2 and e_3 each follow independent standard normal distribution. Due to nomalized variance of v_2, it can be shown that, by considering q_2 without Σ_{22} term, only homoskedasticity is needed for GIME to hold for q_2. Fractional response y_1 fails GIME for q_1 due to correlation between e_2 and $(e_2 + e_3)/\sqrt{2}$. Since relevent random variables are all discrete and have bounded supports, RMSE for all estimators are well-defined and can be used in comparison. Number of repetition is 10^4 and mGMM was estimated by iterative (or continuously updating) method. The simulation program was written in ado/MATA language in STATA 13, and it was executed using HPCC(High Performance Computing Center) resources provided by the iCER(institute for Cyber-Enabled Research) at Michigan State University.

Table 2.1 shows results. Simulation I is a case where a complete set of GIME's hold so that MD,

Table 2.1 Root Mean Squared Error and Standard Deviation ($\eta = 0.6$)

	I	II	III	IV	I	II	III	IV
y_1	bin	bin	frac	frac	bin	bin	frac	frac
v_2	hom	heter	hom	heter	hom	heter	hom	heter
	RMSE($\hat{\alpha}$)				SD($\hat{\alpha}$)			
CF	.075365	.077754	.104740	.104207	.074631	.076346	.083360	.082906
QLIML	.076288	.078776	.065795	.067283	.075181	.076850	.064932	.066215
mGMM	.106961	.110279	.088611	.089067	.088201	.090190	.074308	.075319
cMD	.074520	.077083	.065725	.066771	.074508	.076982	.065704	.066712
MD	.074145	.076336	.065140	.066081	.073898	.076332	.065104	.066068
	RMSE($\hat{\delta}$)				SD($\hat{\delta}$)			
CF	.106819	.116410	.125905	.132584	.106276	.115156	.108713	.117753
QLIML	.107902	.118604	.094155	.103526	.106946	.116472	.093296	.102338
mGMM	.135496	.148364	.114708	.122376	.120365	.130357	.102978	.110010
cMD	.107595	.115777	.094605	.102796	.107592	.115562	.094569	.102651
MD	.107138	.114848	.094074	.102169	.107046	.114801	.094078	.102141
	RMSE($\hat{\eta}$)				SD($\hat{\eta}$)			
CF	.092770	.123793	.107716	.128835	.092549	.123702	.100161	.124508
QLIML	.095036	.128690	.086661	.113820	.094068	.127413	.085600	.112983
mGMM	.108038	.142233	.096547	.124093	.101424	.135053	.091019	.118997
cMD	.093617	.124352	.086388	.112209	.093618	.124358	.086323	.112214
MD	.093134	.123649	.085921	.111645	.093101	.123641	.085903	.111646

cMD, mGMM and QLIML are all asymptotically equivalent. In simulation I, thus, all estiamators are asymptotically equivalent including CF. Other simuations (II, III, IV) have at least one GIME failing, and MD, cMD and mGMM are efficient relative to both QLIML and CF. Standard deviations are also presented along with RMSEs.

There are some points to be mentioned: 1) These results show that there can be cases where minimum distance and its equivalent estimators can outperform CF and QLIML in finite sample. 2) Minimum distance estimators and equivalent ones except mGMM seem to behave quite similar to each other in Simulation I as predicted. 3) MD-QLIML performs remarkably well while asymptotically equivalent mGMM had poor finite sample behavior. Compared to other estimators, MD-QLIML has usually the best performance and, even when it is the second best, the RMSE difference from the best is not large.

CHAPTER 3

SHORT PANEL DATA QUANTILE REGRESSION MODEL WITH SPARSE CORRELATED EFFECTS

3.1 Introduction

Application of quantile regression to panel data is attractive to empirical researchers. Compared to conditional mean regression, quantile regression provides a more thorough description of the population distribution by nature. With its application to panel data, the unobserved individual effects can be accounted for so that a potential source of endogenous variation is eliminated. A natural quantile analogue of a linear panel data model, however, suffers from the well-known incidental parameters problem as in generic nonlinear models (Neyman and Scott, 1954). Rosen (2012) shows that with time dimension fixed, the conditional quantile restriction alone cannot identify the regression coefficients in general. Additional point-identifying restrictions considered in the literature so far assume at least one of the following: (i) infinite time dimension, (ii) pure location-shifting unobserved effects, and (iii) a certain degree of within-group independence of the regression errors (e.g. Koenker, 2004; Rosen, 2012; Lamarche, 2010; Canay, 2011; see Section 2). Depending on the empirical contexts, these assumptions may not be credible for short panel data analysis, and any breakdown of such identifying restrictions will result in an inconsistent estimation. The purpose of this chapter is to study an alternative point-identifying model restriction and feasible estimation procedure for linear panel data quantile regression with a fixed time dimension.

The main contributions of this chapter are as follows. First, I propose a new point-identifying restriction for a linear panel data quantile regression model with a finite time dimension. The new model restriction reasonably accounts for the τ-quantile-specific time-invariant heterogeneity, and allows arbitrary within-group dependence of regression errors. The generalized Chamberlain device is taken analogously as a control function to capture τ-quantile-specific time-invariant endogenous variations. Endogeneity due to an observability pattern in unbalanced panel data can be accounted for as well. Second, asymptotic properties of a non-convex penalized estimator are studied. To treat the high-dimensional nature of the generalized Chamberlain device, a nonconvex penalized estimator is adopted. Compared to exact sparse models for cross-sectional data in Wang, Wu and Li (2012; WWL) and Sherwood and Wang (2016; SW), the model in consideration accounts for an approximation error in the sparse model, and within-group dependence of panel data. A convergence rate and asymptotic distribution of the oracle estimator is studied under both exact and approximate sparsity assumptions. A sparse version of the standard partially linear semiparametric model asymptotics is derived under approximate sparsity. The proposed penalized estimator is shown to have an oracle property in the sense that the estimator based on the true sparse model belongs to local minima of penalized quasi-likelihood with probability tending to one. The lower bound condition on the smallest magnitude of nonzero coefficients, so-called the beta-min condition is relaxed compared to the one given in SW. Third, a transformation of sieve-approximated correlated effects into a generalized Mundlak form is proposed to make the sparsity assumption more plausible in some cases. Given a choice of sieve basis elements, the approximating terms are transformed into time average and deviations. Whenever the sieve elements contain a first-order polynomial term, both a classical Chamberlain and Mundlak form of correlated effects is nested by the transformed approximating terms as a special case of true sparse models. The Monte Carlo simulation shows that, depending on the true model, the estimator using a generalized Chamberlain form can outperform the one using a Mundlak form, and vice versa. Fifth, an empirical application to birth weight analysis demonstrates a convincing case where the proposed estimator works as intended in real data.

The rest of this chapter is organized as follows: Section 3.2 gives a brief literature review of linear panel data quantile regression. In Section 3.3, the new identifying restriction is explained and formalized. Along with sieve-approximated correlated effects, nonconvex penalized estimation and its asymptotic properties are presented in Section 3.4. Simulation results are discussed in Section 3.5. The empirical application to birth weight analysis is in Section 3.6. Section 3.7 contains concluding remarks.

3.2 Literature on Linear Panel Data Quantile Regression

The literature on linear panel data quantile regression model has been growing rapidly in recent years. First, there are several studies where both the time dimension T and sample size N are assumed to be large. Koenker (2004) proposed penalized estimation under the pure location shift restriction and large (T, N) asymptotics. Lamarche (2010) showed that it is unbiased under a zero median condition on fixed effects, and derived an optimal choice of a penalty parameter. Harding and Lamarche (2016) considered a semiparametric correlated effects model in a similar framework to Koenker (2004) and Lamarche (2010). Kato, Galvao Jr. and Montes-Rojas (2012) formally studied asymptotic results when (T, N) tends to infinity. They relaxed the intertemporal independence assumption in Koenker (2004), and found that for asymptotic normality, the rate condition imposed on T is more restrictive than the one found in generic nonlinear models due to non-smoothness of the loss function. This result indicates that its short panel data application is even less appealing.

Second, point-identifying restrictions and estimation methods for the fixed T case are studied. Rosen (2012) showed weak conditional independence of regression errors across time together with some support and tail conditions imply point-identification. Canay (2011) also showed that an alternative conditional independence restriction in the random coefficient framework is sufficient for identification, and he proposed a simple estimation method when unobserved effects are pure location shifters. When the independence assumption is strengthened to i.i.d., Graham, Hahn and Powell (2009) showed that there is no incidental parameters problem since the first-differenced regression errors have a zero conditional median. Abrevaya and Dahl (2008), without explicitly setting up rigorous model restrictions, applied a quantile analogue of a correlated random effect model to analyze the effects of birth inputs on birthweight.

Apart from linear panel data quantile regression models, there are several related works on panel data models. Wooldridge and Zhu (2016, manuscript) proposed high-dimensional probit model with sparse correlated effects under fixed T. Arellano and Bonhomme (2016) considered a class of nonlinear panel data models under fixed T where unobserved heterogeneity is nonparametrically modelled. Graham, Hahn, Poirier, Powell (2015, manuscript) extends the correlated random coefficients representation of linear quantile regression to panel data under fixed T. Chernozhukov, Fernández-Val, Hahn, Newey (2013) studied a general nonseparable model assuming time-homogeneous errors and large (T, N).

3.3 Identification

3.3.1 Generalized Chamberlain Device

One of the essential advantages from using panel data is to resolve the potential endogeneity problem that arises from unobserved time-constant heterogeneity. The unobserved effects are typically specified as unknown coefficient parameters for individual dummy variables. In the linear panel data conditional mean model, such specification is useful: both the differencing method and direct control of dummies yield consistent estimators under mild conditions. Unfortunately, panel data quantile regression with individual dummies suffers from an incidental parameters problem in general.

I propose a generalized Chamberlain device as an alternative approach to achieve elimination of time-invariant endogeneity in the spirit of a control function approach. The idea is to control time-constant endogenous variation (regressor-correlated variation) only, not the whole heterogeneous individual effect in the unobserved error. A well-known example of the conditional mean model clearly demonstrates this idea: Suppose, for $1 \leq i \leq N$ and $1 \leq t \leq T$,

$$y_{it} = \mathbf{x}_{it}\beta + c_i + v_{it} \tag{3.1}$$

where $\mathbf{x}_{it} \in \mathbb{R}^K$, $\mathbf{x}_i \equiv (\mathbf{x}_{i1}, \cdots, \mathbf{x}_{iT})$, $E[v_{it}|\mathbf{x}_i, c_i] = 0$, and $\mathbf{x}_{it}$ is assumed to be time-varying and continuously distributed. Here, the unobserved time-invariant effect is denoted as c_i following Chamberlain (1984). By taking the conditional expectation of y_{it} on $\mathbf{x}_i$, we have

$$E[y_{it}|\mathbf{x}_i] = \mathbf{x}_{it}\beta + g(\mathbf{x}_i) \tag{3.2}$$

for some measurable function $g : \mathbb{R}^{TK} \rightarrow \mathbb{R}$. Note that the unknown arbitrary function g does not depend on the time index t. Then, regression with sieve-approximated $g(\mathbf{x}_i)$, for example, yields a consistent estimator of β. In this sense, the conditional moment restriction (3.2) can be viewed as a control function approach counterpart for the linear panel data model. (see section 19.8.2 of Li and Racine (2007) for details.) Such control function g will be called a "generalized Chamberlain device" or "correlated effect" in this chapter. To date, this approach has not been considered seriously in the linear panel data literature since the methods based on direct control or removal of the individual effect c_i equivalently eliminate potential endogeneity without much difficulty.

The generalized Chamberlain device is taken analogously in quantile regression setting. Suppose, for $1 \leq i \leq N$ and $1 \leq t \leq T$, the structural equation is

$$y_{it} = \mathbf{x}_{it}\beta + u_{it} \tag{3.3}$$

where, for simplicity, $\mathbf{x}_{it}$ is assumed to be time-varying and continuously distributed. We consider balanced panel data from now on unless explicitly mentioned. For each $\tau \in (0, 1)$, we can write

$$Q_\tau (y_{it}|\mathbf{x}_i) = \mathbf{x}_{it}\beta(\tau) + g_t(\mathbf{x}_i, \tau) \tag{3.4}$$

for some measurable function $g_t : \mathbb{R}^{TK} \times (0, 1) \rightarrow \mathbb{R}$. The function $g_t(\mathbf{x}_i, \tau)$ represents τ-quantile-specific endogenous variation contained in u_{it} that is allowed to vary across time given $\mathbf{x}_i$. Unfortunately, such g_t is not separately identifiable from $\mathbf{x}_{it}\beta(\tau)$ in general. Now, assume that any endogenous variation contained in u_{it} is time-constant in the sense that $g_t(\mathbf{x}_i, \tau)$ does not depend on t but is allowed to have a constant level difference across time. Then, for some constants, $k_t(\tau)$s, we have

$$Q_\tau (y_{it}|\mathbf{x}_i) = \mathbf{x}_{it}\beta(\tau) + g(\mathbf{x}_i, \tau) + k_t(\tau) \tag{3.5}$$

where $k_T(\tau) = 0$ is imposed for normalization. Note that (3.5) takes a quantile analogue of (3.2) with the introduction of time effects, and that it formalizes the "time-constant endogeneity" assumption for u_{it}. It does not rely on either additivity of composite error, $c_i + v_{it}$, or the widely used conditional quantile restriction $Q_\tau (v_{it}|\mathbf{x}_i) = 0$.

3.3.2 Model Restriction and Identification

In this subsection, the time-constant endogeneity assumption is used to derive the generalized Chamberlain device for a formal structural equation. Together with the derived control function, a set of model restrictions that attains point-identification is presented.

For each $i = 1, \cdots, N$ and $t = 1, \cdots, T$, we observe $(y_{it}, \mathbf{x}_{it}, \mathbf{z}_i, \mathbf{v}_t)$. The response variable is $y_{it} \in \mathbb{R}$, and the covariates are time/individual-varying variables $\mathbf{x}_{it} \in \mathbb{R}^{K_1}$, time-constant variables $\mathbf{z}_i \in \mathbb{R}^{K_2}$, and individual-constant variables $\mathbf{v}_t \in \mathbb{R}^{K_3}$. The covariates are allowed to contain both continuous and discrete variables which will be notated by tilde and dot accents, respectively. Specifically, $\tilde{\mathbf{x}}_{it} \in \mathbb{R}^{K_1^c}$ and $\tilde{\mathbf{z}}_i \in \mathbb{R}^{K_2^c}$ are continuous while $\dot{\mathbf{x}}_{it} \in \mathbb{R}^{K_1^d}$ and $\dot{\mathbf{z}}_i \in \mathbb{R}^{K_2^d}$ are discrete where we have $K_1 = K_1^c + K_1^d$ and $K_2 = K_2^c + K_2^d$ by construction. For individual-varying variables, we assume random sampling conditional on individual-constant variables.

Assumption 1 (Random Sample) $\{y_{it}, \mathbf{x}_{it}, \mathbf{z}_i\}_{t=1}^T$ are i.i.d. across i conditional on $\{\mathbf{v}_t\}_{t=1}^T$.

Since we consider linear quantile regression models, it is natural to assume the structural equation for y_{it} to be a linear function of the observed covariates. Throughout the paper, the structural equation is defined as follows: For $i = 1, \cdots, N$ and $t = 1, \cdots, T$,

$$y_{it} = \mathbf{x}_{it}\beta + \mathbf{z}_i\eta + \mathbf{v}_t\xi + u_{it} \tag{3.6}$$

where u_{it} is an unobserved error. The equation (3.6) describes the data generating process of the response variable y_{it}, which is typically implied by economic theories and specific empirical contexts. We may also think of it as an equation in the researcher's mind. Following Hurwicz (1950) and Koopmans and Reiersϕl (1950), it constitutes a 'structure' when paired with a joint distribution function of $(\{u_{it}, \mathbf{x}_{it}\}_{t=1}^T, \mathbf{z}_i)$ conditional on $\{\mathbf{v}_t\}_{t=1}^T$

$$F_{\{u_{it},\mathbf{x}_{it}\}_{t=1}^T,\mathbf{z}_i|\{\mathbf{v}_t\}_{t=1}^T}(u_{i1}, \cdots, u_{iT}, \mathbf{x}_{i1}, \cdots, \mathbf{x}_{iT}, \mathbf{z}_i|\mathbf{v}_1, \cdots, \mathbf{v}_T). \tag{3.7}$$

Depending on the model restrictions imposed on (3.6) and (3.7), the interpretation of parameter (β, η, ξ) changes. For example, conditional quantile restrictions on u_{it} with different values of $\tau \in (0,1)$ will change the value and interpretation of (β, η, ξ) in general.

Under the model restriction of a generalized Chamberlain device, neither η nor ξ can be identified. However, following this argument shows that it is important to include time constant regressor $\mathbf{z}_i$ when the control function g is constructed: Taking conditional τ-quantile of y_{it}, we have

$$Q_\tau(y_{it}|\mathbf{x}_i, \mathbf{z}_i, \{\mathbf{v}_t\}_{t=1}^T) = \mathbf{x}_{it}\beta(\tau) + \mathbf{z}_i\eta(\tau) + \mathbf{v}_t\xi(\tau) + f_t(\mathbf{x}_i, \mathbf{z}_i, \{\mathbf{v}_t\}_{t=1}^T, \tau) \tag{3.8}$$

for some measurable function $f_t : \mathbb{R}^{T(K_1+K_3)+K_2} \times (0,1) \to \mathbb{R}$. Note that the effect of $\{\mathbf{v}_t\}_{t=1}^T$ on f_t and that of t on f_t are confounded. Thus, without loss of generality, we can write $f_t(\mathbf{x}_i, \mathbf{z}_i, \{\mathbf{v}_t\}_{t=1}^T, \tau) = h_t(\mathbf{x}_i, \mathbf{z}_i, \tau)$ for some h_t. (The notation neglects randomness arising from $\{\mathbf{v}_t\}_{t=1}^T$ since $\{\mathbf{v}_t\}_{t=1}^T$ is always fixed in this chapter.) The time-constant endogeneity assumption, then, implies $h_t(\mathbf{x}_i, \mathbf{z}_i, \tau) = h(\mathbf{x}_i, \mathbf{z}_i, \tau) + m_t(\tau)$ for some function h and some constant m_t. The conditional quantile of y_{it} can be written as

$$Q_\tau(y_{it}|\mathbf{x}_i, \mathbf{z}_i, \{\mathbf{v}_t\}_{t=1}^T) = \mathbf{x}_{it}\beta(\tau) + g(\mathbf{x}_i, \mathbf{z}_i, \tau) + k_t(\tau) \tag{3.9}$$

where $g(\mathbf{x}_i, \mathbf{z}_i, \tau) = \mathbf{z}_i\eta(\tau) + h(\mathbf{x}_i, \mathbf{z}_i, \tau)$ and $k_t(\tau) = \mathbf{v}_t\xi(\tau) + m_t(\tau)$. Assumption 2 below summarizes the model restriction of the generalized Chamberlain device. From now on, we will drop $\{\mathbf{v}_t\}_{t=1}^T$ in the conditioning and treat k_ts as parameters to be estimated. The parameters' dependence on τ will also be omitted. The time dummies are denoted as d_t for $t = 1, \cdots, T-1$, and they will be considered together with $\mathbf{x}_{it}$ as in $\mathbf{w}_{it} = \begin{bmatrix} \mathbf{x}_{it} & d_1 & \cdots & d_{T-1} \end{bmatrix}$. The kth element of $\mathbf{w}_{it}$ is written as w_{itk}. The corresponding coefficient parameters are defined as $\boldsymbol{\beta} = \left(\beta', \upsilon'\right)' \in \mathbb{R}^{K_1+(T-1)}$.

Assumption 2 (Correlated Effect) There exists a measurable function $g : \mathbb{R}^{K_1T+K_2} \to \mathbb{R}$ such that, for all (i,t)

$$Q_\tau(y_{it}|\mathbf{x}_i, \mathbf{z}_i) = \mathbf{w}_{it}\boldsymbol{\beta} + g(\mathbf{x}_i, \mathbf{z}_i) \tag{3.10}$$

Assumption 2 is a new model restriction that takes a control function approach for the linear panel data quantile regression model. Note that the correlated effect g depends on time-constant variables $\mathbf{z}_i$ that enter the structural equation. Although the causal effects of $\mathbf{z}_i$ on y_{it} are not identified, it is important to include $\mathbf{z}_i$ as arguments of g. Also, there may be some deterministic relationship among $(\mathbf{x}_i, \mathbf{z}_i)$ which should be dealt with. For example, we may have $x_{itk} = x_{i,t',k}$ for some $\left(t, t', k\right)$ with probability one. Then, having both variables is redundant for g, and one should be removed from the specification. Similarly, if there is a functional relationship between the covariates such as $x_{itk_2} = x_{itk_1}^2$, only the one containing finer information, x_{itk_1}, should remain. Throughout the paper, such redundancy is assumed away for simplicity.

The following theorem shows that a certain degree of richness in the support of $\{\mathbf{w}_{it}\}_t$ and well-behaved error distribution is sufficient for point-identification of $\boldsymbol{\beta}$ under Assumption 1–2. Define $\varepsilon_{it} \equiv y_{it} - \mathbf{w}_{it}\boldsymbol{\beta} - g\left(\mathbf{x}_i, \mathbf{z}_i\right)$, and let $f_{it}\left(\varepsilon\right)$ be a density of ε_{it} conditional on $\left(\mathbf{x}_i, \mathbf{z}_i\right)$. The condition imposed on f_{it} below is part (i) and (ii) of Assumption 3 in Section 3.4.2.

Theorem 3.3.1 (Identification of $\boldsymbol{\beta}$) Let $\bar{\mathbf{W}}_i = ((\mathbf{w}_{i2} - \mathbf{w}_{i1})', \cdots, \left(\mathbf{w}_{iT} - \mathbf{w}_{i(T-1)}\right)')'$. Assume Assumption 1–2 and that $f_{it}\left(\varepsilon\right)$ is continuous and uniformly bounded away from 0 and ∞ in a neighborhood of 0. Suppose that the support of $(\mathbf{w}_{i1}, \cdots, \mathbf{w}_{iT})$ contains J points $(\mathbf{w}_{i1}^{(j)}, \cdots, \mathbf{w}_{iT}^{(j)})$ $1 \leq j \leq J$ such that $JT \times (K_1 + T - 1)$ matrix

$$[\ \bar{\mathbf{W}}_i^{(1)\prime} \ \cdots \ \bar{\mathbf{W}}_i^{(J)\prime}\ ]' \tag{3.11}$$

has full column rank, the pmf of $(\dot{\mathbf{x}}_{it}^{(j)}, \dot{\mathbf{x}}_{it'}^{(j)})$ satisfies $p(\dot{\mathbf{x}}_{it}^{(j)}, \dot{\mathbf{x}}_{it'}^{(j)}) > 0\ \forall j$, and the pdf of $(\tilde{\mathbf{x}}_{it}^{(j)}, \tilde{\mathbf{x}}_{it'}^{(j)})$ satisfies $f_{(\tilde{\mathbf{x}}_{it}, \tilde{\mathbf{x}}_{it'})|(\dot{\mathbf{x}}_{it}, \dot{\mathbf{x}}_{it'})}(\tilde{\mathbf{x}}_{it}^{(j)}, \tilde{\mathbf{x}}_{it'}^{(j)}|\dot{\mathbf{x}}_{it}^{(j)}, \dot{\mathbf{x}}_{it'}^{(j)}) > 0\ \forall j$ where $f_{(\tilde{\mathbf{x}}_{it}, \tilde{\mathbf{x}}_{it'})|(\dot{\mathbf{x}}_{it}, \dot{\mathbf{x}}_{it'})}$ has continuous extension at each $(\mathbf{w}_{it}^{(j)}, \mathbf{w}_{it'}^{(j)})$. Then, $\boldsymbol{\beta}$ is identified.

The result above is not surprising since the current specification does not have incidental parameters. It shows that specification with an unknown function common to every individual has more identifying power than one with unknown parameters unique to each individual. For the rest of the paper, we assume point-identification.

3.3.3 Case of Unbalanced Panel Data with Time-constant Endogeneity

When some time periods are missing for some individuals in the observed panel data, potential endogeneity related to observability should be treated as well. We assume the structural equation for all observed units and time periods are homogeneous. Then, with the introduction of selection indicators and auxiliary balanced data, the generalized Chamberlain device can be modified to account for time-constant endogeneity related to observability. The approach can be regarded as a nonparametric version of the correlated random effect models studied by Wooldridge (2009).

First, define selection indicator s_{it} to be a binary function that takes 1 if $(y_{it}, \mathbf{x}_{it}, \mathbf{z}_i)$ is observed for t, 0 otherwise. In addition, consider an auxiliary balanced panel data $(s_{it}y_{it}, s_{it}\mathbf{x}_{it}, s_{it}\mathbf{z}_i, s_{it})$ for $i = 1, \cdots, N$, $t = 1, \cdots, T$. A corresponding structural equation is assumed to be

$$s_{it}y_{it} = s_{it}\mathbf{x}_{it}\beta + s_{it}\mathbf{z}_i\eta + s_{it}\mathbf{v}_t\xi + s_{it}u_{it} \tag{3.12}$$

which is derived by multiplying s_{it} to the original structural equation. The (3.12) is restrictive only for observed time periods and it assumes the structural equations are homogeneous across all observed units and time periods. The conditional τ-quantile of $s_{it}y_{it}$ conditional on $\boldsymbol{S}_T = (\{s_{it}\mathbf{x}_{it}\}_{t=1}^{T}, s_{it}\mathbf{z}_i, \{s_{it}\mathbf{v}_t\}_{t=1}^{T}, \mathbf{s}_i)$ is

$$Q_\tau\left(s_{it}y_{it}|\boldsymbol{S}_T\right) = s_{it}\mathbf{x}_{it}\beta + s_{it}\mathbf{z}_i\eta + s_{it}\mathbf{v}_t\xi + s_{it}g_t(\{s_{it}\mathbf{x}_{it}\}_{t=1}^{T}, s_{it}\mathbf{z}_i, \{s_{it}\mathbf{v}_t\}_{t=1}^{T}, \mathbf{s}_i) \tag{3.13}$$

for some g_t where $\mathbf{s}_i = (s_{i1}, \cdots, s_{iT})$. Then, based on the previous argument, the time-constant endogeneity assumption implies a conditional quantile restriction

$$Q_\tau(s_{it}y_{it}|\{s_{it}\mathbf{x}_{it}\}_{t=1}^{T}, s_{it}\mathbf{z}_i, \mathbf{s}_i) = s_{it}\mathbf{x}_{it}\beta + s_{it}g(\{s_{it}\mathbf{x}_{it}\}_{t=1}^{T}, s_{it}\mathbf{z}_i, \mathbf{s}_i) + s_{it}k_t\left(\mathbf{s}_i\right) \tag{3.14}$$

where $\{s_{it}\mathbf{v}_t\}_{t=1}^{T}$ is omitted and $(g(\cdot), k_t)$ depends on the selection indicator $\mathbf{s}_i$. Since (3.14) is not restrictive when $s_{it} = 0$, we may write an equivalent model restriction for (i, t) with $s_{it} = 1$ as

$$Q_\tau(y_{it}|\{\mathbf{x}_{it}\}_{\{t:s_{it}=1\}}, \mathbf{z}_i, \mathbf{s}_i) = \mathbf{x}_{it}\beta + g(\{\mathbf{x}_{it}\}_{\{t:s_{it}=1\}}, \mathbf{z}_i, \mathbf{s}_i) + k_t\left(\mathbf{s}_i\right). \tag{3.15}$$

Since the generalized Chamberlain device, $g(\{\mathbf{x}_{it}\}_{\{t:s_{it}=1\}}, \mathbf{z}_i, \mathbf{s}_i)$ in (3.15) is a function of selection operator $\mathbf{s}_i$, it now accounts for time-constant endogenous variation due to observability.

Note that $\mathbf{s}_i$ acts as a classification device for the observed pattern of each individual. If there is no endogeneity related to observability, g and k_t will not depend on $\mathbf{s}_i$ when g is assumed to have an additive form (3.16) in the next section.

Identification of β (not $\boldsymbol{\beta}$) can be trivially achieved under conditions in Theorem 3.3.1 for $\{\mathbf{w}_{it}\}_{\{t:s_{it}=1\}}\,|\mathbf{s}_i = \tilde{\mathbf{s}}_i$ such that $P\,(\mathbf{s}_i = \tilde{\mathbf{s}}_i) > 0$ and $\sum_{i=1}^{T} \tilde{s}_{it} \geq 2$. In other words, if there exists a positive fraction of observable units with a certain number of multiple periods, and if the support of regressors are rich enough, the parameter of interest is identified. The estimation procedure will be applied to the auxiliary balanced panel data. When g has an additive form, an essential difference in estimation is that each fraction of individuals with different observability pattern $\mathbf{s}_i$ is allowed to have different additive components of g for each x_{itk} and z_{ik}.

3.4 Estimation

For estimation, the correlated effect g is approximated by sieve spaces. The approximated g has high-dimensionality for three reasons: First, the number of approximating terms is infinite in general. Second, the number of arguments in g is $TK_1 + K_2$, which can grow fast in T. Besides the problem that the truncation choice on sieve approximation can be limited with a large number of arguments, the existence of discrete variables can introduce more nontrivial problems. In particular, if discrete variables $(\dot{\mathbf{x}}_i, \dot{\mathbf{z}}_i)$ have rich enough supports, the total number of approximating terms can be comparable to, or larger than N even after we impose the additive functional form on g and truncate the approximating terms for additive components of continuous variables $(\tilde{\mathbf{x}}_i, \tilde{\mathbf{z}}_i)$. Such "too many regressors" problem arises from the fact that it is not obvious how to truncate approximating terms related to discrete variables in general. Note that N is the maximal number of linearly independent time-invariant regressors. Third, when the panel data is unbalanced, more complex observability patterns result in a larger number of nonparametric nuisance components to be approximated since we allow different functional forms of g for each group of individuals with a different pattern.

The reasons for high-dimensionality mentioned above indicate that the standard sieve truncation via information criteria or cross-validation is not always usable and effective. For high-dimesional models, a penalized estimation of a sparse model with the Least Absolute Shrinkage and Selection Operator (LASSO; Tibshirani, 1996) and its variants is popular due to its prediction accuracy and computational feasibility. For high-dimensional quantile regression models, Belloni and Chernozhukov (2011; BC), Wang, Wu and Li (2012; WWL) and SW (2016) studied properties of penalized estimators with certain penalty functions. Since the nonconvex penalty functions used in WWL (2012) and SW (2016) have oracle property under mild conditions, asymptotic distribution of resulting penalized estimators can be studied via that of oracle estimator. This is a big benefit from using nonconvex penalty function compared to LASSO which has oracle property only under a quite restrictive condition. Another practical benefit is that the relaxed estimation procedure such as "post-Lasso estimation" is not necessary for nonconvex penalized estimators applied to a large sample. In this chapter, two nonconvex penalty functions are considered: Smoothly Clipped Absolute Deviation (SCAD; Fan and Li, 2001) and Minimax Concave Penalty (MCP; Zhang, 2010). For details of a general class of penalty functions, see Fan and Lv (2009) and Lv and Fan (2009) for example. Besides overcoming a nontrivial high-dimensionalty problem, it is expected that the penalized estimator improves over the standard truncated sieve estimators under sparsity assumption (Belloni and Chernozhukov, 2011a). Note that the penalized estimator selects the relevant sparse terms only while the relevant terms can be excluded from the first K elements selected by standard truncated sieve estimators. To make the sparsity assumption more plausible in some cases, a transformation of the approximated correlated effect into a generalized Mundlak form is proposed.

3.4.1 Sieve-approximated Correlated Effect

In this chapter, the specific sieve space in which g belongs is not assumed. The theoretical framework is quite flexible and can accomodate various specifications. As long as the true function g can be sparsely approximated by a collection of terms that satisfies regularity conditions, such a collection of terms can be used. Here, we briefly cover one useful example of the sieve approximation of g with an additive form, smoothness of additive components of g for continuous covariates, and finiteness of support for discrete covariates. While a smoothness assumption is quite standard, the additive function space can be replaced by a tensor product sieve space in general. We may also consider using multiple sieve spaces together so that basis elements can be mixed (Bunea, Tsybakov, Wegkamp, 2007; Belloni, Chen, Chernozhukov, Hansen, 2012).

The additivity requires g to be represented by the sum of the univariate functions of each argument,

$$g(\mathbf{x}_i, \mathbf{z}_i) = g_0 + \sum_{t=1}^{T} \sum_{k=1}^{K_1} g_{tk}^{x}(x_{itk}) + \sum_{k=1}^{K_2} g_k^{z}(z_{ik}) \tag{3.16}$$

where $g_0 \in \mathbb{R}$ is a constant. For identification purposes, $E\left[g_{tk}^{x}(x_{itk})\right] = E\left[g_k^{z}(z_{ik})\right] = 0\ \forall t, k$ is typically assumed but we may instead drop a constant term (if there exists any) in the sieve elements for each g_{tk}^{x} and g_k^{z}. Given additivity, smoothness restriction is imposed on the additive components of g with continuous covariates, $(g_{tk}^{x}\left(\tilde{x}_{itk}\right), g_k^{z}(\tilde{z}_{ik})\Big)$. The Hölder condition of a certain order is the most popular choice. (see Chen, 2007). Finite supports are assumed for components with discrete covariates, $(g_{tk}^{x}\left(\dot{x}_{itk}\right), g_k^{z}(\dot{z}_{ik})\Big)$, which implies that relevant approximating errors will be exactly zeros with large enough N.

For approximating $g_{tk}^{x}(\tilde{\mathbf{x}}_{it})$ and $g_k^{z}(\tilde{\mathbf{z}}_i)$, B-spline elements are frequently used. See Schumaker (2007) for details. The following shows how the B-spline basis can be implemented in practice: Given a knot sequence $0 = t_0 < t_1 < \cdots < t_{J_N} < t_{J_N+1} = 1$, degree p B-spline elements are $\left\{1, x, \cdots, x^p, (x - t_1)_+^p, \cdots, \left(x - t_{J_N}\right)_+^p\right\}$ where the range of continuously distributed x is assumed to be $[0, 1]$ and $(\cdot)_+^p = (\max\{0, \cdot\})^p$. Then, the approximated $g_{tk}^{x}(\tilde{x}_{itk})$, for example, can be written as

$$s_{tk}^{x}(\tilde{x}_{itk}) = \sum_{q=1}^{p} \gamma_{tkq}^{x} \tilde{x}_{itk}^{q} + \sum_{j=1}^{J_N} \gamma_{tk(p+j)}^{x} \left(\tilde{x}_{itk} - t_j\right)_+^p \tag{3.17}$$

where the constant term is removed for identification. In several contexts, nonparametric or semiparametric conditional quantile estimators using B-splines are shown to achieve the optimal convergence rate with $J_N \asymp N^{\frac{1}{2r+1}}$ under regularity conditions where r denotes the degree of Hölder condition. For example, He, Zhu and Fung (2002) showed the result for a univariate semiparametric component in the panel data model with an unspecified dependence structure. If the optimal growth rate is the same for the additive semiparametric model, Assumption 5 in the next subsection is conservatively satisfied when we impose $p_N \asymp N^{\frac{1}{2r+1}}$ and $r > 1$.

For discrete $\dot{x}_{itk}$ and $\dot{z}_{ik}$, we do not rely on a smoothness assumption in general. The corresponding g^x_{tk} and g^z_k function can be exactly expressed as a linear combination of indicator functions that take value 1 on the support elements. Suppose that a discrete random variable $\dot{x}_{itk}$ has realized support elements $\{a_s\}_{s=1}^{S^x_{tk,N}}$ in the given sample. Then, without loss of generality, the function $g^x_{tk}(\dot{x}_{itk})$ can be written (or approximated) as

$$s^x_{tk}(\dot{x}_{itk}) = \sum_{s=1}^{S^x_{tk,N}} \gamma^x_{tks} 1\left[\dot{x}_{itk} = a_s\right] \tag{3.18}$$

where indicator functions $1\left[\dot{x}_{itk} = a_s\right]$ act as sieve basis elements of the function space in which $g^x_{tk}(\dot{x}_{itk})$ lies. To meet the identification condition $E\left(g^x_{tk}(\dot{x}_{itk})\right) = 0$, we may equivalently drop one of the indicator terms. Since $\dot{x}_{itk}$ is assumed to have finite support, for some $\overline{S}^x_{tk} < \infty$, we have $S^x_{tk,N} \xrightarrow{p} \overline{S}^x_{tk}$ as N tends to infinity, and the approximation error becomes zero. However, since there can be multiple discrete variables with $S^x_{tk,N}$ (or $S^z_{k,N}$) whose total is quite large relative to N, model selection is inevitable in some cases. Note that the approximating terms for discrete g components do not have a natural way to regularize the dimension as in the case of splines with increasing knots under smoothness assumption. Unless some additional assumption is employed, all of the terms in (3.18) should be included as regressors in principle.

As discussed at the beginning, sparsity on the approximated correlated effect is assumed for a feasible inference. Sparsity in our context means that only a small number of approximating terms have true nonzero coefficients. In other words, the correlated effect is regular enough so we need only a small number of variables to describe it well. Recently, the sparsity assumption is gaining more credibility as the 'bet on the sparsity' principle is understood better (Hastie, Tibshirani, Wainwright, 2015). Since the validity of sparsity depends on the choice of basis, a specific basis (or mixture of them) should be selected carefully.

Given a set of approximating terms, a transformation into a generalized Mundlak form is proposed for time-varying regressor parts, g_{tk}^{x}s. The idea is to take the time averages and deviations given the common approximating terms of g_{tk}^{x} for $t = 1, \cdots, T$.

Definition 3.4.1 (Generalized Mundlak Form) Suppose the approximated $g_{tk}^{x}(x_{itk})$ of correlated effect is $s_{tk}^{x}(x_{itk}) = \sum_{s=1}^{S} \gamma_{tks} p_{ks}(x_{itk})$ for $t = 1, \cdots, T$. Then, given t_0, define a transformed $\tilde{p}_{ks}(x_{itk})$ as

$$\tilde{p}_{ks}(x_{itk}) = \begin{cases} \frac{1}{T}\sum_{t=1}^{T} p_{ks}(x_{itk}) & t = t_0 \\ p_{ks}(x_{itk}) - \frac{1}{T}\sum_{t=1}^{T} p_{ks}(x_{itk}) & t \neq t_o \end{cases} \tag{3.19}$$

If one of the approximating terms contains a first-order polynomial, that is, $p_{ks}(x_{itk}) = x_{itk}$ for some s, then both classical Chamberlain and Mundlak device is nested in the transformed $\tilde{p}_{tks}(x_{itk})$ as a special case of sparse models. Note that the basis elements in (3.18) also can be easily transformed into a form that contains a first-order polynomial. The choice of t_0 can be avoided if $p_{t_0 ks}\left(x_{it_0 k}\right) - \frac{1}{T}\sum_{t=1}^{T} p_{tks}(x_{itk})$ is also included in the approximating terms which will be defined as 'dictionary variables' in Subsection 3.4.2.

The rationale for the transformation (3.19) is the following. In empirical studies, it is often found that the estimators based on classical Chamberlain and Mundlak devices do not differ much while a Chamberlain device contains many more terms. If this is because the true coefficients of x_{itk} in the Chamberlain device are the same for all t, then the true coefficient of time deviation terms, $x_{itk} - \frac{1}{T}\sum_{t=1}^{T} x_{itk}$, in the generalized Mundlak form (3.19) are zeros. Similarly, if the true coefficients of $p_{ks}(x_{itk})$ in the approximated correlated effect are the same for all t, then the true coefficients of time deviations, $p_{ks}(x_{itk}) - \frac{1}{T}\sum_{t=1}^{T} p_{ks}(x_{itk})$, are zeros, and the generalized Mundlak form has far fewer number of approximating terms. This indicates that the selection over the generalized Mundlak form can have a more sparse submodel on the correlated effects.

3.4.2 Penalized Estimation via Non-convex Penalty Functions

Given the approximating terms for g, a nonconvex penalized estimator is proposed along with its asymptotic properties. The convergence rate and asymptotic distribution of the penalized estimator is studied indirectly by deriving those of an estimator based on the true sparse model. In turn, inference is conducted as if the estimated submodel is the true sparse model. This is mainly justified by two facts: (i) the proposed nonconvex penalized estimator is shown to have oracle property, and (ii) any submodel can be interpreted as an approximation to the true model by construction. In the following, the asymptotic properties of the true sparse estimator are discussed first, and then the oracle property of the penalized estimator is presented. Two approaches are considered: (i) exactly sparse model of g and (ii) approximately sparse model of g.

The approximating terms can be divided into two groups in general. A group to be penalized and another group not to be penalized. Each group of variables is denoted as $\boldsymbol{\pi}\left(\mathbf{x}_i, \mathbf{z}_i\right) \in \mathbb{R}^{p_N}$ and $\bar{\boldsymbol{\pi}}\left(\mathbf{x}_i, \mathbf{z}_i\right) \in \mathbb{R}^{\bar{p}}$, respectively. Note that the number of unpenalized term $\bar{p}$ is fixed and not allowed to depend on N, and that the total number of dictionary variable p_N can be very large relative to the sample size N (i.e. $p_N \gg N$) since ultra-high dimensionality is allowed for the proposed estimator. Then, g can be written as

$$g\left(\mathbf{x}_i, \mathbf{z}_i\right) = \bar{\boldsymbol{\pi}}\left(\mathbf{x}_i, \mathbf{z}_i\right) \bar{\boldsymbol{\gamma}} + \boldsymbol{\pi}\left(\mathbf{x}_i, \mathbf{z}_i\right) \boldsymbol{\gamma} + r_i \tag{3.20}$$

where $(\bar{\boldsymbol{\gamma}}, \boldsymbol{\gamma})$ has a conformable dimension and r_i is an approximation error. The terms to be penalized, $\boldsymbol{\pi}\left(\mathbf{x}_i, \mathbf{z}_i\right)$, will be called ‘dictionary variables’ from now on. There is no hard guideline about whether a given approximating variable should be penalized or not. However, a constant term, g_0 in our setting, is typically not recommended to be penalized and will not be treated as a dictionary variable in this paper. Given the choice of dictionary variables, $\mathbf{w}_{it}$ is redefined as

$$\mathbf{w}_{it} = \left[\begin{array}{ccccc} \mathbf{x}_{it} & d_1 & \cdots & d_{T-1} & \bar{\boldsymbol{\pi}}\left(\mathbf{x}_i, \mathbf{z}_i\right) \end{array}\right] \tag{3.21}$$

where $\mathbf{w}_{it} \in \mathbb{R}^{K_4}$ and $\bar{\boldsymbol{\pi}}\left(\mathbf{x}_i, \mathbf{z}_i\right)$ is assumed to contain a constant as the default. $\boldsymbol{\beta}$ is redefined accordingly. Also, dictionary variables, $\boldsymbol{\pi}\left(\mathbf{x}_i, \mathbf{z}_i\right)$, should be rescaled to have unit (pooled) sample variance. Otherwise, selection will get affected by the scales of the variables.

Under the sparsity assumption, only a small subset of dictionary variables have nonzero true coefficients. The cardinality of sparse coefficients is allowed to increase as N tends to infinity and its growth rate is ruled by a true sparse model given a sequence of dictionary variables. With increasing cardinality, the sparse approximation tends to the true function by construction. The framework is similar to the "approximate sparsity model" proposed in Belloni, Chernozhukov (2011a), Belloni, Chen, Chernozhukov, Hansen (2012) and Belloni, Chernozhukov, Hansen (2014) in its spirit. A difference is that the sparse model is assumed to be exact in the sense that the corresponding approximation error is not explicitly considered. Application of a penalized estimation to high-dimensional nonparametric modeling is also discussed by Fan and Li (2001). Note that, in contrast to the theoretical framework of Sherwood and Wang (2016), the parameter of interest β is not penalized in this chapter. In turn, there is no such pathological case where a parameter of interest is not selected in a penalized estimate.

The estimator based on the true sparse model is often called "oracle estimator" in high-dimensional statistics literature. The corresponding true sparse model will be refered to as "oracle model" in this chapter. Let A be the index set of sparse coefficients given $\boldsymbol{\pi}\left(\mathbf{x}_i, \mathbf{z}_i\right)$,

$$A = A_N = \left\{1 \leq j \leq p_N : \gamma_{oj} \neq 0\right\} \tag{3.22}$$

and its cardinality be $q_N = |A|$. By rearranging $\boldsymbol{\pi}\left(\mathbf{x}_i, \mathbf{z}_i\right)$, we may assume the first q_N elements of $\boldsymbol{\gamma}_o$ are nonzero and the remaining $p_N - q_N$ components are zeros i.e. $\boldsymbol{\gamma}_o = (\boldsymbol{\gamma}'_{oA}, \mathbf{0}'_{p_N - q_N})'$, and similarly, denote $\boldsymbol{\pi}\left(\mathbf{x}_i, \mathbf{z}_i\right) = (\boldsymbol{\pi}_A\left(\mathbf{x}_i, \mathbf{z}_i\right), \boldsymbol{\pi}_{A^c}\left(\mathbf{x}_i, \mathbf{z}_i\right))$. Then, we can define the oracle estimator as follows.

Definition 3.4.2 (Oracle Estimator)

$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}_A) = \underset{(\beta, \boldsymbol{\gamma}_A)}{\arg\min} \frac{1}{N} \sum_{t=1}^{T} \sum_{i=1}^{N} \rho_\tau\left(y_{it} - \mathbf{w}_{it}\boldsymbol{\beta} - \boldsymbol{\pi}_A\left(\mathbf{x}_i, \mathbf{z}_i\right) \boldsymbol{\gamma}_A\right) \tag{3.23}$$

where $\rho_\tau\left(u\right) = u(\tau - 1_{[u<0]})$

Regularity conditions for the oracle estimator and penalized estimator are given for exactly sparse model below. In the following, let $F_{it}\left(\varepsilon\right) = F\left(\varepsilon | \mathbf{x}_i, \mathbf{z}_i\right)$ be a conditional cdf of ε_{it} given

$(\mathbf{x}_i, \mathbf{z}_i)$, and $e_{it} \equiv y_{it} - \mathbf{w}_{it}\boldsymbol{\beta} - \boldsymbol{\pi}_A(\mathbf{x}_i, \mathbf{z}_i)\boldsymbol{\gamma}_A$ be the approximated regression error. The vector of all regressors is written as $\tilde{\mathbf{w}}_{it}^A = (\mathbf{w}_{it}, \boldsymbol{\pi}_A(\mathbf{x}_i, \mathbf{z}_i))$ while its stacked versions are denoted as $\tilde{\mathbf{W}}_i^A = (\tilde{\mathbf{w}}_{i1}^{A\prime}, \cdots, \tilde{\mathbf{w}}_{iT}^{A\prime})'$ and $\tilde{\mathbf{W}}_A = (\tilde{\mathbf{W}}_1^{A\prime}, \cdots, \tilde{\mathbf{W}}_N^{A\prime})'$.

Assumption 3 (Regression Error) (i) ε_{it} has the continuous conditional density function f_{it} (ii) f_{it} is uniformly bounded away from 0 and ∞ in a neighborhood of 0 $\forall t$. (iii) f_{it}' has a uniform upper bound in a neighborhood of 0 $\forall t$.

Assumption 4 (Covariates) (i) $\exists M_1 > 0$ such that $|\tilde{w}_{itk}| \leq M_1 \ \forall (i, t, k)$, (ii) $\exists C_1 > 0$, $C_2 > 0$ such that, with probability one, $C_1 \leq \lambda_{\min}(\frac{1}{N}\tilde{\mathbf{W}}_A'\tilde{\mathbf{W}}_A) \leq \lambda_{\max}(\frac{1}{N}\tilde{\mathbf{W}}_A'\tilde{\mathbf{W}}_A) \leq C_2$.

Assumption 5 (Sparse Model Size) $q_N = O(N^{C_3})$ where $C_3 < \frac{1}{3}$

Assumption 6 (Exact Sparsity) $g(\mathbf{x}_i, \mathbf{z}_i) = \boldsymbol{\pi}_A(\mathbf{x}_i, \mathbf{z}_i)\boldsymbol{\gamma}_A$ for each N.

Assumption 3 is a fairly standard regularity condition on regression error ε_{it}. Note that the within-group dependence of $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \cdots, \varepsilon_{iT})$ conditional on $(\mathbf{x}_i, \mathbf{z}_i)$ is allowed to be arbitrary under Assumption 1 and 3. Assumption 4 imposes boundedness on the regressors and eigenvalues of the Gram matrix $N^{-1}\tilde{\mathbf{W}}_A'\tilde{\mathbf{W}}_A$ in the oracle model. Assumption 5 restricts the growth rate of the oracle model size. It is required for a given dictionary variable sequence to be valid. Assumption 6 is an essential condition that characterizes exactly sparse g. It means that for each sample size, the model with q_N terms exactly describes the correlated effects. Under these conditions, the convergence rate and asymptotic normality results for oracle estimator of $\boldsymbol{\theta}_A = (\boldsymbol{\beta}', \boldsymbol{\gamma}_A')'$ is shown below.

Theorem 3.4.3 (Convergence Rate of Oracle Estimator) Suppose Assumption 1–6. Then,

$$\|\hat{\boldsymbol{\theta}}_A - \boldsymbol{\theta}_{oA}\| = O_p(\sqrt{N^{-1}q_N}) \tag{3.24}$$

Proof. The result follows from Lemma C.1.1 and C.1.4 in Appendix C.1.2. ■

Theorem 3.4.4 (Asymptotic Normality of Oracle Estimator) Suppose Assumption 1–6. Let G_N be an $l \times q_N$ matrix with l fixed and $G_N G_N' \to G$, a positive definite matrix. Then,

$$\sqrt{N} G_N \Sigma_N^{-\frac{1}{2}} (\hat{\boldsymbol{\theta}}_A - \boldsymbol{\theta}_{oA}) \overset{d}{\to} N\left(0_l, G\right) \tag{3.25}$$

where $\psi_\tau\left(\varepsilon_{it}\right) = \tau - I\left(\varepsilon_{it} < 0\right), \Psi_\tau\left(\boldsymbol{\varepsilon}_i\right) = \left(\psi_\tau\left(\varepsilon_{i1}\right), \cdots, \psi_\tau\left(\varepsilon_{iT}\right)\right)', B_N = diag\left(\{\{f_{it}\left(0\right)\}_t\}_i\right)$, $\mathbf{K}_N = \frac{1}{N} \tilde{\mathbf{W}}_A' B_N \tilde{\mathbf{W}}_A$, $\mathbf{S}_N = \frac{1}{N} \sum_{i=1}^N \tilde{\mathbf{W}}_i^{A\prime} \Psi_\tau\left(\boldsymbol{\varepsilon}_i\right) \Psi_\tau\left(\boldsymbol{\varepsilon}_i\right)' \tilde{\mathbf{W}}_i^A$, and $\Sigma_N = \mathbf{K}_N^{-1} \mathbf{S}_N \mathbf{K}_N^{-1}$.

Proof. The result follows from Lemma C.1.1 and C.1.4 in Appendix C.1.2. ■

In Theorem 3.4.3 and 3.4.4, we are not assuming primitive conditions regarding the sieve basis nature of $\boldsymbol{\pi}_A\left(\mathbf{x}_i, \mathbf{z}_i\right)$. It is only assumed that given a collection of dictionary variable $\boldsymbol{\pi}\left(\mathbf{x}_i, \mathbf{z}_i\right)$, true sparse regressors $\boldsymbol{\pi}_A\left(\mathbf{x}_i, \mathbf{z}_i\right)$ exist and satisfy the assumptions given, especially the exact sparse model condition. With additional assumptions on $\boldsymbol{\pi}_A\left(\mathbf{x}_i, \mathbf{z}_i\right)$, it is possible to derive a sparse version of standard partially linear semiparametric model results accounting for nonzero approximation error r_i.

The additional assumptions on $\boldsymbol{\pi}_A\left(\mathbf{x}_i, \mathbf{z}_i\right)$ require some definitions and notations. First, let $\mathcal{G}$ be a function space to which g belongs. For example, if g is assumed to be additive, and if $(\mathbf{x}_i, \mathbf{z}_i)$ contains continuous variables only, then an additive function space, $\mathcal{H}_1 + \cdots + \mathcal{H}_{K_4}$ where $\mathcal{H}$ is the Hölder space of certain degree, is a popular choice. Define $\mathcal{G}^*$ to be the subspace of $\mathcal{G}$ whose elements can be expressed by active elements of the dictionary variable sequence, $\{\boldsymbol{\pi}_A\left(\mathbf{x}_i, \mathbf{z}_i\right)\}_{N=1}^{\infty}$. Then, we can consider a weighted projection of each regressor in $\mathbf{w}_{it}$ onto $\mathcal{G}^*$

$$h_k^* \equiv \underset{h \in \mathcal{G}^*}{\arg\inf} \sum_{i=1}^{N} \sum_{t=1}^{T} E\left[f_{it}\left(0\right)\left(w_{itk} - h\left(\mathbf{x}_i, \mathbf{z}_i\right)\right)^2\right] \tag{3.26}$$

where w_{itk} is the kth element of $\mathbf{w}_{it}$. The corresponding population residual is written as $\Delta_{itk} = w_{itk} - h_k^*$. Stacked version of h_k^* and Δ_{itk} are denoted as follows: $\boldsymbol{h}_i = \left(h_1^*\left(\mathbf{x}_i, \mathbf{z}_i\right), \cdots, h_{K_4}^*\left(\mathbf{x}_i, \mathbf{z}_i\right)\right)$, $\mathbf{1}_T = (1, \cdots, 1)' \in \mathbb{R}^T$, $\mathbf{H} = \left(\boldsymbol{h}_1', \cdots, \boldsymbol{h}_N'\right)' \otimes \mathbf{1}_T \in \mathbb{M}_{NT \times K_4}$, $\boldsymbol{\Delta}_{it} = \left(\Delta_{it1}, \cdots, \Delta_{itK_4}\right)$, $\boldsymbol{\Delta}_i = \left(\boldsymbol{\Delta}_{i1}', \cdots, \boldsymbol{\Delta}_{iT}'\right)'$, $\boldsymbol{\Delta} = \left(\boldsymbol{\Delta}_1', \cdots, \boldsymbol{\Delta}_N'\right)'$ so that $\mathbf{W} = \mathbf{H} + \boldsymbol{\Delta}$ where $\mathbf{W} = (\mathbf{w}_{11}', \mathbf{w}_{12}', \cdots, \mathbf{w}_{NT}')'$. Then, it is easy to check $\hat{h}_k\left(\mathbf{x}_i, \mathbf{z}_i\right) = \boldsymbol{\pi}_A\left(\mathbf{x}_i, \mathbf{z}_i\right) \hat{\varphi}_k$ where $W_k = \left(w_{11k}, \cdots, w_{NTk}\right)'$, $\boldsymbol{\Pi}_A =$

$(\boldsymbol{\pi}_A(\mathbf{x}_1,\mathbf{z}_1)',\cdots,\boldsymbol{\pi}_A(\mathbf{x}_N,\mathbf{z}_N)')'\otimes\mathbf{1}_T$, $\hat{\varphi}_k=\left(\boldsymbol{\Pi}_A'B_N\boldsymbol{\Pi}_A\right)^{-1}\boldsymbol{\Pi}_A'B_NW_k$. Additional conditions on $\boldsymbol{\pi}_A(\mathbf{x}_i,\mathbf{z}_i)$ are given below.

Assumption 7 (Covariates) $\exists M_2>0$ such that $E[\Delta_{itk}^4]\leq M_2\ \forall\,(i,t,k)$

Assumption 8 (Approximate Sparse Correlated Effects) (i) $\sup_i|r_i|=O(N^{-1/2}q_N^{1/2})$

(ii) $N^{-1}\sum_{i=1}^N[h_k^*(\mathbf{x}_i,\mathbf{z}_i)-\hat{h}_k(\mathbf{x}_i,\mathbf{z}_i)]^2=o_p(1)\ \forall k$

Assumption 7 restricts the population residual of w_{itk} projected out of h_k^* to have finite fourth order moment. Assumption 8 is the essential condition that characterizes $\{\boldsymbol{\pi}(\mathbf{x}_i,\mathbf{z}_i)\}$ as the sieve basis elements that attain well-behaved sparse submodel. Part (i) assumes that the order of approximation error is uniformly dominated by $1/\sqrt{N}$. Part (ii) is a high-level assumption assuming that the sample analogue estimator of h_k^* converges to a true function with respect to the empirical L_2-norm. Since the convergence rate is not restricted and $f_{it}(0)$ is uniformly bounded, this is a fairly standard property of the sieve basis elements. For example, when q_N diverges, it can be shown that the uniform approximation property of $\boldsymbol{\pi}_A(\mathbf{x}_1,\mathbf{z}_1)$ (Newey, 1997)

$$\sup_{\mathbf{x}_i,\mathbf{z}_i}\left|h_k^*(\mathbf{x}_i,\mathbf{z}_i)-\boldsymbol{\pi}_A(\mathbf{x}_i,\mathbf{z}_i)\varphi_{o,k}\right|=O(q_N^{-\alpha})\tag{3.27}$$

for some $\alpha>0$ implies Assumption 8 (ii) under Assumption 1–5. With Assumptions 7 and 8 additionally assumed, the theorems below show a sparse version of the standard partially linear semiparametric model asymptotics. Denote $\hat{g}(\mathbf{x}_i,\mathbf{z}_i)=\boldsymbol{\pi}_A(\mathbf{x}_i,\mathbf{z}_i)\hat{\boldsymbol{\gamma}}_A$.

Theorem 3.4.5 (Convergence Rate of Oracle Estimator) Suppose Assumption 1–5, 7 and 8. Then,

$$\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_o\|=O_p(N^{-\frac{1}{2}})\tag{3.28}$$

$$N^{-1}\sum_{i=1}^N[\hat{g}(\mathbf{x}_i,\mathbf{z}_i)-g_o(\mathbf{x}_i,\mathbf{z}_i)]^2=O_p\left(N^{-1}q_N\right)\tag{3.29}$$

Proof. See Appendix C.1.3. ■

Theorem 3.4.6 (Asymptotic Normality of Oracle Estimator) Suppose Assumption 1–5, 7 and 8. Then,

$$\sqrt{N}\Sigma_N^{*-\frac{1}{2}}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_o)\overset{d}{\to} N\left(0,I\right) \tag{3.30}$$

where $\Sigma_N^* = \mathbf{K}_N^{*-1}\mathbf{S}_N^*\mathbf{K}_N^{*-1}$ with $\mathbf{K}_N^* = N^{-1}\boldsymbol{\Delta}'B_N\boldsymbol{\Delta}, \mathbf{S}_N^* = N^{-1}\sum_{i=1}^N \boldsymbol{\Delta}_i'\Psi_\tau\left(\boldsymbol{\varepsilon}_i\right)\Psi_\tau\left(\boldsymbol{\varepsilon}_i\right)'\boldsymbol{\Delta}_i$.

Proof. The result follows from Lemmas C.1.9 and C.1.11 in Appendix C.1.3. ■

The convergence rate of $\hat{\boldsymbol{\beta}}$ is the parametric rate as in the typical partially linear semiparametric models. On the other hand, $\hat{g}\left(\mathbf{x}_i,\mathbf{z}_i\right)$ has a convergence rate of $N^{-1}q_N$ which depends on the sparsity of a true submodel given the dictionary variable sequence. One obvious implication is that the performance of an oracle estimator will depend on the choice of the dictionary variable sequence. Each of Theorem 3.4.4 and 3.4.6 shows an asymptotic distribution result that can be used to approximate the distribution of the oracle estimator in a finite sample. Note that the sample analogue estimator of $N^{-1}\Sigma_N$ and $N^{-1}\Sigma_N^*$ coincide for approximating $\hat{V}(\hat{\boldsymbol{\beta}})$. The computation of variance estimators is presented in Subsubsection 3.4.2.2.

Since true nonzero coefficients are unknown, the sparse set is estimated by penalizing coefficients of all dictionary variable in the sample optimization problem. In multiple contexts, the penalized estimators using nonconvex penalty functions such as SCAD (Fan and Li, 2001)

$$p_\lambda\left(|\gamma|\right)=\begin{cases}\lambda\left|\gamma\right| & 0\le\gamma<\lambda \\ \frac{a\lambda|\gamma|-\left(\gamma^2+\lambda^2\right)/2}{(a-1)} & \lambda\le|\gamma|\le a\lambda \\ \frac{(a+1)\lambda^2}{2} & a\lambda<|\gamma|\end{cases}\quad\text{for some } a>2 \tag{3.31}$$

and MCP (Zhang, 2010)

$$p_\lambda\left(|\gamma|\right)=\begin{cases}\lambda(|\gamma|-\frac{\gamma^2}{2a\lambda}) & 0\le|\gamma|\le a\lambda \\ \frac{a\lambda^2}{2} & a\lambda\le|\gamma|\end{cases}\quad\text{for some } a>1 \tag{3.32}$$

are shown to yield the oracle estimator among the local minima of a penalized objective function with probability tending to one. Such feature is called “a (weak) oracle property” of the nonconvex penalized estimator. To present the oracle property for the current model setting, the penalized estimator is defined as follows.

Definition 3.4.7 (Penalized Estimator)

$$(\hat{\boldsymbol{\beta}}_\lambda, \hat{\boldsymbol{\gamma}}_\lambda) = \underset{(\beta,\boldsymbol{\gamma})}{\arg\min} \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \rho_\tau \left(y_{it} - \mathbf{w}_{it}\boldsymbol{\beta} - \boldsymbol{\pi}\left(\mathbf{x}_i, \mathbf{z}_i\right)\boldsymbol{\gamma} \right) + \sum_{j=1}^{p_N} p_\lambda(|\gamma_j|) \tag{3.33}$$

where $p_\lambda(\cdot)$ is either a SCAD or MCP penalty function.

The oracle property is shown with one additional condition on true sparse coefficients. Assumption 9 below is often called a 'beta-min' condition, which basically assumes that the minimum maginitude of nonzero coefficients in the oracle model is sufficiently large. In our context, the lower bound for the coefficient magnitude can be understood as a truncation cut-off for approximating surrogate function $\boldsymbol{\pi}\left(\mathbf{x}_i, \mathbf{z}_i\right)\boldsymbol{\gamma}$ given a sequence of dictionary variables.

Assumption 9 (Nonzero Coefficients) There exist positve constants C_4 and C_5 such that $C_3 < C_4 < 1$ and $N^{(1-C_4)/2} \min_{1 \le j \le q_N} |\gamma_{oj}| \ge C_5$.

Theorem 3.4.8 (Oracle Property of Penalized Estimator) Suppose Assumption 1–5 and 9 together with Assumption 6 or with Assumption 7 and 8. If $\lambda = o\left(N^{-(1-C_4)/2}\right)$, $N^{-1/2}q_N^{1/2} = o(\lambda)$, and $\log(p_N) = o(N\lambda^2)$, then

$$\lim_{N\to\infty} P((\hat{\boldsymbol{\beta}}_\lambda, \hat{\boldsymbol{\gamma}}_\lambda) \in \varepsilon_N(\lambda)) = 1 \tag{3.34}$$

where $\varepsilon_N(\lambda)$ is the set of local minima of objective function in (3.33).

Proof. See Appendix. ■

The rate condition on q_N and λ in Theorem 3.4.8 is weaker than the one given in Sherwood and Wang (2016). In turn, a necessary requirement on C_4 in the beta-min condition is also weaker.

3.4.2.1 Choice of Thresholding Parameter λ

Lee, Noh and Park (2014; LNP) recently proposed a modified Bayesian Information Criterion for linear quantile regression with cross-sectional data when the dimension of the dictionary variables diverges and the dimension of the true model is a constant. Sherwood and Wang (2016) take LNP's criterion for the case where the dimension of the true model may diverge. Its pooled information version for panel data in the current setting can be considered as follows

$$\text{QBICL}\left(\lambda\right)=2TN\log\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\rho_{\tau}(e_{it}(\hat{\boldsymbol{\beta}}_{\lambda},\hat{\boldsymbol{\gamma}}_{\lambda}))\right)+S_{\lambda}C_{N}\log\left(TN\right) \tag{3.35}$$

where $e_{it}\left(\boldsymbol{\beta},\boldsymbol{\gamma}\right)=y_{it}-\mathbf{w}_{it}\boldsymbol{\beta}-\boldsymbol{\pi}\left(\mathbf{x}_{i},\mathbf{z}_{i}\right)\boldsymbol{\gamma}$, S_{λ} is the degree of freedom of the fitted model and C_{N} is chosen as $\log\left(p_{N}\right)$ in Sherwood and Wang (2016).

Note that the goodness of fit measure in (3.35) is derived from the quasi-likelihood of asymmetric laplace distribution with scaling parameter σ. (See LNP for details.) When $\sigma=1$ is imposed, the resulting measure coincides with the conventional check loss function without logarithm and TN-scaling. Then, an alternative form of high-dimensional BIC can be written as

$$\text{BICL}\left(\lambda\right)=2\sum_{i=1}^{N}\sum_{t=1}^{T}\rho_{\tau}(e_{it}(\hat{\boldsymbol{\beta}}_{\lambda},\hat{\boldsymbol{\gamma}}_{\lambda}))+S_{\lambda}C_{N}\log\left(TN\right). \tag{3.36}$$

To take into account the clustered information of panel data, it is useful to think of clustering as a kind of misspecification problem in quasi-likelihood. In this perspective, generalized BIC (GBIC) and GBIC$_{p}$ studied by Lv and Liu (2014; LL) can be considered. They explicitly incorporate model misspecification using a second-order term in the asymptotic expansion of the Bayesian principle under generalized linear model settings. The final result is claimed to be general enough to be applied to other contexts. Adding the second-order term of GBIC studied by LL to (3.35), we have

$$\text{GQBICL}\left(\lambda\right)=2TN\log\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\rho_{\tau}(e_{it}(\hat{\boldsymbol{\beta}}_{\lambda},\hat{\boldsymbol{\gamma}}_{\lambda}))\right)+S_{\lambda}C_{N}\log\left(TN\right)-\log\det\left(\boldsymbol{H}_{\lambda,N}\right) \tag{3.37}$$

where $\boldsymbol{H}_{\lambda,N}=\hat{K}_{\lambda,N}^{-1}\hat{S}_{\lambda,N}$ is a covariance contrast matrix evaluated at $(\hat{\boldsymbol{\beta}}_{\lambda},\hat{\boldsymbol{\gamma}}_{\lambda})$. Note that the second-order term can be negative. The same modification can be done for BICL. For further details about GBIC and GBIC$_{p}$, see LL (2014). Note that if the correction term $\log\det\left(\boldsymbol{H}_{\lambda,N}\right)$ is asymptotically bounded, then the first two terms in the information criterion will be dominant as N tends to infinity. Thus, if GBIC is indeed a valid criterion for selection consistency, then regular BICs without the correction term must be valid as well in such cases.

3.4.2.2 Computation of Variance Estimators

The sample analogue estimators for the sandwich form of $\mathbf{K}_N^{-1}\mathbf{S}_N\mathbf{K}_N^{-1}$ and $\mathbf{K}_N^{*-1}\mathbf{S}_N^{*}\mathbf{K}_N^{*-1}$ are computed using the set of selected variables, $\hat{A}(\lambda)$, given the penalized estimator $(\hat{\boldsymbol{\beta}}_\lambda, \hat{\boldsymbol{\gamma}}_\lambda)$. This is mainly justfied by the fact that (i) the penalized estimator has an oracle property, and that (ii) any submodel constitutes an approximation of the true model. Here, estimators are constructed following the cluster-robust variance estimator proposed by Wooldridge (2010). Let M^+ denote the Moore-Penrose generalized inverse of M. First, the residual $\hat{e}_{it}$ is simply computed by plugging in estimates $(\hat{\boldsymbol{\beta}}_\lambda, \hat{\boldsymbol{\gamma}}_\lambda)$ in the formula:

$$\hat{e}_{it} = y_{it} - \mathbf{w}_{it}\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\pi}_{\hat{A}(\lambda)}(\mathbf{x}_i, \mathbf{z}_i)\,\hat{\boldsymbol{\gamma}}_\lambda. \tag{3.38}$$

The π_A-projected-out regressor $\hat{\Delta}_{it}$ can be estimated as, for a sequence h_N tending to 0,

$$\hat{\Delta}_{it} = \mathbf{w}_{it} - \boldsymbol{\pi}_{\hat{A},i}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} 1_{[|\hat{e}_{it}|\leq h_N]}\boldsymbol{\pi}'_{\hat{A},i}\boldsymbol{\pi}_{\hat{A},i}\right)^{+}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} 1_{[|\hat{e}_{it}|\leq h_N]}\boldsymbol{\pi}'_{\hat{A},i}\mathbf{w}_{it}\right) \tag{3.39}$$

where the conditional density $f_{it}(0)$ is approximated via uniform kernel. Then, the sample analogue estimator of Σ_N^* can be written as $\hat{\Sigma}_N^* = \hat{\mathbf{K}}_N^{*+}\hat{\mathbf{S}}_N^*\hat{\mathbf{K}}_N^{*+}$ with

$$\hat{\mathbf{K}}_N^* = \frac{1}{2Nh_N}\sum_{i=1}^{N}\sum_{t=1}^{T} 1_{[|\hat{e}_{it}|\leq h_N]}\hat{\Delta}'_{it}\hat{\Delta}_{it} \tag{3.40}$$

$$\hat{\mathbf{S}}_N^* = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{t'=1}^{T}\psi_\tau\left(\hat{e}_{it'}\right)\psi_\tau\left(\hat{e}_{it}\right)\hat{\Delta}'_{it'}\hat{\Delta}_{it}. \tag{3.41}$$

The generalized inverse is used instead of the regular inverse since $\boldsymbol{\pi}_A$ may contain a set of linearly dependent variables given sample size N and threshold parameter λ. Note that the Moore-Penrose inverse coincides with the ordinary inverse whenever the operated matrix is invertible. The estimator $\hat{\Sigma}_N = \hat{\mathbf{K}}_N^+\hat{\mathbf{S}}_N\hat{\mathbf{K}}_N^+$ can be similarly computed by replacing $\hat{\Delta}_{it}$ with $\tilde{w}_{it}$ in (3.40) and (3.41). It can be shown that the estimated variances of $\hat{\boldsymbol{\beta}}$ based on $\hat{\Sigma}_N$ and $\hat{\Sigma}_N^*$ are numerically equivalent if $\hat{\mathbf{K}}_N$ is invertible. For the choice of sequence h_N, see Perente and Santos Silva (2010) for example.

3.5 Monte Carlo Simulation

A set of Monte Carlo simulations is conducted to study the selection performance and estimator performance on simple location shift and location-scale shift models. With a 3-period panel structure, 5 specifications are considered:

$$\begin{aligned}
\text{DGP 1} : y_{it} &= x_{it1} + x_{it2} + x_{i11} + x_{i12} + x_{i13} + x_{i21} + x_{i22} + x_{i23} + u_{it} \\
\text{DGP 2} : y_{it} &= (8 + x_{it1} + x_{it2} + x_{i11} + x_{i12} + x_{i13} + x_{i21} + x_{i22} + x_{i23}) (u_{it} + 1) \\
\text{DGP 3} : y_{it} &= 14 + x_{it1} + x_{it2} + \sum_{k \in K} (x_{i11}^k + x_{i12}^k + x_{i13}^k + x_{i21}^k + x_{i22}^k + x_{i23}^k) + u_{it} \\
\text{DGP 4} : y_{it} &= \{14 + x_{it1} + x_{it2} + \sum_{k \in K} (x_{i11}^k + x_{i12}^k + x_{i13}^k + x_{i21}^k + x_{i22}^k + x_{i23}^k)\} (u_{it} + 1) \\
\text{DGP 5} : y_{it} &= \{14 + x_{it1} + x_{it2} + \sum_{k \in K} (x_{i11}^k + 2x_{i12}^k + x_{i21}^k + 2x_{i22}^k)\} (u_{it} + 1)
\end{aligned}$$

where $T = 3,\ N = 300$ or 1000, $x_{1t} \sim U(-1, 1)\,,\ x_{2t} \sim U(-1, 1)\,,\ u_{it} \sim U(0, 1)\,,\ K = \{1, 2, 7\}$. DGP 1, 3 and 2, 4, 5 are location shift and location-scale shift models, respectively. Note that location-scale shift models have heteroskedastic regression error terms. DGP 1 and 2 impose a Chamberlain specification while DGP 3, 4 and 5 introduce nonlinearity or different coefficients on correlated effect terms across time periods. Note that the rescaled true parameters for higher-order polynomial terms are smaller since all dictionary variables are rescaled to have unit sample variance in estimation. For example, a normalized 1st and 7th order polynomial term x_{i11}, x_{i11}^7 in DGP 3 have true parameter values of about .58 and .26, respectively. In turn, higher-order terms are harder to be selected as relevant variables at given sample sizes. The number of simulated draws is 1000.

Table 3.1-3.2 and Table A.1-A.11 in Appendix C.2 contain the results. Along with QBICL and BICL, their non-high dimensional versions, QBIC and BIC are also considered. AIC1 and AIC2 are AIC counterparts of QBIC and BIC that use different goodness of fit measures. "p_N" denotes the number of dictionary variables, and "q_o" denotes the number of terms in the true models. "TV" and "FV" are defined as the average number of true and false coefficients selected, respectively. "true" means the true model hit rate.

Table 3.1 Selection Performance, DGP 1 and 2

					$\tau = 0.1$			$\tau = 0.5$			$\tau = 0.9$		
Method	p_N	q_o	N	IC	TV	FV	True	TV	FV	True	TV	FV	True
gMund	136	2	300	QBIC	2.00	0.13	0.89	2.00	0.17	0.88	2.00	0.17	0.87
gMund	136	2	300	QBICL	2.00	0.00	1.00	2.00	0.00	1.00	2.00	0.00	1.00
gMund	136	2	300	BIC	2.00	0.00	1.00	2.00	0.00	1.00	2.00	0.00	1.00
gMund	136	2	300	BICL	2.00	0.00	1.00	2.00	0.00	1.00	2.00	0.00	1.00
gMund	136	2	300	AIC1	2.00	2.18	0.35	2.00	3.78	0.29	2.00	4.05	0.23
gMund	136	2	300	AIC2	2.00	0.00	1.00	2.00	0.01	0.99	2.00	0.00	1.00
gMund	136	2	1000	QBIC	2.00	0.08	0.94	2.00	0.10	0.93	2.00	0.11	0.92
gMund	136	2	1000	QBICL	2.00	0.00	1.00	2.00	0.00	1.00	2.00	0.00	1.00
gMund	136	2	1000	BIC	2.00	0.00	1.00	2.00	0.00	1.00	2.00	0.00	1.00
gMund	136	2	1000	BICL	2.00	0.00	1.00	2.00	0.00	1.00	2.00	0.00	1.00
gMund	136	2	1000	AIC1	2.00	2.57	0.39	2.00	7.22	0.23	2.00	6.34	0.24
gMund	136	2	1000	AIC2	2.00	0.00	1.00	2.00	0.01	0.99	2.00	0.00	1.00
gCham	102	6	300	QBIC	6.00	0.18	0.87	6.00	0.19	0.85	6.00	0.22	0.85
gCham	102	6	300	QBICL	6.00	0.00	1.00	6.00	0.00	1.00	6.00	0.00	1.00
gCham	102	6	300	BIC	6.00	0.00	1.00	6.00	0.00	1.00	6.00	0.00	1.00
gCham	102	6	300	BICL	6.00	0.00	1.00	6.00	0.00	1.00	6.00	0.00	1.00
gCham	102	6	300	AIC1	6.00	5.71	0.24	6.00	3.77	0.27	6.00	4.36	0.27
gCham	102	6	300	AIC2	6.00	0.00	1.00	6.00	0.01	0.99	6.00	0.00	1.00
gCham	102	6	1000	QBIC	6.00	0.12	0.92	6.00	0.13	0.91	6.00	0.16	0.90
gCham	102	6	1000	QBICL	6.00	0.00	1.00	6.00	0.00	1.00	6.00	0.00	1.00
gCham	102	6	1000	BIC	6.00	0.00	1.00	6.00	0.00	1.00	6.00	0.00	1.00
gCham	102	6	1000	BICL	6.00	0.00	1.00	6.00	0.00	1.00	6.00	0.00	1.00
gCham	102	6	1000	AIC1	6.00	7.49	0.23	6.00	7.42	0.20	6.00	7.87	0.21
gCham	102	6	1000	AIC2	6.00	0.00	1.00	6.00	0.01	0.99	6.00	0.00	1.00

					$\tau = 0.1$			$\tau = 0.5$			$\tau = 0.9$		
Method	p_N	q_o	N	IC	TV	FV	True	TV	FV	True	TV	FV	True
gMund	136	2	300	QBIC	2.00	0.85	0.55	2.00	1.16	0.40	2.00	1.16	0.44
gMund	136	2	300	QBICL	2.00	0.00	1.00	1.98	0.02	0.98	2.00	0.01	1.00
gMund	136	2	300	BIC	2.00	0.02	0.98	2.00	1.15	0.41	2.00	0.02	0.98
gMund	136	2	300	BICL	2.00	0.00	1.00	1.98	0.02	0.98	2.00	0.01	1.00
gMund	136	2	300	AIC1	2.00	6.63	0.03	2.00	6.75	0.00	2.00	7.91	0.00
gMund	136	2	300	AIC2	2.00	1.42	0.39	2.00	6.65	0.00	2.00	1.74	0.30
gMund	136	2	1000	QBIC	2.00	0.72	0.59	2.00	0.86	0.51	2.00	0.80	0.54
gMund	136	2	1000	QBICL	2.00	0.00	1.00	2.00	0.00	1.00	2.00	0.00	1.00
gMund	136	2	1000	BIC	2.00	0.01	0.99	2.00	0.85	0.51	2.00	0.01	0.99
gMund	136	2	1000	BICL	2.00	0.00	1.00	2.00	0.00	1.00	2.00	0.00	1.00
gMund	136	2	1000	AIC1	2.00	7.59	0.00	2.00	8.03	0.01	2.00	8.90	0.00
gMund	136	2	1000	AIC2	2.00	1.68	0.31	2.00	7.98	0.01	2.00	1.89	0.26
gCham	102	6	300	QBIC	5.85	0.80	0.55	5.61	1.01	0.37	6.00	0.82	0.54
gCham	102	6	300	QBICL	5.74	0.27	0.77	5.46	0.49	0.57	5.94	0.14	0.87
gCham	102	6	300	BIC	5.75	0.28	0.77	5.61	0.98	0.38	5.98	0.15	0.87
gCham	102	6	300	BICL	0.86	0.11	0.06	5.46	0.49	0.57	5.86	0.14	0.87
gCham	102	6	300	AIC1	5.96	7.33	0.02	5.95	6.24	0.01	6.00	6.40	0.01
gCham	102	6	300	AIC2	5.88	1.24	0.44	5.95	6.07	0.01	6.00	1.31	0.40
gCham	102	6	1000	QBIC	6.00	0.46	0.71	6.00	0.48	0.68	6.00	0.51	0.70
gCham	102	6	1000	QBICL	5.99	0.01	0.98	5.96	0.04	0.96	6.00	0.01	0.99
gCham	102	6	1000	BIC	6.00	0.02	0.98	6.00	0.48	0.69	6.00	0.02	0.98
gCham	102	6	1000	BICL	5.99	0.01	0.98	5.96	0.04	0.96	6.00	0.01	0.99
gCham	102	6	1000	AIC1	6.00	8.95	0.01	6.00	7.59	0.01	6.00	8.35	0.01
gCham	102	6	1000	AIC2	6.00	1.31	0.44	6.00	7.55	0.01	6.00	1.36	0.42

Table 3.2 Estimator performance, DGP 1 and 2, $\beta 1$

			$\tau = 0.1$			$\tau = 0.5$			$\tau = 0.9$		
Method	N	IC	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
gMund	300	QBIC	0.0000	0.0197	0.0197	0.0009	0.0345	0.0345	-0.0000	0.0215	0.0215
gMund	300	QBICL	-0.0001	0.0197	0.0197	0.0008	0.0345	0.0345	-0.0000	0.0215	0.0215
gMund	300	BIC	-0.0001	0.0197	0.0197	0.0008	0.0345	0.0345	-0.0000	0.0215	0.0215
gMund	300	BICL	-0.0001	0.0197	0.0197	0.0008	0.0345	0.0345	-0.0000	0.0215	0.0215
gMund	300	AIC1	-0.0001	0.0201	0.0200	0.0021	0.0346	0.0346	-0.0002	0.0215	0.0215
gMund	300	AIC2	-0.0001	0.0197	0.0197	0.0008	0.0345	0.0345	-0.0000	0.0215	0.0215
gMund	1000	QBIC	-0.0002	0.0119	0.0119	0.0013	0.0191	0.0192	-0.0001	0.0114	0.0114
gMund	1000	QBICL	-0.0001	0.0119	0.0119	0.0013	0.0191	0.0191	-0.0001	0.0114	0.0114
gMund	1000	BIC	-0.0001	0.0119	0.0119	0.0013	0.0191	0.0191	-0.0001	0.0114	0.0114
gMund	1000	BICL	-0.0001	0.0119	0.0119	0.0013	0.0191	0.0191	-0.0001	0.0114	0.0114
gMund	1000	AIC1	-0.0001	0.0118	0.0118	0.0015	0.0194	0.0195	-0.0001	0.0114	0.0114
gMund	1000	AIC2	-0.0001	0.0119	0.0119	0.0013	0.0191	0.0191	-0.0001	0.0114	0.0114
gCham	300	QBIC	0.0007	0.0211	0.0211	0.0042	0.0358	0.0360	-0.0005	0.0212	0.0212
gCham	300	QBICL	0.0005	0.0212	0.0212	0.0040	0.0358	0.0360	-0.0006	0.0208	0.0208
gCham	300	BIC	0.0005	0.0212	0.0212	0.0040	0.0358	0.0360	-0.0006	0.0208	0.0208
gCham	300	BICL	0.0005	0.0212	0.0212	0.0040	0.0358	0.0360	-0.0006	0.0208	0.0208
gCham	300	AIC1	0.0014	0.0208	0.0209	0.0058	0.0355	0.0359	-0.0002	0.0212	0.0212
gCham	300	AIC2	0.0005	0.0212	0.0212	0.0039	0.0359	0.0361	-0.0006	0.0208	0.0208
gCham	1000	QBIC	0.0004	0.0116	0.0116	-0.0004	0.0186	0.0186	0.0002	0.0117	0.0117
gCham	1000	QBICL	0.0003	0.0115	0.0115	-0.0005	0.0186	0.0186	0.0002	0.0117	0.0117
gCham	1000	BIC	0.0003	0.0115	0.0115	-0.0005	0.0186	0.0186	0.0002	0.0117	0.0117
gCham	1000	BICL	0.0003	0.0115	0.0115	-0.0005	0.0186	0.0186	0.0002	0.0117	0.0117
gCham	1000	AIC1	0.0004	0.0115	0.0115	-0.0001	0.0186	0.0186	0.0005	0.0117	0.0117
gCham	1000	AIC2	0.0003	0.0115	0.0115	-0.0005	0.0186	0.0186	0.0002	0.0117	0.0117

			$\tau = 0.1$			$\tau = 0.5$			$\tau = 0.9$		
Method	N	IC	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
gMund	300	QBIC	0.0073	0.1551	0.1552	0.0048	0.2746	0.2745	0.0028	0.1648	0.1648
gMund	300	QBICL	0.0066	0.1565	0.1565	0.0064	0.2778	0.2777	0.0013	0.1647	0.1646
gMund	300	BIC	0.0066	0.1562	0.1562	0.0049	0.2747	0.2746	0.0017	0.1651	0.1650
gMund	300	BICL	0.0066	0.1565	0.1565	0.0064	0.2778	0.2777	0.0013	0.1647	0.1646
gMund	300	AIC1	0.0160	0.1559	0.1567	-0.0009	0.2657	0.2656	-0.0033	0.1612	0.1612
gMund	300	AIC2	0.0073	0.1551	0.1552	-0.0007	0.2658	0.2657	0.0028	0.1650	0.1649
gMund	1000	QBIC	0.0066	0.0877	0.0879	-0.0051	0.1479	0.1479	0.0063	0.0854	0.0856
gMund	1000	QBICL	0.0056	0.0877	0.0878	-0.0053	0.1467	0.1467	0.0048	0.0858	0.0859
gMund	1000	BIC	0.0056	0.0878	0.0879	-0.0053	0.1478	0.1478	0.0048	0.0858	0.0859
gMund	1000	BICL	0.0056	0.0877	0.0878	-0.0053	0.1467	0.1467	0.0048	0.0858	0.0859
gMund	1000	AIC1	0.0103	0.0893	0.0898	-0.0026	0.1479	0.1478	0.0036	0.0865	0.0865
gMund	1000	AIC2	0.0071	0.0881	0.0883	-0.0028	0.1480	0.1479	0.0054	0.0851	0.0852
gCham	300	QBIC	0.0105	0.1679	0.1682	0.0367	0.2562	0.2587	-0.0015	0.1625	0.1624
gCham	300	QBICL	0.0148	0.1668	0.1674	0.0503	0.2738	0.2783	0.0027	0.1627	0.1627
gCham	300	BIC	0.0141	0.1669	0.1674	0.0367	0.2565	0.2589	0.0003	0.1626	0.1625
gCham	300	BICL	1.0494	0.4584	1.1450	0.0508	0.2745	0.2790	0.0077	0.1622	0.1623
gCham	300	AIC1	0.0135	0.1640	0.1645	0.0289	0.2590	0.2605	-0.0070	0.1570	0.1571
gCham	300	AIC2	0.0101	0.1660	0.1663	0.0284	0.2591	0.2605	-0.0022	0.1611	0.1610
gCham	1000	QBIC	0.0033	0.0882	0.0883	-0.0004	0.1477	0.1476	-0.0006	0.0880	0.0880
gCham	1000	QBICL	0.0040	0.0879	0.0879	0.0017	0.1490	0.1489	-0.0002	0.0891	0.0891
gCham	1000	BIC	0.0038	0.0878	0.0878	-0.0004	0.1482	0.1481	-0.0002	0.0891	0.0891
gCham	1000	BICL	0.0042	0.0879	0.0880	0.0017	0.1490	0.1489	-0.0002	0.0891	0.0891
gCham	1000	AIC1	0.0047	0.0870	0.0871	-0.0008	0.1487	0.1486	-0.0035	0.0880	0.0880
gCham	1000	AIC2	0.0039	0.0875	0.0875	-0.0006	0.1488	0.1487	-0.0007	0.0878	0.0878

There are some useful findings to be mentioned. First, the root mean squared error (RMSE) of β estimates decreases and TV increases as the sample size increases for all DGPs, quantiles and information criteria considered. Second, FV decreases as the sample size increases in DGP 1, 2 and 3 with non-high dimensional BIC type criteria. With AIC or in other models, FV may increase. In DGP 4 and 5, FV increases for BIC type criteria but not as much as AICs. Third, TV seems to be the key element that determines the estimator performance. Neither FV nor the true hit rate seems to matter much. This can be easily seen by comparing AICs with BICs. AIC typically involves a much higher FV and lower true hit rate but often shows the smallest bias, SD and RMSE. Fourth, the estimators using a generalized Mundlak form and generalized Chamberlain form can outperform the others. When the coefficients on the correlated effect terms across time periods are constant as in DGP 1, 2, 3, and 4, the Mundlak form yields a smaller number of nonzero terms to be selected and the corresponding estimator often has better performance. But when the coefficients are different across time as in DGP 5, the generalized Chamberlain form can have a more sparse selected submodel than the generalized Mundlak form and the corresponding estimator often outperforms.

3.6 Application: The Effect of Smoking on Birth Outcomes

In this section, the proposed estimator is applied to an empirical example of birth weight analysis. The data in use is the matched panel[1] #3 of Abrevaya (2006) where mean regression analysis was done accounting for unobserved individual moms' heterogeneity. First, the median regression with a correlated effect shows convincing evidence that the correlated effect estimator works well as intended. Second, the corresponding other quantile regression results show that for lower quantiles, the impact of smoking on birth weight is smaller in terms of absolute magnitute but can be larger relative to fitted quantile birth weights. Third, some computational issues are

[1] The data do not have a panel structure in the strict sense since each mom is observed at a different time point when she gave a birth. Still, the data set is a clustered in general and the proposed method is valid as long as the set of assumptions hold.

reported regarding optimization of the nonconvex objective function.

The matched panel data #3 of Abrevaya (2006) contains information on 129,569 two-birth moms and 12,360 three-birth moms. Note that in these data, the quantile regression with individual fixed effects will cost an additional 141,929 dummies. One of the main benefits of the correlated effect estimator is to reduce the number of additional terms while individual heterogeneity is treated well. The results below show that less than 400 additional terms are spent to obtain reasonable estimates. The number of additional terms is less than 0.3‰ that of fixed effect estimator.

The structural equation is taken from Abrevaya (2006) as

$$\begin{aligned} BW_{ib} &= smoke_{ib} + male_{ib} + age_{ib} + age_{ib}^2 + adeqcode2_{ib} + adeqcode3_{ib} \\ &\quad + novisit_{ib} + pretri2_{ib} + pretri3_{ib} + \_Inlbnl\_1_{ib} + \cdots + \_Inlbnl\_15_{ib} \\ &\quad + \_Iyear\_1_{ib} + \cdots + \_Iyear\_8_{ib} + const + \varepsilon_{ib} \end{aligned} \tag{3.42}$$

where i is an individual mother index, b is an observed birth index, "adeqcode#" is the Kessner index of #, "novisit" is an indicator of no prenatal visit during pregnancy, "pretri#" is an indicator of the first prenatal visit in #th trimester, and other terms are live birth order and year effects. The observed birth index b corresponds to time index t in our setting. All right-hand-side variables in (3.42) constitute $\mathbf{x}_{ib}$ following the previous notation.

Besides $\mathbf{x}_{ib}$, there are several within-group-constant variables, $\mathbf{z}_i$ that are used to construct the correlated effects: binary dummies for high school graduate, some college experience, college graduate, marital status, being black and state of residence. Since all variables except "age" are treated as discrete, the sparsity assumption is essentially imposed on the "age" component of the correlated effect. For the sieve approximation, polynomials of order up to 10th are used as default.

To account for potential endogeneity due to observability, the selection indicator will be used to construct the correlated effects. Basically, there are two observed patterns: 2-birth mothers and 3-birth mothers. Then, the vector of the selection indicator can be written as $\mathbf{s}_i = (1, 1, 0)$ or $(1, 1, 1)$. For notational convenience, denote $\mathbf{s}_i^{[2]} = (1, 1, 0)$ and $\mathbf{s}_i^{[3]} = (1, 1, 1)$. Then, the

Table 3.3 Birthweight, mean and median regression, all moms (unit:grams)

	Mean		Median			
	OLS	FE	Pooled	CE1	CE2	CE3
Smoke	-243.27	-144.04	-238.49	-138.26	-138.55	-140.34
	(3.20)	(4.75)	(3.79)	(6.31)	(6.39)	(6.45)
Male	126.70	133.58	131.27	138.87	139.38	139.45
	(1.88)	(2.08)	(2.10)	(2.51)	(2.54)	(2.52)
Age	7.06	-15.98	2.59	-13.37	-8.32	-8.74
	(1.77)	(3.96)	(2.10)	(5.38)	(27.04)	(4.57)
Age^2	-0.12	0.32	-0.04	0.35	0.26	2.75
	(0.03)	(0.05)	(0.04)	(0.07)	(0.29)	(0.07)
High-school graduate	60.52		64.19			
	(4.12)		(4.98)			
Some College	91.34		96.65			
	(4.52)		(5.55)			
College Graduate	100.89		102.84			
	(4.73)		(5.79)			
Married	64.43		55.23			
	(3.65)		(4.32)			
Black	-252.04		-239.28			
	(4.36)		(5.07)			
Kessner index = 2	-100.93	-84.43	-81.71	-79.17	-74.12	-69.96
	(4.19)	(4.45)	(4.56)	(5.66)	(6.14)	(5.55)
Kessner index = 3	-176.48	-143.91	-149.85	-163.42	-154.35	-150.94
	(10.20)	(10.28)	(12.48)	(15.67)	(15.52)	(15.50)
No prenatal visit	-26.49	-42.35	7.87	-32.02	-47.25	-52.70
	(18.00)	(16.57)	(21.29)	(24.99)	(27.03)	(26.88)
First prenatal visit in 2nd trimester	89.12	66.56	72.21	67.38	62.80	58.66
	(4.96)	(5.27)	(5.40)	(6.73)	(9.27)	(6.67)
First prenatal visit in 3rd trimester	154.66	111.90	119.48	109.92	111.45	118.90
	(12.03)	(12.49)	(14.34)	(18.82)	(22.84)	(18.41)
Information Criterion	-	-	-	BIC	BIC	BICL
# of Dictionary Var.	-	-	-	301	451	301
# of Selected Var.	-	-	-	298	300	169

conditional quantiles with the generalized Chamberlain device can be succinctly written as

$$Q_\tau\left(s_{ib}y_{ib}|\left\{\mathbf{x}_{it}\right\}_{t=1}^{2},\mathbf{z}_i,\mathbf{s}_i=\mathbf{s}_i^{[2]}\right)=s_{ib}\mathbf{x}_{ib}\beta+s_{ib}g_2\left(\mathbf{x}_{i1},\mathbf{x}_{i2},\mathbf{z}_i\right)+s_{ib}k_{2b} \tag{3.43}$$

$$Q_\tau\left(s_{ib}y_{ib}|\left\{\mathbf{x}_{it}\right\}_{t=1}^{3},\mathbf{z}_i,\mathbf{s}_i=\mathbf{s}_i^{[3]}\right)=s_{ib}\mathbf{x}_{ib}\beta+s_{ib}g_3\left(\left\{\mathbf{x}_{it}\right\}_{t=1}^{3},\mathbf{z}_i\right)+s_{ib}k_{3b} \tag{3.44}$$

for some g_2, g_3, k_{2b}s and k_{3b}s. Assuming additivity of gs, following transformation is considered

$$g_2\left(\mathbf{x}_{i1},\mathbf{x}_{i2},\mathbf{z}_i\right)+k_{2b}=g_{2,0}+\sum_{b=1}^{3}\sum_{k=1}^{K_1}g_{2bk}^{x}\left(x_{ibk}\right)+\sum_{k=1}^{K_2}g_{2k}^{z}\left(z_{ik}\right)+k_{2b} \tag{3.45}$$

$$g_3\left(\left\{\mathbf{x}_{it}\right\}_{t=1}^{3},\mathbf{z}_i\right)+k_{3b}=g_{2,0}+\sum_{b=1}^{3}\sum_{k=1}^{K_1}g_{2bk}^{x}\left(x_{ibk}\right)+\sum_{k=1}^{K_2}g_{2k}^{z}\left(z_{ik}\right)+k_{2b} \tag{3.46}$$

$$+h_{3,0}+\sum_{b=1}^{3}\sum_{k=1}^{K_1}h_{bk}^{x}\left(x_{ibk}\right)+\sum_{k=1}^{K_2}h_{2k}^{z}\left(z_{ik}\right)+l_b \tag{3.47}$$

where $g_{2bk}^{x}\left(x_{ibk}\right)=0$ for $b=3$, $h_{3,0}=g_{3,0}-g_{2,0}$, $h_{bk}^{x}\left(x_{ibk}\right)=g_{3bk}^{x}\left(x_{ibk}\right)-g_{2bk}^{x}\left(x_{ibk}\right)$, $h_{2k}^{z}\left(z_{ik}\right)=g_{3k}^{z}\left(z_{ik}\right)-g_{2k}^{z}\left(z_{ik}\right)$ and $l_b=k_{3b}-k_{2b}$. In estimation, the interaction of the 3-birth mom dummy and approximating terms for g_{2bk}^{x} and g_{2k}^{z} $(b=1,2)$ are included. The constant term $g_{2,0}$, $h_{3,0}$ and time effects k_{2b}, l_b are also included but not penalized. l_2 is excluded to avoid multicollinearity. If there is no systematic difference in g components between the two-birth mothers and three-birth mothers, then the corresponding h compoents will be zeros and there will be fewer terms selected in the final estimates.

In Table 3.3, the OLS estimator, FE estimator, pooled median regression estimator and CE estimators are compared. BIC and BICL were used to choose the threshold parameters of the CE1/CE2 and CE3 estimates, respectively where the candidate threshold parameters were chosen to be 50 equi-spaced points between 0 and 0.01. The CE2 estimator uses polynomials of order upto 40th. The "rqPen" and "pracma" packages for R were used for computing penalized estimates and Moore-Penrose inverse matrices, respectively. As noted by Abrevaya (2006), the FE coefficient estimate on the “smoke” variable has approximately a 100g lower magnitude than the OLS estimate, which is consistent with the basic omitted variables story. The CE coefficient estimates on the “smoke” variable also has a lower magnitude than the pooled median regression estimate by a

Table 3.4 Birthweight, quantile regression with CE, all moms, (unit:grams)

	Quantile				
	0.1	0.25	0.5	0.75	0.9
Smoke	-129.99	-136.67	-138.26	-148.34	-152.86
	(10.75)	(7.29)	(6.31)	(7.63)	(9.80)
Male	102.44	122.63	138.87	150.31	162.12
	(3.79)	(2.91)	(2.51)	(2.87)	(3.41)
Age	-1.25	-6.75	-13.37	-9.05	-4.47
	(8.34)	(5.38)	(5.38)	(5.27)	(3.04)
Age^2	1.94	0.26	0.35	0.25	1.12
	(0.13)	(0.08)	(0.07)	(0.08)	(0.05)
Kessner index = 2	-121.12	-106.11	-79.17	-69.06	-75.41
	(10.29)	(6.90)	(5.66)	(6.45)	(8.65)
Kessner index = 3	-255.81	-220.18	-163.42	-138.40	-120.40
	(27.03)	(17.05)	(15.67)	(17.80)	(19.91)
No prenatal visit	-169.44	-40.80	-32.02	-9.20	-88.82
	(49.00)	(27.40)	(24.99)	(30.33)	(48.29)
First prenatal visit in 2nd trimester	92.05	88.90	67.38	53.81	69.25
	(12.20)	(7.97)	(6.73)	(7.59)	(10.72)
First prenatal visit in 3rd trimester	260.68	178.45	109.92	97.94	112.61
	(36.80)	(20.16)	(18.82)	(21.59)	(25.68)
# of Dictionary Var.	301	301	301	301	301
# of Selected Var.	150	246	298	269	141

similar amount. Moreover, the FE and CE coefficient estimates on other variables show similar patterns of changes from the OLS and pooled median regression, respectively. For example, the coefficient estimates of OLS/FE on age and age^2 alternate in sign and similar patterns are found in the pooled median regression and CE estimates on age and age^2. Overall, the CE estimates are quite close to the FE estimates, and they can be regarded as a median analogue of the FE estimates. Note that this is sensible because, considering the nature of dependent variables, we expect the conditional distribution of regression errors are fairly symmetric, and because the CE estimator takes the control function analogue of the FE estimator. For the unconditional distribution, the mean/median pairs of birth weights for $b = 1, 2$ and 3 are 3426g/3430g, 3482g/3487g and 3517g/3520g, respectively. Figure A.1 in Appendix C.2 shows a frequency histrogram for pooled birth weights across all births.

Table 3.5 Coefficient Estimates on 'Smoke' using Different ICs, (unit:grams)

	Quantile				
	0.1	0.25	0.5	0.75	0.9
QBIC	-132.51	-145.97	-145.27	-147.74	-152.41
	[88]	[75]	[45]	[40]	[93]
QBICL	-140.61	-158.1	-153.97	-152.33	-245.82
	[37]	[37]	[30]	[27]	[30]
BIC	-129.99	-136.66	-138.26	-148.34	-152.86
	[150]	[246]	[298]	[269]	[141]
BICL	-129.99	-140.18	-140.34	-146.47	-152.86
	[150]	[179]	[169]	[244]	[141]

Table 3.4 contains the CE estimates for 10, 25, 50, 75 and 90 percentiles.[2] The same set of dictionary variables is used with BIC for all cases. Evidently, the magnitude of the coefficient estimates on "smoke" variable declines as the percentile decreases. Note that the pooled quantile regression results in Appendix C.2 shows an exact opposite relationship, which indicates that the impact of smoking is more severely overestimated in the pooled regressions for the lower quantiles. Although the absolute magnitude of impact declines, its proportionate impact can be larger for lower quantiles. For example, relative to the fitted values for a two-birth mom who had two female babies at age 27, and 28 (with all other dummy variables equal to zero), the proportionate impacts of smoking for 10, 25, 50, 75 and 90 percentiles are -5.13%, -4.60%, -4.03%, -3.98%, and -3.92%, respectively.

There are some computational issues that need to be addressed. First, the main computational challenge in the SCAD or MCP penalized estimator lies in the non-convex nature of the objective function. The numerical algorithms studied so far use some version of an approximated objective function. In this chapter, all estimates are computed using iterative quantile regression on an augmented data set based on local linear approximation of the SCAD panelty function (Sherwood and Wang, 2016). Second, when Sherwood and Wang (2016)'s iterative quantile regression method

[2] The estimates in Table A.13 were computed with classical CRE (after dropping linearly dependent terms). CRE estimator is less robust than the proposed CE estimator by construction.

is used on the given data set, the selection path can vanish for small enough threshold parameters at τ close enough to 0 or 1. That is, the penalized estimator is essentially not computable for a small enough λ at high-end or low-end quantiles. The .1 and .9 quantile results could have more selected terms if there was no such problems. Table 3.5 shows the coefficient estimates for the “smoke” variable based on four different Bayesian-type information criteria. The bracketed numbers are the number of selected variables out of 301 dictionary variables. BIC and BICL do not have any difference at the 10 and 90 percentiles. Unreported results show that .15 and .85 quantiles do not suffer from the problem. It seems that the numerical algorithm for the quantile regression matters. For the given data set, Koenker and d’Orey’s (1987; KO) algorithm yields more stable results than Armstrong, Frome and Kung’s (1979; AFK) algorithm.

3.7 Concluding Remarks

I propose a new model restriction and estimation procedure for a linear panel data quantile regression model with fixed T. By introducing a nonparametric correlated effect, the new model restriction reasonably accounts for the τ-quantile-specific time-invariant heterogeneity and allows arbitrary within-group dependence of regression errors. A non-convex penalized estimation procedure is employed under the sparsity assumption on the correlated effect. To make the sparsity assumption more plausible in some cases, a transformation of the approximated correlated effect into a generalized Mundlak form is proposed.

There are interesting questions to be answered in future research. First, it would be useful to study Bayesian type information criteria that allow diverging p_N and q_N and attain selection consistency under a certain degree of misspecification. Second, the numerical algorithm for a nonconvex objective function can be improved for more stable and efficient computation. Third, extending the current framework to account for a censored response variable and time-varying endogeneity is another interesting direction to pursue. The extended estimator is expected to have similar advantages to the estimator studied in this chapter.

APPENDICES

APPENDIX A

AN APPENDIX FOR CHAPTER 1

A.1 Assumptions

Assumptions (1) $(\mathbf{y}_{i1}, \mathbf{y}_{i2}, \mathbf{z}_i)$ are i.i.d. (2) $\Theta \underset{cpt}{\subset} \mathbb{R}^p$ (3) $q_1 : \Theta \times W \to \mathbb{R}$ and $q_2 : \Theta_2 \times W \to \mathbb{R}$ where $(\mathbf{y}_{i1}, \mathbf{y}_{i2}, \mathbf{z}_i) \in W$ (4) $\theta_o \in int\,(\Theta)$ and let $\mathcal{N}$ be a neighborhood of θ_o (5) with probability one, $q_1\,(\mathbf{y}_{i1}, \mathbf{y}_{i2}, \mathbf{z}_i, \theta_1, \theta_2)$ and $q_2\,(\mathbf{y}_{i2}, \mathbf{z}_i, \theta_2)$ are continuously differentiable at each $\theta \in \Theta$ and twice continuosly differentiable in $\mathcal{N}$. (6) $E\left[\sup_{\theta\in\Theta}\left\|\begin{matrix}\frac{\partial(q_1+q_2)}{\partial\theta}\\ \frac{\partial q_2}{\partial\theta_2}\end{matrix}\right\|\right] < \infty$ (7) each element of $\frac{\partial q_1(\mathbf{y}_{i1},\mathbf{y}_{i2},\mathbf{z}_i,\theta_{o1},\theta_{o2})}{\partial\theta}$ and $\frac{\partial q_2(\mathbf{y}_{i2},\mathbf{z}_i,\theta_{o2})}{\partial\theta_2}$ has finite second moment. (8) $E\left[\sup_{\theta\in\mathcal{N}}\left\|\begin{matrix}\frac{\partial(q_1+q_2)}{\partial\theta\partial\theta'}\\ \frac{\partial q_2}{\partial\theta_2\partial\theta'}\end{matrix}\right\|\right] < \infty$ (9) (QLIML f.o.c.) $\{\theta_o\} = \left\{\theta \in \Theta : E\left[\frac{\partial(q_1+q_2)}{\partial\theta}\right] = 0\right\}$ (10) (CF f.o.c.) $\{\theta_o\} = \left\{\theta \in \Theta : E\left[\begin{matrix}\frac{\partial q_1(\theta)}{\partial\theta_1}\\ \frac{\partial q_2(\theta_2)}{\partial\theta_2}\end{matrix}\right] = 0\right\}$ (11) (QLIML rank condition) $E\left[\frac{\partial(q_1+q_2)(\theta_o)}{\partial\theta\partial\theta'}\right]$ and $V\left[\frac{\partial(q_1+q_2)(\theta_o)}{\partial\theta}\right]$ are invertible. (12) (CF rank condition) $E\left[\begin{matrix}\frac{\partial q_1(\theta_o)}{\partial\theta_1\partial\theta'}\\ \frac{\partial q_2(\theta_{o2})}{\partial\theta_2\partial\theta'}\end{matrix}\right]$ and $V\left[\begin{matrix}\frac{\partial q_1(\theta_o)}{\partial\theta_1}\\ \frac{\partial q_2(\theta_{o2})}{\partial\theta_2}\end{matrix}\right]$ are invertible.

A.2 Proofs

A.2.1 Proof of Proposition 1.3.4

Under regularity conditions, it suffices to show the followings (Newey and McFadden, 1994):

(a) $E\left[\sup_{\theta\in\Theta}\left\|\begin{matrix}\frac{\partial q_1(\theta)}{\partial\theta_1}\\ \frac{\partial q_1(\theta)}{\partial\theta_{22}}\\ \frac{\partial q_2(\theta_2)}{\partial\theta_2}\end{matrix}\right\|\right] < \infty$ (b) $E\left[\sup_{\theta\in\mathcal{N}}\left\|\begin{matrix}\frac{\partial q_1(\theta)}{\partial\theta_1\partial\theta'}\\ \frac{\partial q_1(\theta)}{\partial\theta_{22}\partial\theta'}\\ \frac{\partial q_2(\theta_2)}{\partial\theta_2\partial\theta'}\end{matrix}\right\|\right] < \infty$

(c) $\{\theta_o\} = \left\{\theta \in \Theta : E\left[\begin{array}{c}\frac{\partial q_1(\theta)}{\partial \theta_1}\\ \frac{\partial q_1(\theta)}{\partial \theta_{22}}\\ \frac{\partial q_2(\theta_2)}{\partial \theta_2}\end{array}\right]\Bigg]\right\}$ (d) $E\left[\begin{array}{c}\frac{\partial q_1(\theta_o)}{\partial \theta_1 \partial \theta'}\\ \frac{\partial q_1(\theta_o)}{\partial \theta_{22} \partial \theta'}\\ \frac{\partial q_2(\theta_{o2})}{\partial \theta_2 \partial \theta'}\end{array}\right]$ has full column rank.

(e) $V\left(\begin{array}{c}\frac{\partial q_1(\theta_o)}{\partial \theta_1}\\ \frac{\partial q_1(\theta_o)}{\partial \theta_{22}}\\ \frac{\partial q_2(\theta_{o2})}{\partial \theta_2}\end{array}\right)$ is invertible.

(c) and (e) are direct implications of definition of GMM-QLIML. (d) can be shown from Assumption 12 since adding extra rows does not affect column rank. (a) and (b) are implied by triangular inequality together with Assumption 6 and Assumption 8. i.e. $\left\|\frac{\partial q_1(\theta)}{\partial \theta_{22}}\right\| \leq \left\|\frac{\partial q_1(\theta)}{\partial \theta_{22}} + \frac{\partial q_2(\theta_2)}{\partial \theta_{22}}\right\| + \left\|\frac{\partial q_2(\theta_2)}{\partial \theta_{22}}\right\|$ and $\left\|\frac{\partial q_1(\theta)}{\partial \theta_{22} \partial \theta'}\right\| \leq \left\|\frac{\partial q_1(\theta)}{\partial \theta_{22} \partial \theta'} + \frac{\partial q_2(\theta_2)}{\partial \theta_{22} \partial \theta'}\right\| + \left\|\frac{\partial q_2(\theta_2)}{\partial \theta_{22} \partial \theta'}\right\|$

A.2.2 Proof of Proposition 1.3.6

(a) and (b) are directly implied by following lemma.

Lemma A.2.1 Let G be a linear span of moment functions $\{g_i\}$ in (1.8). Optimal GMM based on each maximal linearly independent set at true parameter values in G yields asymptotically equivalent estimator.

Proof. Suppose $\{\tilde{g}_i(\theta_o)\}$ and $\{\hat{g}_i(\theta_o)\}$ are maximal linearly independent subsets in linear span of moment functions $\{g_i(\theta_o)\}$. First, the number of moments in $\{\tilde{g}_i(\theta_o)\}$ and $\{\hat{g}_i(\theta_o)\}$ is the same since both $\{\tilde{g}_i(\theta_o)\}$ and $\{\hat{g}_i(\theta_o)\}$ are basis for $span(\{g_i(\theta_o)\})$ and $span(\{g_i(\theta_o)\})$ is a finite dimensional vector space. Second, there exists an invertible linear map $A(\theta_o)$ such that $A(\theta_o)\left[\begin{array}{c}\tilde{g}_1(\theta_o)\\ \vdots\\ \tilde{g}_k(\theta_o)\end{array}\right] = \left[\begin{array}{c}\hat{g}_1(\theta_o)\\ \vdots\\ \hat{g}_k(\theta_o)\end{array}\right]$ by definition of basis. By proposition 3.4, efficient GMM based on $\{\tilde{g}_i(\theta_o)\}$ and $\{\hat{g}_i(\theta_o)\}$ are well-defined and asymptotic normal. Then, it is easy to see

$$
\begin{aligned}
& E\left[\frac{\partial \hat{g}(\theta_o)}{\partial \theta}\right]' V\left(\frac{\partial \hat{g}(\theta_o)}{\partial \theta}\right)^{-1} E\left[\frac{\partial \hat{g}(\theta_o)}{\partial \theta}\right] \\
& = E\left[\frac{\partial \tilde{g}(\theta_o)}{\partial \theta}\right]' A(\theta_o)' \left(A(\theta_o)'\right)^{-1} V\left(\frac{\partial \tilde{g}(\theta_o)}{\partial \theta}\right)^{-1} \left(A(\theta_o)\right)^{-1} A(\theta_o)\, E\left[\frac{\partial \tilde{g}(\theta_o)}{\partial \theta}\right] \\
& = E\left[\frac{\partial \tilde{g}(\theta_o)}{\partial \theta}\right]' V\left(\frac{\partial \tilde{g}(\theta_o)}{\partial \theta}\right)^{-1} E\left[\frac{\partial \tilde{g}(\theta_o)}{\partial \theta}\right]
\end{aligned}
$$

■

A.2.3 Statement (d),(e) and Proof of Proposition 1.3.7

(d) $V^S_{GMM-QLIML} = V^S_{QLIML}$ iff

$$
\begin{aligned}
& E_o\left[\frac{\partial q^o_{i2}}{\partial \theta_{22}\partial \theta'_S}\right] - cov_o\left(\frac{\partial q^o_{i2}}{\partial \theta_{22}}, \begin{bmatrix} \frac{\partial q^o_{i1}}{\partial \theta_1} \\ \frac{\partial q^o_{i1}}{\partial \theta_2} + \frac{\partial q^o_{i2}}{\partial \theta_2} \end{bmatrix}\right) W^*_o E_o \begin{bmatrix} \frac{\partial q^o_{i1}}{\partial \theta_1 \partial \theta'_S} \\ \frac{\partial q^o_{i1}}{\partial \theta_2 \partial \theta'_S} + \frac{\partial q^o_{i2}}{\partial \theta_2 \partial \theta'_S} \end{bmatrix} \\
& = \left[E_o\left[\frac{\partial q^o_{i2}}{\partial \theta_{22}\partial \theta'_{-S}}\right] - cov_o\left(\frac{\partial q^o_{i2}}{\partial \theta_{22}}, \begin{bmatrix} \frac{\partial q^o_{i1}}{\partial \theta_1} \\ \frac{\partial q^o_{i1}}{\partial \theta_2} + \frac{\partial q_{i2}}{\partial \theta_2} \end{bmatrix}\right) W^*_o E_o \begin{bmatrix} \frac{\partial q^o_{i1}}{\partial \theta_1 \partial \theta'_{-S}} \\ \frac{\partial q^o_{i1}}{\partial \theta_2 \partial \theta'_{-S}} + \frac{\partial q^o_{i2}}{\partial \theta_2 \partial \theta'_{-S}} \end{bmatrix} \right] \\
& \times \left[E_o \begin{bmatrix} \frac{\partial q^o_{i1}}{\partial \theta_1 \partial \theta'_{-S}} \\ \frac{\partial q^o_{i1}}{\partial \theta_2 \partial \theta'_{-S}} + \frac{\partial q^o_{i2}}{\partial \theta_2 \partial \theta'_{-S}} \end{bmatrix}' W^*_o E_o \begin{bmatrix} \frac{\partial q^o_{i1}}{\partial \theta_1 \partial \theta'_{-S}} \\ \frac{\partial q^o_{i1}}{\partial \theta_2 \partial \theta'_{-S}} + \frac{\partial q^o_{i2}}{\partial \theta_2 \partial \theta'_{-S}} \end{bmatrix} \right]^{-1} \\
& \times \left[E_o \begin{bmatrix} \frac{\partial q^o_{i1}}{\partial \theta_1 \partial \theta'_{-S}} \\ \frac{\partial q^o_{i1}}{\partial \theta_2 \partial \theta'_{-S}} + \frac{\partial q^o_{i2}}{\partial \theta_2 \partial \theta'_{-S}} \end{bmatrix}' W^*_o E_o \begin{bmatrix} \frac{\partial q^o_{i1}}{\partial \theta_1 \partial \theta'_S} \\ \frac{\partial q^o_{i1}}{\partial \theta_2 \partial \theta'_S} + \frac{\partial q^o_{i2}}{\partial \theta_2 \partial \theta'_S} \end{bmatrix} \right]
\end{aligned}
$$

where

$$
W^*_o = V_o \begin{pmatrix} \frac{\partial q^o_{i1}}{\partial \theta_1} \\ \frac{\partial q^o_{i1}}{\partial \theta_2} + \frac{\partial q^o_{i2}}{\partial \theta_2} \end{pmatrix}^{-1}
$$

(e) $V^S_{GMM-QLIML} = V^S_{CF}$ iff

$$
E_o\left[\frac{\partial q^o_{i1}}{\partial \theta_{22}\partial \theta'_S}\right] - cov_o\left(\frac{\partial q^o_{i1}}{\partial \theta_{22}}, \begin{bmatrix} \frac{\partial q^o_{i1}}{\partial \theta_1} \\ \frac{\partial q^o_{i2}}{\partial \theta_2} \end{bmatrix}\right) V_o \begin{pmatrix} \frac{\partial q^o_{i1}}{\partial \theta_1} \\ \frac{\partial q^o_{i2}}{\partial \theta_2} \end{pmatrix}^{-1} E_o \begin{bmatrix} \frac{\partial q^o_{i1}}{\partial \theta_1 \partial \theta'_S} \\ \frac{\partial q^o_{i2}}{\partial \theta_2 \partial \theta'_S} \end{bmatrix}
$$

$$
\begin{aligned}
&= \left[E_o \left[\frac{\partial q_{i1}^o}{\partial \theta_{22} \partial \theta'_{-S}} \right] - cov_o \left(\frac{\partial q_{i1}^o}{\partial \theta_{22}}, \begin{bmatrix} \frac{\partial q_{i1}^o}{\partial \theta_1} \\ \frac{\partial q_{i2}^o}{\partial \theta_2} \end{bmatrix} \right) V_o \begin{pmatrix} \frac{\partial q_{i1}^o}{\partial \theta_1} \\ \frac{\partial q_{i2}^o}{\partial \theta_2} \end{pmatrix}^{-1} E_o \begin{bmatrix} \frac{\partial q_{i1}^o}{\partial \theta_1 \partial \theta'_{-S}} \\ \frac{\partial q_{i2}^o}{\partial \theta_2 \partial \theta'_{-S}} \end{bmatrix} \right] \\
&\times \left[E_o \begin{bmatrix} \frac{\partial q_{i1}^o}{\partial \theta_1 \partial \theta'_{-S}} \\ \frac{\partial q_{i2}^o}{\partial \theta_2 \partial \theta'_{-S}} \end{bmatrix}' V_o \begin{pmatrix} \frac{\partial q_{i1}^o}{\partial \theta_1} \\ \frac{\partial q_{i2}^o}{\partial \theta_2} \end{pmatrix}^{-1} E_o \begin{bmatrix} \frac{\partial q_{i1}^o}{\partial \theta_1 \partial \theta'_{-S}} \\ \frac{\partial q_{i2}^o}{\partial \theta_2 \partial \theta'_{-S}} \end{bmatrix} \right]^{-1} \\
&\times \left[E_o \begin{bmatrix} \frac{\partial q_{i1}^o}{\partial \theta_1 \partial \theta'_{-S}} \\ \frac{\partial q_{i2}^o}{\partial \theta_2 \partial \theta'_{-S}} \end{bmatrix}' V_o \begin{pmatrix} \frac{\partial q_{i1}^o}{\partial \theta_1} \\ \frac{\partial q_{i2}^o}{\partial \theta_2} \end{pmatrix}^{-1} E_o \begin{bmatrix} \frac{\partial q_{i1}^o}{\partial \theta_1 \partial \theta'_S} \\ \frac{\partial q_{i2}^o}{\partial \theta_2 \partial \theta'_S} \end{bmatrix} \right]
\end{aligned}
$$

Proof

(a) $V_{GMM-QLIML} \preceq V_{CF}$ is trivial. To see $V_{GMM-QLIML} \preceq V_{QLIML}$, note first that, at true parameter value,

$$
\begin{bmatrix} \frac{\partial}{\partial \theta_1} q_1 (\theta_1, \theta_2) \\ \frac{\partial}{\partial \theta_2} q_1 (\theta_1, \theta_2) + \frac{\partial}{\partial \theta_2} q_2 (\theta_2) \end{bmatrix}
$$

is linearly independent by Assumption 11. Thus, there exists an extension to a basis

$$
\begin{bmatrix} \frac{\partial}{\partial \theta_1} q_1 (\theta_1, \theta_2) \\ \frac{\partial}{\partial \theta_2} q_1 (\theta_1, \theta_2) + \frac{\partial}{\partial \theta_2} q_2 (\theta_2) \\ \frac{\partial}{\partial \theta_2^*} q_1 (\theta) \end{bmatrix} \tag{A.1}
$$

which is an invertible linear transformation of (1.9). Hence, the result follows by Lemma C.1.

(b) Apply BQSW redundancy condition to (1.9).

(c) Apply BQSW redundancy condition to (A.1).

(d),(e) Apply BQSW partial redundancy results to (1.9) and (A.1) , respectively.

A.2.4 Proof of Corollary 1.3.9

(a) Suppose GIMEs:

$$V_o\left(\frac{\partial q_1^o}{\partial\theta}\right) = \tau E_o\left[-\frac{\partial q_1^o}{\partial\theta\partial\theta'}\right], V_o\left(\frac{\partial q_2^o}{\partial\theta_2}\right) = \tau E_o\left[-\frac{\partial q_2^o}{\partial\theta_2\partial\theta_2'}\right],$$

$$V_o\left(\begin{bmatrix} \frac{\partial q_1^o}{\partial\theta_1} \\ \frac{\partial q_1^o}{\partial\theta_2} + \frac{\partial q_2^o}{\partial\theta_2} \end{bmatrix}\right) = \tau E_o\begin{bmatrix} -\frac{\partial q_1^o}{\partial\theta_1\partial\theta'} \\ -\frac{\partial q_1^o}{\partial\theta_2\partial\theta'} - \frac{\partial q_2^o}{\partial\theta_2\partial\theta'} \end{bmatrix}$$

It is implied that $cov_o\left(\frac{\partial q_1^o}{\partial\theta_1}, \frac{\partial q_2^o}{\partial\theta_2}\right) = 0$. Then the condition (c) in Proposition 1.3.7 follows. To see $V_{QLIML} \neq V_{CF}$, consider a submatrix of

$$cov_o\left(\frac{\partial q_{i1}}{\partial\theta_{22}}, \frac{\partial q_{i1}}{\partial\theta_2}\right) - cov_o\left(\frac{\partial q_{i1}}{\partial\theta_{22}}, \frac{\partial q_{i1}}{\partial\theta_1}\right) V_o\left(\frac{\partial q_{i1}}{\partial\theta_1}\right)^{-1} cov_o\left(\frac{\partial q_{i1}}{\partial\theta_1}, \frac{\partial q_{i1}}{\partial\theta_2}\right),$$

$$\begin{aligned}
&cov_o\left(\frac{\partial q_{i1}}{\partial\theta_{22}}, \frac{\partial q_{i1}}{\partial\theta_{22}}\right) - cov_o\left(\frac{\partial q_{i1}}{\partial\theta_{22}}, \frac{\partial q_{i1}}{\partial\theta_1}\right) V_o\left(\frac{\partial q_{i1}}{\partial\theta_1}\right)^{-1} cov_o\left(\frac{\partial q_{i1}}{\partial\theta_1}, \frac{\partial q_{i1}}{\partial\theta_{22}}\right) \\
&= V_o\left(\frac{\partial q_{i1}}{\partial\theta_{22}}\right) - cov_o\left(\frac{\partial q_{i1}}{\partial\theta_{22}}, \frac{\partial q_{i1}}{\partial\theta_1}\right) V_o\left(\frac{\partial q_{i1}}{\partial\theta_1}\right)^{-1} V_o\left(\frac{\partial q_{i1}}{\partial\theta_1}\right) V_o\left(\frac{\partial q_{i1}}{\partial\theta_1}\right)^{-1} cov_o\left(\frac{\partial q_{i1}}{\partial\theta_1}, \frac{\partial q_{i1}}{\partial\theta_{22}}\right) \\
&= V_o\left(\frac{\partial q_{i1}}{\partial\theta_{22}}\right) - V\left(cov_o\left(\frac{\partial q_{i1}}{\partial\theta_{22}}, \frac{\partial q_{i1}}{\partial\theta_1}\right) V_o\left(\frac{\partial q_{i1}}{\partial\theta_1}\right)^{-1} \frac{\partial q_{i1}}{\partial\theta_1}\right)
\end{aligned}$$

The last expression can be interpreted as difference of outer products of $\frac{\partial q_{i1}}{\partial\theta_{22}}$ and $L\left(\frac{\partial q_{i1}}{\partial\theta_{22}}\middle| \frac{\partial q_{i1}}{\partial\theta_1}\right)$. Since $\frac{\partial q_{i1}}{\partial\theta_{22}}$ and $\frac{\partial q_{i1}}{\partial\theta_1}$ are assumed to be linearly independent at true parameter, it cannot be zero.

(b) Suppose GIMEs. Without loss of generality, $\tau = 1$ is assumed.Then, the result (e) of Proposition 1.3.7 implies

$$\begin{aligned}
(LHS) &= E_o\left[\frac{\partial q_{i1}^o}{\partial\theta_{22}\partial\theta_{11}'}\right] - cov_o\left(\frac{\partial q_{i1}^o}{\partial\theta_{22}}, \begin{bmatrix} \frac{\partial q_{i1}^o}{\partial\theta_1} \\ \frac{\partial q_{i2}^o}{\partial\theta_2} \end{bmatrix}\right) V_o\begin{pmatrix} \frac{\partial q_{i1}^o}{\partial\theta_1} \\ \frac{\partial q_{i2}^o}{\partial\theta_2} \end{pmatrix}^{-1} E_o\begin{bmatrix} \frac{\partial q_{i1}^o}{\partial\theta_1\partial\theta_{11}'} \\ \frac{\partial q_{i2}^o}{\partial\theta_2\partial\theta_{11}'} \end{bmatrix} \\
&= 0_{p_{22}\times p_{11}}
\end{aligned}$$

For RHS,

$$
\begin{aligned}
&(1st \text{ part of } RHS)\\
&= E\left[\frac{\partial q_{i1}}{\partial\theta_{22}\partial\left(\theta'_{12},\theta'_2\right)}\right] - cov\left(\frac{\partial q_{i1}}{\partial\theta_{22}}, \left[\begin{array}{c}\frac{\partial q_{i1}}{\partial\theta_1}\\ \frac{\partial q_{i2}}{\partial\theta_2}\end{array}\right]\right) V\left(\begin{array}{c}\frac{\partial q_{i1}}{\partial\theta_1}\\ \frac{\partial q_{i2}}{\partial\theta_2}\end{array}\right)^{-1} E\left[\begin{array}{c}\frac{\partial q_{i1}}{\partial\theta_1\partial\left(\theta'_{12},\theta'_2\right)}\\ \frac{\partial q_{i2}}{\partial\theta_2\partial\left(\theta'_{12},\theta'_2\right)}\end{array}\right]\\
&= \left[\begin{array}{cc} cov\left(\frac{\partial q_{i1}}{\partial\theta_{22}},\frac{\partial q_{i1}}{\partial\theta_{12}}\right) & cov\left(\frac{\partial q_{i1}}{\partial\theta_{22}},\frac{\partial q_{i1}}{\partial\theta_2}\right)\end{array}\right]\\
&-\left[\begin{array}{cc} cov\left(\frac{\partial q_{i1}}{\partial\theta_{22}},\frac{\partial q_{i1}}{\partial\theta_1}\right) & 0_{p_{22}\times p_2}\end{array}\right]\left[\begin{array}{cc} V\left(\frac{\partial q_{i1}}{\partial\theta_1}\right)^{-1} & 0_{p_1\times p_2}\\ 0_{p_2\times p_1} & V\left(\frac{\partial q_{i2}}{\partial\theta_2}\right)^{-1}\end{array}\right]\\
&\times\left[\begin{array}{cc} cov\left(\frac{\partial q_{i1}}{\partial\theta_1},\frac{\partial q_{i1}}{\partial\theta_{12}}\right) & cov\left(\frac{\partial q_{i1}}{\partial\theta_1},\frac{\partial q_{i1}}{\partial\theta_2}\right)\\ 0_{p_2\times p_{12}} & V\left(\frac{\partial q_{i2}}{\partial\theta_2}\right)\end{array}\right]\\
&= \left[\begin{array}{cc} 0_{p_{22}\times p_{11}} & cov\left(\frac{\partial q_{i1}}{\partial\theta_{22}},\frac{\partial q_{i1}}{\partial\theta_2}\right) - cov\left(\frac{\partial q_{i1}}{\partial\theta_{22}},\frac{\partial q_{i1}}{\partial\theta_1}\right) V\left(\frac{\partial q_{i1}}{\partial\theta_1}\right)^{-1} cov\left(\frac{\partial q_{i1}}{\partial\theta_1},\frac{\partial q_{i1}}{\partial\theta_2}\right)\end{array}\right]
\end{aligned}
$$

and

$$
\begin{aligned}
&(2nd \text{ part of RHS})\\
&= \left[\begin{array}{cc} cov\left(\frac{\partial q_{i1}}{\partial\theta_1},\frac{\partial q_{i1}}{\partial\theta_{12}}\right) & cov\left(\frac{\partial q_{i1}}{\partial\theta_1},\frac{\partial q_{i1}}{\partial\theta_2}\right)\\ 0_{p_2\times p_{12}} & V\left(\frac{\partial q_{i2}}{\partial\theta_2}\right)\end{array}\right]'\left[\begin{array}{cc} V\left(\frac{\partial q_{i1}}{\partial\theta_1}\right)^{-1} & 0_{p_1\times p_2}\\ 0_{p_2\times p_1} & V\left(\frac{\partial q_{i2}}{\partial\theta_2}\right)^{-1}\end{array}\right]\\
&\times\left[\begin{array}{cc} cov\left(\frac{\partial q_{i1}}{\partial\theta_1},\frac{\partial q_{i1}}{\partial\theta_{12}}\right) & cov\left(\frac{\partial q_{i1}}{\partial\theta_1},\frac{\partial q_{i1}}{\partial\theta_2}\right)\\ 0_{p_2\times p_{12}} & V\left(\frac{\partial q_{i2}}{\partial\theta_2}\right)\end{array}\right]\\
&= \left[\begin{array}{cc} V\left(\frac{\partial q_{i1}}{\partial\theta_{12}}\right) & cov\left(\frac{\partial q_{i1}}{\partial\theta_{12}},\frac{\partial q_{i1}}{\partial\theta_2}\right)\\ cov\left(\frac{\partial q_{i1}}{\partial\theta_2},\frac{\partial q_{i1}}{\partial\theta_{12}}\right) & cov\left(\frac{\partial q_{i1}}{\partial\theta_2},\frac{\partial q_{i1}}{\partial\theta_1}\right) V\left(\frac{\partial q_{i1}}{\partial\theta_1}\right)^{-1} cov\left(\frac{\partial q_{i1}}{\partial\theta_1},\frac{\partial q_{i1}}{\partial\theta_2}\right) + V\left(\frac{\partial q_{i2}}{\partial\theta_2}\right)\end{array}\right]
\end{aligned}
$$

and

$$
\begin{aligned}
&(3rd \text{ part of RHS})\\
&= \begin{bmatrix} cov\left(\frac{\partial q_{i1}}{\partial\theta_1},\frac{\partial q_{i1}}{\partial\theta_{12}}\right) & cov\left(\frac{\partial q_{i1}}{\partial\theta_1},\frac{\partial q_{i1}}{\partial\theta_2}\right) \\ 0_{p_2\times p_{12}} & V\left(\frac{\partial q_{i2}}{\partial\theta_2}\right) \end{bmatrix}' \begin{bmatrix} V\left(\frac{\partial q_{i1}}{\partial\theta_1}\right)^{-1} & 0_{p_1\times p_2} \\ 0_{p_2\times p_1} & V\left(\frac{\partial q_{i2}}{\partial\theta_2}\right)^{-1} \end{bmatrix}\\
&\times \begin{bmatrix} cov\left(\frac{\partial q_{i1}}{\partial\theta_1},\frac{\partial q_{i1}}{\partial\theta_{11}}\right) \\ 0_{p_2\times p_{11}} \end{bmatrix}\\
&= \begin{bmatrix} cov\left(\frac{\partial q_{i1}}{\partial\theta_{12}},\frac{\partial q_{i1}}{\partial\theta_{11}}\right) \\ cov\left(\frac{\partial q_{i1}}{\partial\theta_2},\frac{\partial q_{i1}}{\partial\theta_{11}}\right) \end{bmatrix}
\end{aligned}
$$

Consider the inverse of the second part:

$$
\begin{aligned}
&\begin{bmatrix} V\left(\frac{\partial q_{i1}}{\partial\theta_{12}}\right) & cov\left(\frac{\partial q_{i1}}{\partial\theta_{12}},\frac{\partial q_{i1}}{\partial\theta_2}\right) \\ cov\left(\frac{\partial q_{i1}}{\partial\theta_2},\frac{\partial q_{i1}}{\partial\theta_{12}}\right) & cov\left(\frac{\partial q_{i1}}{\partial\theta_2},\frac{\partial q_{i1}}{\partial\theta_1}\right) V\left(\frac{\partial q_{i1}}{\partial\theta_1}\right)^{-1} cov\left(\frac{\partial q_{i1}}{\partial\theta_1},\frac{\partial q_{i1}}{\partial\theta_2}\right) + V\left(\frac{\partial q_{i2}}{\partial\theta_2}\right) \end{bmatrix}^{-1}\\
&= \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}
\end{aligned}
$$

Then (RHS) can be expressed as

$$
\begin{aligned}
&(RHS)\\
&= \begin{bmatrix} 0_{p_{22}\times p_{11}} & cov\left(\frac{\partial q_{i1}}{\partial\theta_{22}},\frac{\partial q_{i1}}{\partial\theta_2}\right) - cov\left(\frac{\partial q_{i1}}{\partial\theta_{22}},\frac{\partial q_{i1}}{\partial\theta_1}\right) V\left(\frac{\partial q_{i1}}{\partial\theta_1}\right)^{-1} cov\left(\frac{\partial q_{i1}}{\partial\theta_1},\frac{\partial q_{i1}}{\partial\theta_2}\right) \end{bmatrix}\\
&\times \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} \times \begin{bmatrix} cov\left(\frac{\partial q_{i1}}{\partial\theta_{12}},\frac{\partial q_{i1}}{\partial\theta_{11}}\right) \\ cov\left(\frac{\partial q_{i1}}{\partial\theta_2},\frac{\partial q_{i1}}{\partial\theta_{11}}\right) \end{bmatrix}\\
&= \left[cov\left(\frac{\partial q_{i1}}{\partial\theta_{22}},\frac{\partial q_{i1}}{\partial\theta_2}\right) - cov\left(\frac{\partial q_{i1}}{\partial\theta_{22}},\frac{\partial q_{i1}}{\partial\theta_1}\right) V\left(\frac{\partial q_{i1}}{\partial\theta_1}\right)^{-1} cov\left(\frac{\partial q_{i1}}{\partial\theta_1},\frac{\partial q_{i1}}{\partial\theta_2}\right)\right]\\
&\times \left[R_{21} \times cov\left(\frac{\partial q_{i1}}{\partial\theta_{12}},\frac{\partial q_{i1}}{\partial\theta_{11}}\right) + R_{22} \times cov\left(\frac{\partial q_{i1}}{\partial\theta_2},\frac{\partial q_{i1}}{\partial\theta_{11}}\right)\right]
\end{aligned}
$$

where

$$\begin{aligned}
R_{21} &= -\left[cov\left(\frac{\partial q_{i1}}{\partial\theta_2},\frac{\partial q_{i1}}{\partial\theta_1}\right) V\left(\frac{\partial q_{i1}}{\partial\theta_1}\right)^{-1} cov\left(\frac{\partial q_{i1}}{\partial\theta_1},\frac{\partial q_{i1}}{\partial\theta_2}\right) + V\left(\frac{\partial q_{i2}}{\partial\theta_2}\right)\right]^{-1} cov\left(\frac{\partial q_{i1}}{\partial\theta_{12}},\frac{\partial q_{i1}}{\partial\theta_2}\right) \\
&\quad \times \left(V\left(\frac{\partial q_{i1}}{\partial\theta_{12}}\right) - \left[cov\left(\frac{\partial q_{i1}}{\partial\theta_{12}},\frac{\partial q_{i1}}{\partial\theta_2}\right) M_o^{-1} cov\left(\frac{\partial q_{i1}}{\partial\theta_2},\frac{\partial q_{i1}}{\partial\theta_{12}}\right) cov\left(\frac{\partial q_{i1}}{\partial\theta_{12}},\frac{\partial q_{i1}}{\partial\theta_2}\right)\right]\right)^{-1} \\
R_{22} &= \left[M_o - cov\left(\frac{\partial q_{i1}}{\partial\theta_2},\frac{\partial q_{i1}}{\partial\theta_{12}}\right) V\left(\frac{\partial q_{i1}}{\partial\theta_{12}}\right)^{-1} cov\left(\frac{\partial q_{i1}}{\partial\theta_{12}},\frac{\partial q_{i1}}{\partial\theta_2}\right)\right]^{-1} \\
M_o &= cov\left(\frac{\partial q_{i1}}{\partial\theta_2},\frac{\partial q_{i1}}{\partial\theta_1}\right) V\left(\frac{\partial q_{i1}}{\partial\theta_1}\right)^{-1} cov\left(\frac{\partial q_{i1}}{\partial\theta_1},\frac{\partial q_{i1}}{\partial\theta_2}\right) + V\left(\frac{\partial q_{i2}}{\partial\theta_2}\right)
\end{aligned}$$

A.2.5 Proof of Proposition 1.3.10

For

$$\begin{bmatrix} m_1(\zeta_1,\zeta_2) \\ m_2(\zeta_1,\zeta_2) \end{bmatrix}$$

asymptotically equivalent linearized moment functions are

$$\begin{aligned}
&\begin{bmatrix} l_{i1}(\zeta_1,\zeta_2) \\ l_{i2}(\zeta_1,\zeta_2) \end{bmatrix} \\
&= \begin{bmatrix} m_{i1}(\zeta_{o1},\zeta_{o2}) + E\left[\frac{\partial m_{i1}(\zeta_{o1},\zeta_{o2})}{\partial\zeta_1'}\right](\zeta_1-\zeta_{o1}) + E\left[\frac{\partial m_{i1}(\zeta_{o1},\zeta_{o2})}{\partial\zeta_2'}\right](\zeta_2-\zeta_{o2}) \\ m_{i2}(\zeta_{o1},\zeta_{o2}) + E\left[\frac{\partial m_{i2}(\zeta_{o1},\zeta_{o2})}{\partial\zeta_1'}\right](\zeta_1-\zeta_{o1}) + E\left[\frac{\partial m_{i2}(\zeta_{o1},\zeta_{o2})}{\partial\zeta_2'}\right](\zeta_2-\zeta_{o2}) \end{bmatrix}
\end{aligned}$$

By subtracting

$$E\left[\frac{\partial m_{i1}(\zeta_{o1},\zeta_{o2})}{\partial\zeta_2'}\right]\left(E\left[\frac{\partial m_{i2}(\zeta_{o1},\zeta_{o2})}{\partial\zeta_2'}\right]\right)^{-1} l_{i2}(\zeta_1,\zeta_2)$$

from $l_{i1}(\zeta_1,\zeta_2)$, we get

$$l_{i1}'(\zeta_1) = l_{i1}(\zeta_1,\zeta_2) - E\left[\frac{\partial m_{i1}(\zeta_{o1},\zeta_{o2})}{\partial\zeta_2'}\right]\left(E\left[\frac{\partial m_{i2}(\zeta_{o1},\zeta_{o2})}{\partial\zeta_2'}\right]\right)^{-1} l_{i2}(\zeta_1,\zeta_2)$$

where $l_{i1}'(\zeta_1)$ is a function of ζ_1 only. Then standard asymptotics from $l_{i1}'(\zeta_1)$ yields

$$V_{QLIML}^{\zeta_1} = A_1^{-1} B_1 A_1^{-1}$$

where

$$A_1 = E\left[\frac{\partial m_{i1}(\zeta_{o1}, \zeta_{o2})}{\partial \zeta_1'}\right] - E\left[\frac{\partial m_{i1}(\zeta_{o1}, \zeta_{o2})}{\partial \zeta_2'}\right]\left(E\left[\frac{\partial m_{i2}(\zeta_{o1}, \zeta_{o2})}{\partial \zeta_2'}\right]\right)^{-1} E\left[\frac{\partial m_{i2}(\zeta_{o1}, \zeta_{o2})}{\partial \zeta_1'}\right]$$

$$B_1 = V\left(m_{i1}(\zeta_{o1}, \zeta_{o2}) - E\left[\frac{\partial m_{i1}(\zeta_{o1}, \zeta_{o2})}{\partial \zeta_2'}\right]\left(E\left[\frac{\partial m_{i2}(\zeta_{o1}, \zeta_{o2})}{\partial \zeta_2'}\right]\right)^{-1} m_{i2}(\zeta_{o1}, \zeta_{o2})\right)$$

Since $E\left[\frac{\partial m_{i1}(\zeta_{o1},\zeta_{o2})}{\partial \zeta_2'}\right] = 0$, this simply reduces to

$$A_1 = E\left[\frac{\partial m_{i1}(\zeta_{o1}, \zeta_{o2})}{\partial \zeta_1'}\right]$$

$$B_1 = V(m_{i1}(\zeta_{o1}, \zeta_{o2}))$$

and the result is the same for the case of

$$\begin{bmatrix} m_1(\zeta_1, \zeta_2) \\ m_3(\zeta_1, \zeta_2) \end{bmatrix}$$

A.2.6 Proof of Corollary 1.3.11

First, consider following reparameterizations with η and $\sigma_{11|2}$

$$\eta \equiv \Sigma_{22}^{-1}\Sigma_{21}$$

$$\sigma_{11|2} \equiv \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

where $\theta_1 = \left(\alpha', \delta_1', \eta', \sigma_{11|2}\right)'$, $\theta_2 = \left(vec(\boldsymbol{\delta}_2)', vec(\Sigma_{22})\right)'$ (Similar proof can be done with original scores as well.) These modifications do not change parameter estimates (other than Σ_{11} and Σ_{21}) of any methods studied in this chapter. It is due to the fact that the first two reparameterization do not impose any restriction on parameter space. Now, the quasi-scores of q_1

are modified to

$$\frac{\partial q_1}{\partial \theta_1} = \begin{bmatrix} \sigma_{11|2}^{-1} e_i\left(\theta\right)\mathbf{x}' \\ \frac{1}{2}\sigma_{11|2}^{-2} h_i\left(\theta\right) \end{bmatrix}$$

$$\frac{\partial q_1}{\partial \theta_2} = \begin{bmatrix} -\sigma_{11|2}^{-1} e_i\left(\theta\right)\left[\eta \otimes \mathbf{z}'\right] \\ 0_{\frac{r(r+1)}{2}\times 1} \end{bmatrix}$$

where $\mathbf{x} = \begin{bmatrix} \mathbf{y}_2 & \mathbf{z}_1 & \mathbf{v}_2\left(\boldsymbol{\delta}_2\right) \end{bmatrix}$ and $\left(e_i\left(\theta\right), h_i\left(\theta\right)\right)$ is defined correspondingly. Moment functions for LIML and CF are

$$E\begin{bmatrix} \sigma_{11|2}^{-1}\left(y_1 - \mathbf{y}_2\boldsymbol{\alpha} - \mathbf{z}_1\delta_1 - \mathbf{v}_2\eta\right)\mathbf{y}_2' \\ \sigma_{11|2}^{-1}\left(y_1 - \mathbf{y}_2\boldsymbol{\alpha} - \mathbf{z}_1\delta_1 - \mathbf{v}_2\eta\right)\mathbf{z}_1' \\ \sigma_{11|2}^{-1}\left(y_1 - \mathbf{y}_2\boldsymbol{\alpha} - \mathbf{z}_1\delta_1 - \mathbf{v}_2\eta\right)\mathbf{v}_2' \\ \frac{1}{2}\sigma_{11|2}^{-2}\left[\left(y_1 - \mathbf{y}_2\boldsymbol{\alpha} - \mathbf{z}_1\delta_1 - \mathbf{v}_2\eta\right)^2 - \sigma_{11|2}\right] \\ vec\left(\mathbf{z}'\left(\mathbf{y}_2 - \mathbf{z}\boldsymbol{\delta}_2\right)\Sigma_{22}^{-1}\right) - vec\left(\mathbf{z}'\sigma_{11|2}^{-1}\left(y_1 - \mathbf{y}_2\boldsymbol{\alpha} - \mathbf{z}_1\delta_1 - \mathbf{v}_2\eta\right)\eta'\right) \\ \frac{1}{2}L_r vec\left(\Sigma_{22}^{-1}\mathbf{v}_{i2}\left(\boldsymbol{\delta}_2\right)'\mathbf{v}_{i2}\left(\boldsymbol{\delta}_2\right)\Sigma_{22}^{-1} - \Sigma_{22}^{-1}\right) \end{bmatrix} = 0 \qquad \text{(A.2)}$$

and

$$E\begin{bmatrix} \sigma_{11|2}^{-1}\left(y_1 - \mathbf{y}_2\boldsymbol{\alpha} - \mathbf{z}_1\delta_1 - \mathbf{v}_2\eta\right)\mathbf{y}_2' \\ \sigma_{11|2}^{-1}\left(y_1 - \mathbf{y}_2\boldsymbol{\alpha} - \mathbf{z}_1\delta_1 - \mathbf{v}_2\eta\right)\mathbf{z}_1' \\ \sigma_{11|2}^{-1}\left(y_1 - \mathbf{y}_2\boldsymbol{\alpha} - \mathbf{z}_1\delta_1 - \mathbf{v}_2\eta\right)\mathbf{v}_2' \\ \frac{1}{2}\sigma_{11|2}^{-2}\left[\left(y_1 - \mathbf{y}_2\boldsymbol{\alpha} - \mathbf{z}_1\delta_1 - \mathbf{v}_2\eta\right)^2 - \sigma_{11|2}\right] \\ vec\left(\mathbf{z}'\left(\mathbf{y}_2 - \mathbf{z}\boldsymbol{\delta}_2\right)\Sigma_{22}^{-1}\right) \\ \frac{1}{2}L_r vec\left(\Sigma_{22}^{-1}\mathbf{v}_{i2}\left(\boldsymbol{\delta}_2\right)'\mathbf{v}_{i2}\left(\boldsymbol{\delta}_2\right)\Sigma_{22}^{-1} - \Sigma_{22}^{-1}\right) \end{bmatrix} = 0 \qquad \text{(A.3)}$$

If $\eta = 0$, the result is trivial. Suppose that there exists at least one nonzero element of η. Also, assume over-identification. By substitution of $\mathbf{y}_2$ and an invertible linear transformation, these can be equivalently expressed as

$$E\begin{bmatrix} \left(y_1-\mathbf{y}_2\boldsymbol{\alpha}-\mathbf{z}_1\delta_1-\mathbf{v}_2\eta\right)\delta_{22}'\mathbf{z}_2' \\ \left(y_1-\mathbf{y}_2\boldsymbol{\alpha}-\mathbf{z}_1\delta_1-\mathbf{v}_2\eta\right)\mathbf{z}_1' \\ \left(y_1-\mathbf{y}_2\boldsymbol{\alpha}-\mathbf{z}_1\delta_1-\mathbf{v}_2\eta\right)\mathbf{v}_2' \\ \left(y_1-\mathbf{y}_2\boldsymbol{\alpha}-\mathbf{z}_1\delta_1-\mathbf{v}_2\eta\right)^2-\sigma_{11|2} \\ vec\left(\mathbf{z}'\left(\mathbf{y}_2-\mathbf{z}\boldsymbol{\delta}_2\right)\Sigma_{22}^{-1}\right)-vec\left(\mathbf{z}'\sigma_{11|2}^{-1}\left(y_1-\mathbf{y}_2\boldsymbol{\alpha}-\mathbf{z}_1\delta_1-\mathbf{v}_2\eta\right)\eta'\right) \\ L_r vec\left(\Sigma_{22}^{-1}\mathbf{v}_{i2}\left(\boldsymbol{\delta}_2\right)'\mathbf{v}_{i2}\left(\boldsymbol{\delta}_2\right)\Sigma_{22}^{-1}-\Sigma_{22}^{-1}\right) \end{bmatrix}=0 \qquad (A.4)$$

and

$$E\begin{bmatrix} \left(y_1-\mathbf{y}_2\boldsymbol{\alpha}-\mathbf{z}_1\delta_1-\mathbf{v}_2\eta\right)\delta_{22}'\mathbf{z}_2' \\ \left(y_1-\mathbf{y}_2\boldsymbol{\alpha}-\mathbf{z}_1\delta_1-\mathbf{v}_2\eta\right)\mathbf{z}_1' \\ \left(y_1-\mathbf{y}_2\boldsymbol{\alpha}-\mathbf{z}_1\delta_1-\mathbf{v}_2\eta\right)\mathbf{v}_2' \\ \left(y_1-\mathbf{y}_2\boldsymbol{\alpha}-\mathbf{z}_1\delta_1-\mathbf{v}_2\eta\right)^2-\sigma_{11|2} \\ vec\left(\mathbf{z}'\left(\mathbf{y}_2-\mathbf{z}\boldsymbol{\delta}_2\right)\Sigma_{22}^{-1}\right) \\ L_r vec\left(\Sigma_{22}^{-1}\mathbf{v}_{i2}\left(\boldsymbol{\delta}_2\right)'\mathbf{v}_{i2}\left(\boldsymbol{\delta}_2\right)\Sigma_{22}^{-1}-\Sigma_{22}^{-1}\right) \end{bmatrix}=0 \qquad (A.5)$$

respectivley. We can show (A.4) and (A.5) can be transformed by an invertible linear transformation into following expressions

$$E\begin{bmatrix} \left(y_1-\mathbf{y}_2\boldsymbol{\alpha}-\mathbf{z}_1\delta_1\right)\delta_{22}'\mathbf{z}_2' \\ \left(y_1-\mathbf{y}_2\boldsymbol{\alpha}-\mathbf{z}_1\delta_1\right)\mathbf{z}_1' \\ \left(y_1-\mathbf{y}_2\boldsymbol{\alpha}-\mathbf{z}_1\delta_1-\mathbf{v}_2\eta\right)\mathbf{v}_2' \\ \left(y_1-\mathbf{y}_2\boldsymbol{\alpha}-\mathbf{z}_1\delta_1-\mathbf{v}_2\eta\right)^2-\sigma_{11|2} \\ vec\left(\mathbf{z}_1'\left(\mathbf{y}_2-\mathbf{z}\boldsymbol{\delta}_2\right)\Sigma_{22}^{-1}\right) \\ vec\left(\mathbf{z}_2'\left(\mathbf{y}_2-\mathbf{z}\boldsymbol{\delta}_2\right)\Sigma_{22}^{-1}\right)-vec\left(\mathbf{z}_2'\sigma_{11|2}^{-1}\left(y_1-\mathbf{y}_2\boldsymbol{\alpha}-\mathbf{z}_1\delta_1-\mathbf{v}_2\eta\right)\eta'\right) \\ L_r vec\left(\Sigma_{22}^{-1}\mathbf{v}_{i2}\left(\boldsymbol{\delta}_2\right)'\mathbf{v}_{i2}\left(\boldsymbol{\delta}_2\right)\Sigma_{22}^{-1}-\Sigma_{22}^{-1}\right) \end{bmatrix}=0 \qquad (A.6)$$

and

$$E\begin{bmatrix} \left(y_1-\mathbf{y}_2\boldsymbol{\alpha}-\mathbf{z}_1\delta_1\right)\delta_{22}'\mathbf{z}_2' \\ \left(y_1-\mathbf{y}_2\boldsymbol{\alpha}-\mathbf{z}_1\delta_1\right)\mathbf{z}_1' \\ \left(y_1-\mathbf{y}_2\boldsymbol{\alpha}-\mathbf{z}_1\delta_1-\mathbf{v}_2\eta\right)\mathbf{v}_2' \\ \left(y_1-\mathbf{y}_2\boldsymbol{\alpha}-\mathbf{z}_1\delta_1-\mathbf{v}_2\eta\right)^2-\sigma_{11|2} \\ vec\left(\mathbf{z}'\left(\mathbf{y}_2-\mathbf{z}\boldsymbol{\delta}_2\right)\Sigma_{22}^{-1}\right) \\ L_r vec\left(\Sigma_{22}^{-1}\mathbf{v}_{i2}\left(\boldsymbol{\delta}_2\right)'\mathbf{v}_{i2}\left(\boldsymbol{\delta}_2\right)\Sigma_{22}^{-1}-\Sigma_{22}^{-1}\right) \end{bmatrix}=0 \qquad \text{(A.7)}$$

repectively. Then the result follows by Proposition 1.3.11 and by similar argument as in Lemma C.1. Equivalence in CF case is clear since

$$E\left[vec\left(\mathbf{z}'\left(\mathbf{y}_2-\mathbf{z}\boldsymbol{\delta}_2\right)\Sigma_{22}^{-1}\right)\right]=0$$

implies

$$E\left[\begin{bmatrix} \delta_{22}'\mathbf{z}_2' \\ \mathbf{z}_1' \end{bmatrix}\mathbf{v}_2\eta\right]=0.$$

Too see (A.4) implies (A.6), note the second $k_2 r$ part of the fifth moments implies

$$E\left[vec\left(\delta_{22}'\mathbf{z}_2'\left(\mathbf{y}_2-\mathbf{z}\boldsymbol{\delta}_2\right)\Sigma_{22}^{-1}\right)-vec\left(\delta_{22}'\mathbf{z}_2'\sigma_{11|2}^{-1}\left(y_1-\mathbf{y}_2\boldsymbol{\alpha}-\mathbf{z}_1\delta_1-\mathbf{v}_2\eta\right)\eta'\right)\right]=0$$

and, by adding the first moments after multiplying $\sigma_{11|2}^{-1}$ and elements of η, we have

$$E\left[vec\left(\delta_{22}'\mathbf{z}_2'\left(\mathbf{y}_2-\mathbf{z}\boldsymbol{\delta}_2\right)\Sigma_{22}^{-1}\right)\right]=0$$

which implies $E\left[\delta_{22}'\mathbf{z}_2'\underbrace{\mathbf{v}_2\eta}_{\text{scalar}}\right]=0$. Similarly, by adding second moment after multiplying $\sigma_{11|2}^{-1}$ and elements of η to the first $k_1 r$ part of the fifth moments, we have

$$E\left[vec\left(\mathbf{z}_1'\left(\mathbf{y}_2-\mathbf{z}\boldsymbol{\delta}_2\right)\Sigma_{22}^{-1}\right)\right]=0$$

and it implies $E\left[\mathbf{z}_1'\mathbf{v}_2\eta\right]=0$. The converse can be shown as following:

$$\begin{aligned} &E\left[vec\left(\delta_{22}^t\mathbf{z}_2^t\left(\mathbf{y}_2-\mathbf{z}\boldsymbol{\delta}_2\right)\Sigma_{22}^{-1}\right)-vec\left(\delta_{22}^t\mathbf{z}_2^t\sigma_{11|2}^{-1}\left(y_1-\mathbf{y}_2\boldsymbol{\alpha}-\mathbf{z}_1\delta_1-\mathbf{v}_2\eta\right)\eta'\right)\right] \\ &=E\left[vec\left(\delta_{22}^t\mathbf{z}_2^t\left(\mathbf{y}_2-\mathbf{z}\boldsymbol{\delta}_2\right)\Sigma_{22}^{-1}\right)-vec\left(\delta_{22}^t\mathbf{z}_2^t\sigma_{11|2}^{-1}\left(-\mathbf{v}_2\eta\right)\eta'\right)\right] \\ &=E\left[vec\left(\delta_{22}^t\mathbf{z}_2^t\mathbf{v}_2\Sigma_{22}^{-1}\right)+vec\left(\delta_{22}^t\mathbf{z}_2^t\mathbf{v}_2\eta\eta'\right)\sigma_{11|2}^{-1}\right]=0 \end{aligned}$$

multiplying $\Sigma_{22}\eta$ from right, we have

$$E\left[vec\left(\delta_{22}^{t}\mathbf{z}_{2}^{t}\mathbf{v}_{2}\eta\right)+vec\left(\delta_{22}^{t}\mathbf{z}_{2}^{t}\mathbf{v}_{2}\eta\qquad\underbrace{\eta'\Sigma_{22}\eta}_{\text{strict positive scalar if }\eta\neq0}\sigma_{11|2}^{-1}\right)\right]=0$$

which implies

$$E\left[\delta_{22}^{t}\mathbf{z}_{2}^{t}\mathbf{v}_{2}\eta\right]=0$$

And, again, seeing $E\left[vec\left(\mathbf{z}_{1}'\left(\mathbf{y}_{2}-\mathbf{z}\boldsymbol{\delta}_{2}\right)\Sigma_{22}^{-1}\right)\right]=0$ implies $E\left[\mathbf{z}_{1}'\mathbf{v}_{2}\eta\right]=0$ delivers the result. And invertibility of $E\left[\frac{\partial m_{i2}(\zeta_{o1},\zeta_{o2})}{\partial\zeta_{2}'}\right]$ and $E\left[\frac{\partial m_{i3}(\zeta_{o1},\zeta_{o2})}{\partial\zeta_{2}'}\right]$ can be easily derived from identification conditions from LIML and CF. Hence, it is shown that there exist such $T_1(\theta)$ and $T_2(\theta)$ in Proposition 1.3.10.

A.2.7 Proof of Proposition 1.3.12

Lemma H.1 below proves the results when $g_1(w,\theta,\lambda)$ and $g_2(w,\theta,\lambda)$ are taken properly:

$$\text{QLIML}: g_2=\begin{bmatrix}\frac{\partial q_1}{\partial\lambda_1'}\\ \frac{\partial q_1}{\partial\lambda_2'}+\frac{\partial q_2}{\partial\lambda_2'}\end{bmatrix},\quad \text{CF}: g_2=\begin{bmatrix}\frac{\partial q_1}{\partial\lambda_1'}\\ \frac{\partial q_2}{\partial\lambda_2'}\end{bmatrix},\quad \text{GMM-QLIML}: g_2=\begin{bmatrix}\frac{\partial q_1}{\partial\lambda_1'}\\ \frac{\partial q_2}{\partial\lambda_2'}\end{bmatrix}$$

with g_1 chosen to be the rest of moment functions in each GMM-interpreted estimator.

Lemma A.2.2 Let $\theta\in\mathbb{R}^p$, $\lambda\in\mathbb{R}^r$ and $g=\left(g_1(w,\theta,\lambda)',g_2(w,\theta,\lambda)'\right)'$ be $\mathbb{R}^{q+r}-$valued moment functions with $q\geq p$. Assume regularity conditions for well-definedness of relevant GMM estimators below. Suppose

$$E\begin{bmatrix}\frac{\partial g_1(\theta_o,\lambda_o)}{\partial(\theta',\lambda')}\\ \frac{\partial g_2(\theta_o,\lambda_o)}{\partial(\theta',\lambda')}\end{bmatrix}=\begin{bmatrix}G_{q\times p}^{11} & 0_{q\times r}\\ 0_{r\times p} & G_{r\times r}^{22}\end{bmatrix}$$

where both G^{22} and $V(g(w,\theta_o,\lambda_o))$ are invertible. Then, the asymptotic variance of optimal GMM estimator of θ based on $\left(g_1(\theta,\lambda)',g_2(\theta,\lambda)'\right)'$ is the same as that of optimal GMM estimator of θ based on $g_1(\theta,\lambda_o)$ treating λ_o as a known value.

Proof. Let

$$G = \begin{bmatrix} G^{11}_{q\times p} & 0_{q\times r} \\ 0_{r\times p} & G^{22}_{r\times r} \end{bmatrix}$$

$$V^{-1} = V\begin{pmatrix} g_1(w,\theta_o,\lambda_o) \\ g_2(w,\theta_o,\lambda_o) \end{pmatrix}^{-1} = \begin{bmatrix} V_{11,q\times q} & V_{12,q\times r} \\ V_{21,r\times q} & V_{22,r\times r} \end{bmatrix}^{-1} = \begin{bmatrix} B^{1}_{q\times q} & B^{12}_{q\times r} \\ B^{21}_{r\times q} & B^{22}_{r\times r} \end{bmatrix}$$

Then

$$\begin{aligned} G'V^{-1}G &= \begin{bmatrix} G^{11}_{q\times p} & 0_{q\times r} \\ 0_{r\times p} & G^{22}_{r\times r} \end{bmatrix}' \begin{bmatrix} B^{1}_{q\times q} & B^{12}_{q\times r} \\ B^{21}_{r\times q} & B^{22}_{r\times r} \end{bmatrix} \begin{bmatrix} G^{11}_{q\times p} & 0_{q\times r} \\ 0_{r\times p} & G^{22}_{r\times r} \end{bmatrix} \\ &= \begin{bmatrix} G^{11\prime}B^{1} & G^{11\prime}B^{12} \\ G^{22\prime}B^{21} & G^{22\prime}B^{22} \end{bmatrix} \begin{bmatrix} G^{11}_{q\times p} & 0_{q\times r} \\ 0_{r\times p} & G^{22}_{r\times r} \end{bmatrix} \\ &= \begin{bmatrix} G^{11\prime}B^{1}G^{11} & G^{11\prime}B^{12}G^{22} \\ G^{22\prime}B^{21}G^{11} & G^{22\prime}B^{22}G^{22} \end{bmatrix} \end{aligned}$$

Now, it suffices to show that

$$\left(G^{11\prime}B^{1}G^{11} - G^{11\prime}B^{12}G^{22}\left[G^{22\prime}B^{22}G^{22}\right]^{-1}G^{22\prime}B^{21}G^{11}\right)^{-1} = \left(G^{11\prime}V_{11}^{-1}G^{11}\right)^{-1}$$

This is true since

$$\begin{aligned} &G^{11\prime}B^{1}G^{11} - G^{11\prime}B^{12}G^{22}\left[G^{22\prime}B^{22}G^{22}\right]^{-1}G^{22\prime}B^{21}G^{11} \\ &= G^{11\prime}\left[B^{1} - B^{12}G^{22}\left(G^{22}\right)^{-1}\left(B^{22}\right)^{-1}\left(G^{22\prime}\right)^{-1}G^{22\prime}B^{21}\right]G^{11} \\ &= G^{11\prime}\left[B^{1} - B^{12}\left(B^{22}\right)^{-1}B^{21}\right]G^{11} \\ &= G^{11\prime}V_{11}^{-1}G^{11} \end{aligned}$$

■

A.2.8 Proof of Proposition 1.3.13

Define Schur complements as

$$A/\left(A_{22}+C_{22}\right) \equiv A_{11}-A_{12}\left(A_{22}+C_{22}\right)^{-1} A_{21}$$

$$A/A_{11} \equiv A_{22}+C_{22}-A_{21}A_{11}^{-1}A_{12}$$

where $A=\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22}+C_{22} \end{bmatrix}$, $A_{11}=E\left[-\frac{\partial q_1}{\partial\theta_1\partial\theta_1'}\right]$, $A_{12}=A_{21}^t=E\left[-\frac{\partial q_1}{\partial\theta_1\partial\theta_2'}\right]$, $A_{22}=E\left[-\frac{\partial q_1}{\partial\theta_2\partial\theta_2'}\right]$ and $C_{22}=E\left[-\frac{\partial q_2}{\partial\theta_2\partial\theta_2'}\right]$.

Assume GIMEs i.e. $V\left(\frac{\partial q_1}{\partial\theta_1}\right)=\tau_1 A_{11}$, $cov\left(\frac{\partial q_1}{\partial\theta_1},\frac{\partial q_1}{\partial\theta_2}\right)=\tau_1 A_{12}$, $V\left(\frac{\partial q_1}{\partial\theta_2}\right)=\tau_1 A_{22}$, $V\left(\frac{\partial q_2}{\partial\theta_2}\right)=\tau_2 C_{22}$ and $cov\left(\frac{\partial q_1}{\partial\theta},\frac{\partial q_2}{\partial\theta_2}\right)=0$. Then, what needs to be shown is

$$V_{CF}^{\theta_1}-V_{QLIML}^{\theta_1}=A_{11}^{-1}A_{12}\left[\tau_2 W_1+\left(\tau_1-\tau_2\right)W_2\right]A_{21}A_{11}^{-1}$$

where

$$W_1=C_{22}^{-1}-\left[A/A_{11}\right]^{-1}$$

$$W_2=-\left[A/A_{11}\right]^{-1}\left(A_{22}-A_{21}A_{11}^{-1}A_{12}\right)\left[A/A_{11}\right]^{-1}$$

First, by argument used in the proof of Proposition 1.3.10 (i), the variance difference is

$$\begin{aligned} & V_{CF}^{\theta_1}-V_{QLIML}^{\theta_1} \\ & =A_{11}^{-1}B_2A_{11}^{-1}-\left[A_{11}-A_{12}\left(A_{22}+C_{22}\right)^{-1}A_{21}\right]^{-1}B_1\left[A_{11}-A_{12}\left(A_{22}+C_{22}\right)^{-1}A_{21}\right]^{-1} \end{aligned}$$

where

$$\begin{aligned} B_1 &= V\left[\frac{\partial q_1}{\partial\theta_1}-A_{12}\left(A_{22}+C_{22}\right)^{-1}\left(\frac{\partial q_1}{\partial\theta_2}+\frac{\partial q_2}{\partial\theta_2}\right)\right] \\ &= \tau_1 A_{11}+A_{12}\left(A_{22}+C_{22}\right)^{-1}\left(\tau_1 A_{22}+\tau_2 C_{22}\right)\left(A_{22}+C_{22}\right)^{-1}A_{21} \\ &\quad -2\tau_1 A_{12}\left(A_{22}+C_{22}\right)^{-1}A_{21} \\ &= \tau_1 A_{11}+A_{12}\left(A_{22}+C_{22}\right)^{-1}\left(\tau_1 A_{22}+\tau_1 C_{22}-\tau_1 C_{22}+\tau_2 C_{22}\right)\left(A_{22}+C_{22}\right)^{-1}A_{21} \\ &\quad -2\tau_1 A_{12}\left(A_{22}+C_{22}\right)^{-1}A_{21} \end{aligned}$$

$$
\begin{aligned}
&= \tau_1 \left[A_{11} - A_{12} \left(A_{22} + C_{22} \right)^{-1} A_{21} \right] \\
&+ \left(\tau_2 - \tau_1 \right) A_{12} \left(A_{22} + C_{22} \right)^{-1} C_{22} \left(A_{22} + C_{22} \right)^{-1} A_{21}
\end{aligned}
$$

and

$$
\begin{aligned}
B_2 &= V \left[\frac{\partial q_1^o}{\partial \theta_1} - A_{12} C_{22}^{-1} \frac{\partial q_2^o}{\partial \theta_2} \right] \\
&= \tau_1 A_{11} + \tau_2 A_{12} C_{22}^{-1} A_{21}
\end{aligned}
$$

Since

$$
\begin{aligned}
&\left[A_{11} - A_{12} \left(A_{22} + C_{22} \right)^{-1} A_{21} \right]^{-1} \\
&= A_{11}^{-1} + A_{11}^{-1} A_{12} \left(\left[A_{22} + C_{22} \right] - A_{21} A_{11}^{-1} A_{12} \right)^{-1} A_{21} A_{11}^{-1}
\end{aligned}
$$

the difference can be rearranged as

$$
\begin{aligned}
&A_{11}^{-1} B_2 A_{11}^{-1} - \left[A_{11} - A_{12} \left(A_{22} + C_{22} \right)^{-1} A_{21} \right]^{-1} B_1 \left[A_{11} - A_{12} \left(A_{22} + C_{22} \right)^{-1} A_{21} \right]^{-1} \\
&= A_{11}^{-1} \left(\tau_1 A_{11} + \tau_2 A_{12} C_{22}^{-1} A_{21} \right) A_{11}^{-1} - \tau_1 \left[A_{11} - A_{12} \left(A_{22} + C_{22} \right)^{-1} A_{21} \right]^{-1} \\
&- \left(\tau_2 - \tau_1 \right) \left[A_{11} - A_{12} \left(A_{22} + C_{22} \right)^{-1} A_{21} \right]^{-1} A_{12} D A_{21} \left[A_{11} - A_{12} \left(A_{22} + C_{22} \right)^{-1} A_{21} \right]^{-1} \\
&= A_{11}^{-1} A_{12} \left(\tau_2 C_{22}^{-1} - \tau_1 \left(\left[A_{22} + C_{22} \right] - A_{21} A_{11}^{-1} A_{12} \right)^{-1} \right) A_{21} A_{11}^{-1} \\
&+ \left(\tau_1 - \tau_2 \right) \left[A_{11} - A_{12} \left(A_{22} + C_{22} \right)^{-1} A_{21} \right]^{-1} A_{12} D A_{21} \left[A_{11} - A_{12} \left(A_{22} + C_{22} \right)^{-1} A_{21} \right]^{-1} \\
&= \tau_2 A_{11}^{-1} A_{12} \left(C_{22}^{-1} - \left(\left[A_{22} + C_{22} \right] - A_{21} A_{11}^{-1} A_{12} \right)^{-1} \right) A_{21} A_{11}^{-1} \\
&+ \left(\tau_1 - \tau_2 \right) \left[A_{11} - A_{12} \left(A_{22} + C_{22} \right)^{-1} A_{21} \right]^{-1} A_{12} D A_{21} \left[A_{11} - A_{12} \left(A_{22} + C_{22} \right)^{-1} A_{21} \right]^{-1} \\
&- \left(\tau_1 - \tau_2 \right) A_{11}^{-1} A_{12} \left(\left[A_{22} + C_{22} \right] - A_{21} A_{11}^{-1} A_{12} \right)^{-1} A_{21} A_{11}^{-1}
\end{aligned}
$$

where $D = \left(A_{22} + C_{22} \right)^{-1} C_{22} \left(A_{22} + C_{22} \right)^{-1}$. Now, it suffices to show

$$
\begin{aligned}
&\left[A_{11} - A_{12} \left(A_{22} + C_{22} \right)^{-1} A_{21} \right]^{-1} A_{12} D A_{21} \left[A_{11} - A_{12} \left(A_{22} + C_{22} \right)^{-1} A_{21} \right]^{-1} \\
&- A_{11}^{-1} A_{12} \left(\left[A_{22} + C_{22} \right] - A_{21} A_{11}^{-1} A_{12} \right)^{-1} A_{21} A_{11}^{-1} \\
&= -A_{11}^{-1} A_{12} \left[A / A_{11} \right]^{-1} \left(A_{22} - A_{21} A_{11}^{-1} A_{12} \right) \left[A / A_{11} \right]^{-1} A_{21} A_{11}^{-1}
\end{aligned}
$$

The first term:

$$\left[A_{11}-A_{12}\left(A_{22}+C_{22}\right)^{-1}A_{21}\right]^{-1}A_{12}DA_{21}\left[A_{11}-A_{12}\left(A_{22}+C_{22}\right)^{-1}A_{21}\right]^{-1}$$
$$=A_{11}^{-1}\left[A_{11}+A_{12}\left[A/A_{11}\right]^{-1}A_{21}\right]A_{11}^{-1}A_{12}DA_{21}A_{11}^{-1}\left[A_{11}+A_{12}\left[A/A_{11}\right]^{-1}A_{21}\right]A_{11}^{-1}$$
$$=A_{11}^{-1}\left[A_{11}A_{11}^{-1}A_{12}+A_{12}\left[A/A_{11}\right]^{-1}A_{21}A_{11}^{-1}A_{12}\right]$$
$$\times D\left[A_{21}A_{11}^{-1}A_{11}+A_{21}A_{11}^{-1}A_{12}\left[A/A_{11}\right]^{-1}A_{21}\right]A_{11}^{-1}$$
$$=A_{11}^{-1}A_{12}\left[I+\left[A/A_{11}\right]^{-1}A_{21}A_{11}^{-1}A_{12}\right]D\left[I+A_{21}A_{11}^{-1}A_{12}\left[A/A_{11}\right]^{-1}\right]A_{21}A_{11}^{-1}$$
$$=A_{11}^{-1}A_{12}\left[A/A_{11}\right]^{-1}\left[\left[A/A_{11}\right]+A_{21}A_{11}^{-1}A_{12}\right]D\left[\left[A/A_{11}\right]+A_{21}A_{11}^{-1}A_{12}\right]\left[A/A_{11}\right]^{-1}A_{21}A_{11}^{-1}$$
$$=A_{11}^{-1}A_{12}\left[A/A_{11}\right]^{-1}\left[A_{22}+C_{22}\right]D\left[A_{22}+C_{22}\right]\left[A/A_{11}\right]^{-1}A_{21}A_{11}^{-1}$$
$$=A_{11}^{-1}A_{12}\left[A/A_{11}\right]^{-1}C_{22}\left[A/A_{11}\right]^{-1}A_{21}A_{11}^{-1}$$

The second term:

$$A_{11}^{-1}A_{12}\left(\left[A_{22}+C_{22}\right]-A_{21}A_{11}^{-1}A_{12}\right)^{-1}A_{21}A_{11}^{-1}$$
$$=A_{11}^{-1}A_{12}\left[A/A_{11}\right]^{-1}\left[A/A_{11}\right]\left[A/A_{11}\right]^{-1}A_{21}A_{11}^{-1}$$

Hence the result.

To see positive semi-definiteness of W_1,

$$C_{22}^{-1}-\left[A/A_{11}\right]^{-1}\succeq 0$$

$\Longleftrightarrow$

$$\left[A/A_{11}\right]-C_{22}$$
$$=A_{22}-A_{21}A_{11}^{-1}A_{12}\succeq 0$$

$\underset{\text{under } A_{11}\succ 0}{\Longleftrightarrow}$

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}\succeq 0$$

where the last equivalence is from Schur complement condition for positive semi-definiteness. Since

$$A_{11} = \frac{1}{\tau_1} V_o \left(\frac{\partial q_1^o}{\partial \theta_1} \right) \succ 0 \quad \left(\because \frac{\partial q_1}{\partial \theta_1} \text{ linearly independent at } \theta_o \right)$$

and

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \frac{1}{\tau_1} V_o \left(\frac{\partial q_1^o}{\partial \theta} \right) \succeq 0$$

the statement is shown. Negative semi-definiteness of W_2 is also proved similarly.

A.2.9 Proof of Proposition 1.3.14

Note that

$$E \left[\frac{\partial q_1^o}{\partial \theta_1 \partial vech \left(\Sigma_{22} \right)'} \right] = 0$$
$$E \left[\frac{\partial q_2^o}{\partial vec \left(\boldsymbol{\delta}_2 \right) \partial vech \left(\Sigma_{22} \right)'} \right] = 0$$

Thus we can treat Σ_{22} as a known value by Proposition 1.3.12. Then, redefined $\tilde{q}_2$ is also a member of linear exponential family (Gouriéroux, Monfort and Trognon ,1984). Now, it suffices to show GLM variance assumptions imply GIMEs with corresponding scaling factor in linear exponential family. Let $\mathbf{m}(\theta) \equiv G(\mathbf{y}_2, \mathbf{z}_1, \mathbf{v}_2, \theta_1)$. Based on general form of score and Hessian (Wooldridge, 2010), it is easy to see

$$\begin{aligned}
E_o \left[-\frac{\partial q_1}{\partial \theta \partial \theta'} \right] &= E_o \left[\frac{1}{V_q \left(y_1 | \mathbf{y}_2, \mathbf{z} \right)} \frac{\partial \mathbf{m}(\theta)}{\partial \theta} \frac{\partial \mathbf{m}(\theta)}{\partial \theta'} \right] \\
E_o \left[\frac{\partial q_1}{\partial \theta} \frac{\partial q_1}{\partial \theta'} \right] &= E_o \left[\frac{\left[y_1 - \mathbf{m}(\theta) \right]^2}{V_q \left(y_1 | \mathbf{y}_2, \mathbf{z} \right)^2} \frac{\partial \mathbf{m}(\theta)}{\partial \theta} \frac{\partial \mathbf{m}(\theta)}{\partial \theta'} \right] \\
&= E_o \left[\frac{E_o \left[\left[y_1 - \mathbf{m}(\theta) \right]^2 \middle| \mathbf{y}_2, \mathbf{z} \right]}{V_q \left(y_1 | \mathbf{y}_2, \mathbf{z} \right)^2} \frac{\partial \mathbf{m}(\theta)}{\partial \theta} \frac{\partial \mathbf{m}(\theta)}{\partial \theta'} \right] \\
&= \tau_1 E_o \left[-\frac{\partial q_1}{\partial \theta \partial \theta'} \right]
\end{aligned}$$

And GIMEs for $\tilde{q}_2$ can be shown similarly.

$$\begin{aligned}
E_o\left[-\frac{\partial \tilde{q}_2}{\partial vec\left(\boldsymbol{\delta}_2\right)\partial vec\left(\boldsymbol{\delta}_2\right)'}\right] &= E_o\left[\Sigma_{22}^{-1}\otimes \mathbf{z}'\mathbf{z}\right] \\
V_o\left(\frac{\partial \tilde{q}_2}{\partial vec\left(\boldsymbol{\delta}_2\right)}\right) &= E_o\left[\left[I_r\otimes \mathbf{z}'\right]\Sigma_{22}^{-1}\mathbf{v}_2'\mathbf{v}_2\Sigma_{22}^{-1}\left[I_r\otimes \mathbf{z}\right]\right] \\
&= E_o\left[\left[I_r\otimes \mathbf{z}'\right]\left[\Sigma_{22}^{-1}E_o\left[\mathbf{v}_2'\mathbf{v}_2|\mathbf{z}\right]\Sigma_{22}^{-1}\otimes \mathbf{z}\right]\right] \\
&= \tau_2 E_o\left[\left[I_r\otimes \mathbf{z}'\right]\left[\Sigma_{22}^{-1}\otimes \mathbf{z}\right]\right] \\
&= \tau_2 E_o\left[-\frac{\partial \tilde{q}_2}{\partial vec\left(\boldsymbol{\delta}_2\right)\partial vec\left(\boldsymbol{\delta}_2\right)'}\right]
\end{aligned}$$

Orthogonality of scores holds under correct specification of conditional means since

$$\begin{aligned}
E_o\left[\frac{\partial q_1}{\partial \theta}\frac{\partial q_2}{\partial \theta_2'}\right] &= E_o\left[E_o\left[\left.\frac{\partial q_1}{\partial \theta}\right|\mathbf{y}_2,\mathbf{z}\right]\frac{\partial q_2}{\partial \theta_2'}\right] \\
&= E_o\left[\frac{E_o\left[y_1-\mathbf{m}\left(\theta\right)|\,\mathbf{y}_2,\mathbf{z}\right]}{V_q\left(y_1|\,\mathbf{y}_2,\mathbf{z}\right)}\frac{\partial \mathbf{m}\left(\theta\right)}{\partial \theta}\frac{\partial q_2}{\partial \theta_2'}\right] \\
&= 0
\end{aligned}$$

Then, by Proposition 1.3.13, QLIML is efficient relative to CF for θ_1.

APPENDIX B

AN APPENDIX FOR CHAPTER 2

B.1 Regularity conditions

Assumptions (consistency and asymptotic normality of QLIML and CF estimator)

(1) $(\mathbf{y}_{i1}, \mathbf{y}_{i2}, \mathbf{z}_i)$ are i.i.d.

(2) $\Theta \underset{cpt}{\subset} \mathbb{R}^p$

(3) $q_1 : \Theta \times W \to \mathbb{R}$ and $q_2 : \Theta_2 \times W \to \mathbb{R}$ where $(y_{i1}, \mathbf{y}_{i2}, \mathbf{z}_i) \in W$

(4) $\theta_o \in int\,(\Theta)$ and let $\mathcal{N}$ be a neighborhood of θ_o

(5) $q_{i1}\,(\theta_1, \theta_2)$ and $q_{i2}\,(\theta_2)$ are continuously differentiable at each $\theta \in \Theta$ with probability one.

(6) $E\left[\sup_{\theta\in\Theta}\left\| \begin{array}{c} \frac{\partial(q_1+q_2)}{\partial\theta} \\ \frac{\partial q_2}{\partial\theta_2} \end{array} \right\|\right] < \infty$ (7) $E\left[\left\| \begin{array}{c} \frac{\partial q_1(\theta_o)}{\partial\theta} \\ \frac{\partial q_2(\theta_{o2})}{\partial\theta_2} \end{array} \right\|^2\right] < \infty$

(8) $E\left[\begin{array}{c} \frac{\partial(q_1(\theta)+q_2(\theta_2))}{\partial\theta} \\ \frac{\partial q_2(\theta_2)}{\partial\theta_2} \end{array} \right]$ is differentiable with respect to θ at θ_o

(9) (QLIML orthogonality) $\{\theta_o\} = \left\{\theta \in \Theta : E\left[\frac{\partial(q_1+q_2)}{\partial\theta}\right] = 0\right\}$

(10) (CF orthogonality) $\{\theta_o\} = \left\{\theta \in \Theta : E\left[\begin{array}{c} \frac{\partial q_1(\theta)}{\partial\theta_1} \\ \frac{\partial q_2(\theta_2)}{\partial\theta_2} \end{array} \right] = 0\right\}$

(11) (QLIML rank) $\frac{\partial}{\partial\theta'}E\left[\frac{\partial(q_1(\theta)+q_2(\theta))}{\partial\theta}\right]\Bigg|_{\theta=\theta_o}$ and $V\left[\frac{\partial(q_1(\theta)+q_2(\theta))}{\partial\theta}\Bigg|_{\theta=\theta_o}\right]$ are invertible.

(12) (CF rank) $\frac{\partial}{\partial\theta'}E\left[\begin{array}{c} \frac{\partial q_1(\theta)}{\partial\theta_1} \\ \frac{\partial q_2(\theta)}{\partial\theta_2} \end{array} \right]\Bigg|_{\theta=\theta_o}$ and $V\left[\begin{array}{c} \frac{\partial q_1(\theta_o)}{\partial\theta_1} \\ \frac{\partial q_2(\theta_{o2})}{\partial\theta_2} \end{array} \right]$ are invertible.

(13) (stochastic differentiability) For $j = 1, 2$ and any $\delta_N \to 0$,

$$\sup_{\|\theta-\theta_o\|\le\delta_N} \frac{\sqrt{N}\left\|\hat{g}_N^j\,(\theta) - \hat{g}_N^j\,(\theta_o) - E\left[\hat{g}_N^j\,(\theta)\right]\right\|}{1+\sqrt{N}\,\|\theta-\theta_o\|} \xrightarrow{p} 0$$

where

$$\hat{g}_N^1(\theta) = \sum_{i=1}^{N} \left[\frac{\partial (q_{i1}(\theta) + q_{i2}(\theta))}{\partial \theta} \right] \text{ and } \hat{g}_N^2(\theta) = \sum_{i=1}^{N} \begin{bmatrix} \frac{\partial q_{i1}(\theta)}{\partial \theta_1} \\ \frac{\partial q_{i2}(\theta)}{\partial \theta_2} \end{bmatrix}$$

Assumptions (consistency and asymptotic normality of minimum distance estimators)

(14) $\gamma(\theta_1, \theta_{o2}) = \gamma_o$ if and only if $\theta_1 = \theta_{o1}$

(15) $\frac{\partial \gamma(\theta_{o1}, \theta_{o2})}{\partial \theta_1}$ has full rank

(16) $\gamma : \Theta \to \Gamma$ is continuous at each $\theta \in \Theta$ and continuously differentiable in $\mathcal{N}$

(17) (asymptotic normality of reduced form parameters)

$$\sqrt{N}\left(\left(\hat{\gamma}', \hat{\theta}_2'\right)' - \left(\gamma_o', \theta_{o2}'\right)'\right) \xrightarrow{d} N\left(0, \left(A_R' B_R^{-1} A_R\right)^{-1}\right)$$

where

$$A_R = \frac{\partial}{\partial\left(\gamma', \theta_2'\right)} E \begin{bmatrix} \frac{\partial q_1(\gamma, \theta_2)}{\partial \gamma} \\ \frac{\partial q_2(\theta_2)}{\partial \theta_2} \end{bmatrix} \Bigg|_{(\gamma, \theta_2) = (\gamma_o, \theta_{o2})} \text{ and } B_R = V \begin{bmatrix} \frac{\partial q_1(\gamma_o, \theta_{o2})}{\partial \gamma} \\ \frac{\partial q_2(\theta_{o2})}{\partial \theta_2} \end{bmatrix}$$

Assumption (18) each component of $\frac{\partial q_2(\theta_{o2})}{\partial \theta_2}$ cannot be expressed as linear combination of $\frac{\partial q_1(\theta)}{\partial \theta}$.

B.2 Proof of Proposition 2.3.3

First, it is shown that a well-defined minimum distance estimator and its linearized version are asymptotically equivalent.

Lemma B.2.1 (linearized minimum distance estimator) Assume (1) $\gamma_o - h(\theta) \neq 0$ if $\theta \neq \theta_o$ (2) h is continuously differentiable in θ where $\theta \in \mathbb{R}^p$, $\gamma \in \mathbb{R}^g$ and $g \geq p$ (3) $\frac{\partial h}{\partial \theta}(\theta_o)$ has full column rank (4) $\sqrt{N}(\hat{\gamma} - \gamma_o) \xrightarrow{d} N(0, \Omega_o)$. Then, efficient MD on a linearized link function yields an estimator asymptotically equivalent to efficient MD with original link function.

Proof. Consider a first-order expansion of h around θ_o

$$h(\theta) \approx h(\theta_o) + \frac{\partial h(\theta_o)}{\partial \theta'}(\theta - \theta_o)$$

The minimization problem is

$$\min_{\theta} \left(\hat{\gamma} - h(\theta_o) - \frac{\partial h(\theta_o)}{\partial \theta'}(\theta - \theta_o)\right)' \hat{W} \left(\hat{\gamma} - h(\theta_o) - \frac{\partial h(\theta_o)}{\partial \theta'}(\theta - \theta_o)\right) \tag{A.1}$$

where

$$\hat{W} \xrightarrow{p} W_o$$

The first order condition is

$$\frac{\partial h(\theta_o)}{\partial \theta} \hat{W} \left(\hat{\gamma} - h(\theta_o) - \frac{\partial h(\theta_o)}{\partial \theta'}(\theta - \theta_o)\right) = 0$$

Then

$$\begin{aligned}
\hat{\theta} &= \left[\frac{\partial h(\theta_o)}{\partial \theta} \hat{W} \frac{\partial h(\theta_o)}{\partial \theta'}\right]^{-1} \frac{\partial h(\theta_o)}{\partial \theta} \hat{W} \left(\hat{\gamma} - h(\theta_o) + \frac{\partial h(\theta_o)}{\partial \theta'}\theta_o\right) \\
&= \theta_o + \left[\frac{\partial h(\theta_o)}{\partial \theta} \hat{W} \frac{\partial h(\theta_o)}{\partial \theta'}\right]^{-1} \frac{\partial h(\theta_o)}{\partial \theta} \hat{W} (\hat{\gamma} - h(\theta_o))
\end{aligned}$$

The asymptotic distribution is

$$\begin{aligned}
\sqrt{N}\left(\hat{\theta} - \theta_o\right) &= \left[\frac{\partial h(\theta_o)}{\partial \theta} \hat{W} \frac{\partial h(\theta_o)}{\partial \theta'}\right]^{-1} \frac{\partial h(\theta_o)}{\partial \theta} \hat{W} \sqrt{N} (\hat{\gamma} - h(\theta_o)) \\
&= \left[H_o' W_o H_o\right]^{-1} H_o' W_o \sqrt{N} (\hat{\gamma} - h(\theta_o)) + o_p(1) \\
&= \left[H_o' W_o H_o\right]^{-1} H_o' W_o \sqrt{N} (\hat{\gamma} - \gamma_o) + o_p(1)
\end{aligned}$$

Then the optimal weighting matrix is such that

$$W_o = \Omega_o^{-1}$$

Hence the proof ■

Now, consider an auxiliary model.

Lemma B.2.2 There exists an auxiliary asymptotic model whose GLS estimator is asymptotically equivalent to efficient linearized MD.

Proof. Consider

$$\underbrace{\hat{\gamma} - h(\theta_o) + \frac{\partial h(\theta_o)}{\partial \theta'}\theta_o}_{y} = \underbrace{\frac{\partial h(\theta_o)}{\partial \theta'}}_{X}\theta + u_n$$

where

$$E[u_n] = 0$$

$$V(u_n) = \Omega_o$$

Then GLS estimator of θ in this auxiliary asymptotic model is asymptotically equivalent to efficient linearized MD since GLS estimator solves (A.1). Here, mean and variance of u_n are just imposed restrictions in auxiliary model, not derived from original model. (Gouriéroux, Monfort, Trognon, 1985) ■

Next, above two lemma will be applied to a minimum distance problem with partitioned link function. Consider

$$\{(\theta_{o1}, \theta_{o2})\} = \left\{ (\theta_1, \theta_2) : \begin{bmatrix} \gamma_{o1} \\ \gamma_{o2} \end{bmatrix} = \begin{bmatrix} h_1(\theta_1, \theta_2) \\ h_2(\theta_1, \theta_2) \end{bmatrix} \right\}$$

where $\gamma = (\gamma_1', \gamma_2')'$, $h = (h_1', h_2')'$, $\gamma_1 \in \mathbb{R}^{g_1}$, $\gamma_2 \in \mathbb{R}^{g_2}$, $\theta_1 \in \mathbb{R}^{p_1}$, $\theta_2 \in \mathbb{R}^{p_2}$, and $p_1 = g_2$. Then, by first order expansion, a linearized partioned model is

$$\begin{bmatrix} h_1(\theta_1, \theta_2) \\ h_2(\theta_1, \theta_2) \end{bmatrix} \approx \begin{bmatrix} h_1(\theta_{o1}, \theta_{o2}) \\ h_2(\theta_{o1}, \theta_{o2}) \end{bmatrix} + \begin{bmatrix} \frac{\partial h_1(\theta_o)}{\partial \theta_1'} & \frac{\partial h_1(\theta_o)}{\partial \theta_2'} \\ \frac{\partial h_2(\theta_o)}{\partial \theta_1'} & \frac{\partial h_2(\theta_o)}{\partial \theta_2'} \end{bmatrix} \left(\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} - \begin{bmatrix} \theta_{o1} \\ \theta_{o2} \end{bmatrix} \right)$$

and the corresponding auxiliary asymptotic model is

$$\begin{aligned} &\begin{bmatrix} \hat{\gamma}_1 \\ \hat{\gamma}_2 \end{bmatrix} - \begin{bmatrix} h_1(\theta_{o1}, \theta_{o2}) \\ h_2(\theta_{o1}, \theta_{o2}) \end{bmatrix} + \begin{bmatrix} \frac{\partial h_1(\theta_o)}{\partial \theta_1'} & \frac{\partial h_1(\theta_o)}{\partial \theta_2'} \\ \frac{\partial h_2(\theta_o)}{\partial \theta_1'} & \frac{\partial h_2(\theta_o)}{\partial \theta_2'} \end{bmatrix} \begin{bmatrix} \theta_{o1} \\ \theta_{o2} \end{bmatrix} \\ &= \begin{bmatrix} \frac{\partial h_1(\theta_o)}{\partial \theta_1'} & \frac{\partial h_1(\theta_o)}{\partial \theta_2'} \\ \frac{\partial h_2(\theta_o)}{\partial \theta_1'} & \frac{\partial h_2(\theta_o)}{\partial \theta_2'} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} + \begin{bmatrix} u_{1n} \\ u_{2n} \end{bmatrix} \end{aligned}$$

where

$$E\begin{bmatrix} u_{1n} \\ u_{2n} \end{bmatrix} = 0$$

$$V\begin{bmatrix} u_{1n} \\ u_{2n} \end{bmatrix} = \Omega_o$$

Define

$$\begin{aligned}
y_1 &\equiv \hat{\gamma}_1 - h_1\left(\theta_{o1}, \theta_{o2}\right) + \frac{\partial h_1\left(\theta_o\right)}{\partial \theta_1'}\theta_{o1} + \frac{\partial h_1\left(\theta_o\right)}{\partial \theta_2'}\theta_{o2} \\
y_2 &\equiv \hat{\gamma}_2 - h_2\left(\theta_{o1}, \theta_{o2}\right) + \frac{\partial h_2\left(\theta_o\right)}{\partial \theta_1'}\theta_{o1} + \frac{\partial h_2\left(\theta_o\right)}{\partial \theta_2'}\theta_{o2} \\
X_1 &\equiv \frac{\partial h_1\left(\theta_o\right)}{\partial \theta_1'} \\
Z_1 &\equiv \frac{\partial h_1\left(\theta_o\right)}{\partial \theta_2'} \\
X_2 &\equiv \frac{\partial h_2\left(\theta_o\right)}{\partial \theta_1'} \\
Z_2 &\equiv \frac{\partial h_2\left(\theta_o\right)}{\partial \theta_2'};\ \text{invertible } p_2 \times p_2
\end{aligned}$$

where

$$E\begin{bmatrix} u_{1n} \\ u_{2n} \end{bmatrix} = 0$$

$$V\begin{bmatrix} u_{1n} \\ u_{2n} \end{bmatrix} = \Omega_o$$

Define

$$\begin{aligned}
y_1 &\equiv \hat{\gamma}_1 - h_1(\theta_{o1}, \theta_{o2}) + \frac{\partial h_1(\theta_o)}{\partial \theta_1'} \theta_{o1} + \frac{\partial h_1(\theta_o)}{\partial \theta_2'} \theta_{o2} \\
y_2 &\equiv \hat{\gamma}_2 - h_2(\theta_{o1}, \theta_{o2}) + \frac{\partial h_2(\theta_o)}{\partial \theta_1'} \theta_{o1} + \frac{\partial h_2(\theta_o)}{\partial \theta_2'} \theta_{o2} \\
X_1 &\equiv \frac{\partial h_1(\theta_o)}{\partial \theta_1'} \\
Z_1 &\equiv \frac{\partial h_1(\theta_o)}{\partial \theta_2'} \\
X_2 &\equiv \frac{\partial h_2(\theta_o)}{\partial \theta_1'} \\
Z_2 &\equiv \frac{\partial h_2(\theta_o)}{\partial \theta_2'}; \text{ invertible } p_2 \times p_2
\end{aligned}$$

Then we have a system

$$\begin{cases} y_1 = X_1\theta_1 + Z_1\theta_2 + u_{1n} \\ y_2 = X_2\theta_1 + Z_2\theta_2 + u_{2n} \end{cases}$$

Since Z_2 invertible, we have

$$Z_2^{-1} y_2 = Z_2^{-1} X_2 \theta_1 + \theta_2 + Z_2^{-1} u_{2n}$$

$\Longleftrightarrow$

$$\theta_2 = Z_2^{-1} y_2 - Z_2^{-1} X_2 \theta_1 - Z_2^{-1} u_{2n}$$

Thus

$$y_1 = X_1\theta_1 + Z_1 \left(Z_2^{-1} y_2 - Z_2^{-1} X_2 \theta_1 - Z_2^{-1} u_{2n} \right) + u_{1n}$$

$\Longleftrightarrow$

$$y_1 - Z_1 Z_2^{-1} y_2 = \left(X_1 - Z_1 Z_2^{-1} X_2 \right) \theta_1 + u_{1n} - Z_1 Z_2^{-1} u_{2n}$$

Then an equivalent system is

$$\begin{cases} y_1 - Z_1 Z_2^{-1} y_2 = \left(X_1 - Z_1 Z_2^{-1} X_2 \right) \theta_1 + u_{1n} - Z_1 Z_2^{-1} u_{2n} \\ y_2 = X_2\theta_1 + Z_2\theta_2 + u_{2n} \end{cases}$$

Moreover, define

$$\begin{aligned} u_{1n}^* &\equiv u_{1n} - Z_1 Z_2^{-1} u_{2n} \\ y_1^* &\equiv y_1 - Z_1 Z_2^{-1} y_2 \\ X_1^* &\equiv X_1 - Z_1 Z_2^{-1} X_2 \\ u_{2n}^* &\equiv u_{2n} - L\left(u_{2n} | u_1^*\right) = u_{2n} - A u_1^* \\ y_2^* &= y_2 - A y_1^* \\ X_2^* &= X_2 - A X_1^* \end{aligned}$$

where the linear projection $L(\cdot|\cdot)$ is defined in auxiliary population space. Then we have another equivalent system

$$\begin{cases} y_1^* = X_1^* \theta_1 + u_{1n}^* \\ y_2^* = X_2^* \theta_1 + Z_2 \theta_2 + u_{2n}^* \end{cases}$$

Since u_{1n}^* and u_{2n}^* are orthogonal here, GLS on the first part only is equivalent to joint GLS for θ_1. From now on, It will be proved that concentrated MD is asymptotically equivalent to running GLS on the first part only.

Consistency The concentrating equation is

$$\hat{\gamma}_2 - h_2(\theta_1, \theta_2) = 0$$

and, by implicit function theorem, we know

$$\theta_2 = \varphi_n(\theta_1)$$

is well-defined and continuously differentiable at each θ_1. The concentrated MD is derived from minimizing distance

$$\hat{\gamma}_1 - h_1(\theta_1, \varphi_n(\theta_1))$$

and consistency of concentrated MD easily follows by the fact that $\varphi_n(\theta_1)$ is well-defined and smooth enough for each θ_1.

Optimal Weight Calculation First, from the concentration identity, we can take differentiation on both handsides

$$\frac{\partial h_2\left(\theta_1, \varphi_n\left(\theta_1\right)\right)}{\partial \theta_1'} + \underbrace{\frac{\partial h_2\left(\theta_1, \varphi_n\left(\theta_1\right)\right)}{\partial \theta_2'}}_{\text{invertible}} \frac{\partial \varphi_n\left(\theta_1\right)}{\partial \theta_1'} = 0$$

Hence

$$\frac{\partial \varphi_n\left(\theta_1\right)}{\partial \theta_1'} = -\left[\frac{\partial h_2\left(\theta_1, \varphi_n\left(\theta_1\right)\right)}{\partial \theta_2'}\right]^{-1} \frac{\partial h_2\left(\theta_1, \varphi_n\left(\theta_1\right)\right)}{\partial \theta_1'}$$

In the minimization problem,

$$\min_{\theta_1} \left[\hat{\gamma}_1 - h_1\left(\theta_1, \varphi_n\left(\theta_1\right)\right)\right]' \hat{W} \left[\hat{\gamma}_1 - h_1\left(\theta_1, \varphi_n\left(\theta_1\right)\right)\right]$$

taking first order condition, we have

$$\begin{aligned}
0 &= \left[\frac{\partial}{\partial \theta_1'} h_1\left(\hat{\theta}_1, \varphi_n\left(\hat{\theta}_1\right)\right) + \frac{\partial}{\partial \theta_2'} h_1\left(\hat{\theta}_1, \varphi_n\left(\hat{\theta}_1\right)\right) \frac{\partial \varphi_n\left(\hat{\theta}_1\right)}{\partial \theta_1}\right]' \hat{W} \left[\hat{\gamma}_1 - h_1\left(\hat{\theta}_1, \varphi_n\left(\hat{\theta}_1\right)\right)\right] \\
&= \underbrace{\left[\frac{\partial}{\partial \theta_1'} h_1\left(\hat{\theta}_1, \varphi_n\left(\hat{\theta}_1\right)\right) - \frac{\partial}{\partial \theta_2'} h_1\left(\hat{\theta}_1, \varphi_n\left(\hat{\theta}_1\right)\right) \left[\frac{\partial h_2\left(\hat{\theta}_1, \varphi_n\left(\hat{\theta}_1\right)\right)}{\partial \theta_2'}\right]^{-1} \frac{\partial h_2\left(\hat{\theta}_1, \varphi_n\left(\hat{\theta}_1\right)\right)}{\partial \theta_1'}\right]'}_{\equiv H_n\left(\hat{\theta}_1\right)} \\
&\quad \times \hat{W} \times \left[\hat{\gamma}_1 - h_1\left(\hat{\theta}_1, \varphi_n\left(\hat{\theta}_1\right)\right)\right]
\end{aligned}$$

Note

$$\begin{aligned}
& h_1\left(\hat{\theta}_1, \varphi_n\left(\hat{\theta}_1\right)\right) \\
&= h_1\left(\theta_{o1}, \varphi_n\left(\theta_{o1}\right)\right) \\
&+ \underbrace{\left[\frac{\partial}{\partial \theta_1'} h_1\left(\theta_1^*, \varphi_n\left(\theta_1^*\right)\right) - \frac{\partial}{\partial \theta_2'} h_1\left(\theta_1^*, \varphi_n\left(\theta_1^*\right)\right) \left[\frac{\partial h_2\left(\theta_1^*, \varphi_n\left(\theta_1^*\right)\right)}{\partial \theta_2'}\right]^{-1} \frac{\partial h_2\left(\theta_1^*, \varphi_n\left(\theta_1^*\right)\right)}{\partial \theta_1'}\right]}_{\equiv H_n\left(\theta_1^*\right)} \\
&\times \left(\hat{\theta}_1 - \theta_{o1}\right)
\end{aligned}$$

where

$$\theta_1^* \text{ lies on the segment connecting } \hat{\theta}_1 \text{ and } \theta_{o1}$$

Hence

$$\begin{aligned} 0 &= H_n\left(\hat{\theta}_1\right)' \hat{W}\left[\hat{\gamma}_1 - h_1\left(\hat{\theta}_1, \varphi_n\left(\hat{\theta}_1\right)\right)\right] \\ &= H_n\left(\hat{\theta}_1\right)' \hat{W}\left[\hat{\gamma}_1 - h_1\left(\theta_{o1}, \varphi_n\left(\theta_{o1}\right)\right) - H_n\left(\theta_1^*\right)\left(\hat{\theta}_1 - \theta_{o1}\right)\right] \end{aligned}$$

$\Rightarrow$

$$\begin{aligned} &\sqrt{N}\left(\hat{\theta}_1 - \theta_{o1}\right) \\ &= \left[H_n\left(\hat{\theta}_1\right)' \hat{W} H_n\left(\theta_1^*\right)\right]^{-1} H_n\left(\theta\right)' \hat{W}\sqrt{N}\left[\hat{\gamma}_1 - h_1\left(\theta_{o1}, \varphi_n\left(\theta_{o1}\right)\right)\right] \\ &= \left[H_o' W_o H_o\right]^{-1} H_o' W_o \sqrt{N}\left[\hat{\gamma}_1 - h_1\left(\theta_{o1}, \varphi_n\left(\theta_{o1}\right)\right)\right] + o_p(1) \\ &= \left[H_o' W_o H_o\right]^{-1} H_o' W_o \sqrt{N}\left[\hat{\gamma}_1 - h_1\left(\theta_{o1}, \theta_{o2}\right) + h_1\left(\theta_{o1}, \theta_{o2}\right) - h_1\left(\theta_{o1}, \varphi_n\left(\theta_{o1}\right)\right)\right] + o_p(1) \\ &= \left[H_o' W_o H_o\right]^{-1} H_o' W_o \left[\sqrt{N}\left(\hat{\gamma}_1 - \gamma_{o1}\right) + \sqrt{N}\left(h_1\left(\theta_{o1}, \theta_{o2}\right) - h_1\left(\theta_{o1}, \varphi_n\left(\theta_{o1}\right)\right)\right)\right] + o_p(1) \end{aligned}$$

Note

$$h_1\left(\theta_{o1}, \varphi_n\left(\theta_{o1}\right)\right) = h_1\left(\theta_{o1}, \theta_{o2}\right) + \frac{\partial h_1\left(\theta_{o1}, \theta_2^{**}\right)}{\partial \theta_2'}\left(\varphi_n\left(\theta_{o1}\right) - \theta_{o2}\right)$$

and

$$\begin{aligned} 0 &= \hat{\gamma}_2 - h_2\left(\theta_{o1}, \varphi_n\left(\theta_{o1}\right)\right) \\ &= \hat{\gamma}_2 - h_2\left(\theta_{o1}, \theta_{o2}\right) - \frac{\partial h_2\left(\theta_{o1}, \theta_2^{***}\right)}{\partial \theta_2'}\left(\varphi_n\left(\theta_{o1}\right) - \theta_{o2}\right) \end{aligned}$$

where

$$\theta_2^{**} \text{ and } \theta_2^{***} \text{ lie on the segment connecting } \varphi_n\left(\theta_{o1}\right) \text{ and } \theta_{o2}$$

which implies

$$\begin{aligned} \left(\varphi_n\left(\theta_{o1}\right) - \theta_{o2}\right) &= \left[\frac{\partial h_2\left(\theta_{o1}, \theta_2^{***}\right)}{\partial \theta_2'}\right]^{-1}\left(\hat{\gamma}_2 - h_2\left(\theta_{o1}, \theta_{o2}\right)\right) \\ &\left[\frac{\partial h_2\left(\theta_{o1}, \theta_2^{***}\right)}{\partial \theta_2'}\right]^{-1}\left(\hat{\gamma}_2 - \gamma_{o2}\right) \end{aligned}$$

Thus

$$\begin{aligned}
&\sqrt{N}\left(\hat{\theta}_1-\theta_{o1}\right)\\
&=\left[H_o'W_oH_o\right]^{-1}H_o'W_o\\
&\times\left[\sqrt{N}\left(\hat{\gamma}_1-\gamma_{o1}\right)-\sqrt{N}\left(\frac{\partial h_1\left(\theta_{o1},\theta_2^{**}\right)}{\partial\theta_2'}\left[\frac{\partial h_2\left(\theta_{o1},\theta_2^{***}\right)}{\partial\theta_2'}\right]^{-1}\left(\hat{\gamma}_2-\gamma_{o2}\right)\right)\right]+o_p\left(1\right)\\
&=\left[H_o'W_oH_o\right]^{-1}H_o'W_o\left[\sqrt{N}\left(\hat{\gamma}_1-\gamma_{o1}\right)-\frac{\partial h_1\left(\theta_o\right)}{\partial\theta_2'}\left[\frac{\partial h_2\left(\theta_o\right)}{\partial\theta_2'}\right]^{-1}\sqrt{N}\left(\hat{\gamma}_2-\gamma_{o2}\right)\right]+o_p\left(1\right)\\
&=\left[H_o'W_oH_o\right]^{-1}H_o'W_o\underbrace{\left[\begin{array}{cc} I_{g_1} & -\frac{\partial h_1(\theta_o)}{\partial\theta_2'}\left[\frac{\partial h_2(\theta_o)}{\partial\theta_2'}\right]^{-1}\end{array}\right]\sqrt{N}\left(\left[\begin{array}{c}\hat{\gamma}_1-\gamma_{o1}\\ \hat{\gamma}_2-\gamma_{o2}\end{array}\right]\right)}_{\text{inverse of asymp. var. of this expression is optimal weight}}+o_p\left(1\right)
\end{aligned}$$

where

$$H_o=\frac{\partial}{\partial\theta_1'}h_1\left(\theta_o\right)-\frac{\partial}{\partial\theta_2'}h_1\left(\theta_o\right)\left[\frac{\partial h_2\left(\theta_o\right)}{\partial\theta_2'}\right]^{-1}\frac{\partial h_2\left(\theta_o\right)}{\partial\theta_1'}$$

Therefore, the optimal weight is

$$W_o=\left(\left[\begin{array}{cc} I_{g_1} & -\frac{\partial h_1(\theta_o)}{\partial\theta_2'}\left[\frac{\partial h_2(\theta_o)}{\partial\theta_2'}\right]^{-1}\end{array}\right]\Omega_o\left[\begin{array}{cc} I_{g_1} & -\frac{\partial h_1(\theta_o)}{\partial\theta_2'}\left[\frac{\partial h_2(\theta_o)}{\partial\theta_2'}\right]^{-1}\end{array}\right]'\right)^{-1}$$

Linearized cMD and auxiliary asymptotic model The linearized cMD on

$$\begin{aligned}
&\hat{\gamma}_1-h_1\left(\theta_{o1},\varphi_n\left(\theta_{o1}\right)\right)-\left[\frac{\partial}{\partial\theta_1'}h_1\left(\theta_o\right)-\frac{\partial}{\partial\theta_2'}h_1\left(\theta_o\right)\left[\frac{\partial h_2\left(\theta_o\right)}{\partial\theta_2'}\right]^{-1}\frac{\partial h_2\left(\theta_o\right)}{\partial\theta_1'}\right]\left(\theta_1-\theta_{o1}\right)\\
&=\hat{\gamma}_1-h_1\left(\theta_{o1},\varphi_n\left(\theta_{o1}\right)\right)-H_o\left(\theta_1-\theta_{o1}\right)
\end{aligned}$$

with weights calculated above is asymptotically equivalent to concentrated MD. The corresponding auxiliary asymptotic model is

$$\hat{\gamma}_1-h_1\left(\theta_{o1},\varphi_n\left(\theta_{o1}\right)\right)+H_o\theta_{o1}=H_o\theta+u_{1n}-Z_1Z_2^{-1}u_{2n}$$

where $u_{1n}-Z_1Z_2^{-1}u_{2n}$ is derived from optimal weight calculation. Moreover, since we know

$$\left(X_1-Z_1Z_2^{-1}X_2\right)=H_o$$

to see its equivalence to

$$y_1 - Z_1 Z_2^{-1} y_2 = \left(X_1 - Z_1 Z_2^{-1} X_2\right) \theta_1 + u_{1n} - Z_1 Z_2^{-1} u_{2n}$$

it suffices to show that we can replace $\hat{\gamma}_1 - h_1(\theta_{o1}, \varphi_n(\theta_{o1})) + H_o \theta_{o1}$ with $y_1 - Z_1 Z_2^{-1} y_2$. Note that we have

$$\begin{aligned}
& y_1 - Z_1 Z_2^{-1} y_2 \\
&= \hat{\gamma}_1 - h_1(\theta_{o1}, \theta_{o2}) + \frac{\partial h_1(\theta_o)}{\partial \theta_1'} \theta_{o1} + \frac{\partial h_1(\theta_o)}{\partial \theta_2'} \theta_{o2} \\
&- \frac{\partial h_1(\theta_o)}{\partial \theta_2'} \left[\frac{\partial h_2(\theta_o)}{\partial \theta_2'}\right]^{-1} \left[\hat{\gamma}_2 - h_2(\theta_{o1}, \theta_{o2}) + \frac{\partial h_2(\theta_o)}{\partial \theta_1'} \theta_{o1} + \frac{\partial h_2(\theta_o)}{\partial \theta_2'} \theta_{o2}\right] \\
&= \hat{\gamma}_1 - h_1(\theta_{o1}, \theta_{o2}) + \frac{\partial h_1(\theta_o)}{\partial \theta_1'} \theta_{o1} \\
&- \frac{\partial h_1(\theta_o)}{\partial \theta_2'} \left[\frac{\partial h_2(\theta_o)}{\partial \theta_2'}\right]^{-1} \left[\hat{\gamma}_2 - h_2(\theta_{o1}, \theta_{o2}) + \frac{\partial h_2(\theta_o)}{\partial \theta_1'} \theta_{o1}\right]
\end{aligned}$$

along with

$$\begin{aligned}
& \hat{\gamma}_1 - h_1(\theta_{o1}, \varphi_n(\theta_{o1})) + H_o \theta_{o1} \\
&= \hat{\gamma}_1 - h_1(\theta_{o1}, \theta_{o2}) + h_1(\theta_{o1}, \theta_{o2}) - h_1(\theta_{o1}, \varphi_n(\theta_{o1})) \\
&+ \left[\frac{\partial}{\partial \theta_1'} h_1(\theta_o) - \frac{\partial}{\partial \theta_2'} h_1(\theta_o) \left[\frac{\partial h_2(\theta_o)}{\partial \theta_2'}\right]^{-1} \frac{\partial h_2(\theta_o)}{\partial \theta_1'}\right] \theta_{o1} \\
&= \hat{\gamma}_1 - h_1(\theta_{o1}, \theta_{o2}) + \frac{\partial h_1(\theta_o)}{\partial \theta_1'} \theta_{o1} - \frac{\partial h_1(\theta_o)}{\partial \theta_2'} \left[\frac{\partial h_2(\theta_o)}{\partial \theta_2'}\right]^{-1} \frac{\partial h_2(\theta_o)}{\partial \theta_1'} \theta_{o1} \\
&+ h_1(\theta_{o1}, \theta_{o2}) - h_1(\theta_{o1}, \varphi_n(\theta_{o1})) \\
&= \hat{\gamma}_1 - h_1(\theta_{o1}, \theta_{o2}) + \frac{\partial h_1(\theta_o)}{\partial \theta_1'} \theta_{o1} - \frac{\partial h_1(\theta_o)}{\partial \theta_2'} \left[\frac{\partial h_2(\theta_o)}{\partial \theta_2'}\right]^{-1} \frac{\partial h_2(\theta_o)}{\partial \theta_1'} \theta_{o1} \\
&+ \frac{\partial h_1\left(\theta_{o1}, \theta_2^{**}\right)}{\partial \theta_2'} \left[\frac{\partial h_2\left(\theta_{o1}, \theta_2^{***}\right)}{\partial \theta_2'}\right]^{-1} (\hat{\gamma}_2 - \gamma_{o2}) \\
&= \hat{\gamma}_1 - h_1(\theta_{o1}, \theta_{o2}) + \frac{\partial h_1(\theta_o)}{\partial \theta_1'} \theta_{o1} \\
&- \frac{\partial h_1(\theta_o)}{\partial \theta_2'} \left[\frac{\partial h_2(\theta_o)}{\partial \theta_2'}\right]^{-1} \left[\hat{\gamma}_2 - h_2(\theta_{o1}, \theta_{o2}) + \frac{\partial h_2(\theta_o)}{\partial \theta_1'} \theta_{o1}\right] + o_p\left(n^{-\frac{1}{2}}\right)
\end{aligned}$$

Hence the result.

B.3 Proof of Proposition 2.3.7

First, note that

$$V_{MD-QLIML}^{-1} = H_o' A_R' B_R^{-1} A_R H_o$$

where

$$\begin{aligned}
H_o &= \begin{bmatrix} \frac{\partial}{\partial \theta_1'} \gamma\left(\theta_o\right) & \frac{\partial}{\partial \theta_2'} \gamma\left(\theta_o\right) \\ 0 & I_{p_2} \end{bmatrix} \\
A_R &= \left. \frac{\partial}{\partial\left(\gamma', \theta_2'\right)} E \begin{bmatrix} \frac{\partial q_1(\gamma,\theta_2)}{\partial \gamma} \\ \frac{\partial q_2(\theta_2)}{\partial \theta_2} \end{bmatrix} \right|_{(\gamma,\theta_2)=(\gamma_o,\theta_{o2})} \\
B_R &= V \begin{pmatrix} \frac{\partial}{\partial \gamma} q_{i1}\left(\gamma_o, \theta_{o2}\right) \\ \frac{\partial}{\partial \theta_2} q_{i2}\left(\theta_{o2}\right) \end{pmatrix}
\end{aligned}$$

Next, by product rule of differentiation, we have

$$\begin{aligned}
& \left. \frac{\partial}{\partial \theta'} E \begin{bmatrix} \frac{\partial}{\partial \gamma} q_1\left(\gamma\left(\theta\right), \theta_2\right) \\ \frac{\partial}{\partial \theta_2} q_2\left(\theta_2\right) \end{bmatrix} \right|_{\theta=\theta_o} \\
= & \underbrace{\left. \frac{\partial}{\partial\left(\gamma', \theta_2'\right)} E \begin{bmatrix} \frac{\partial q_1(\gamma,\theta_2)}{\partial \gamma} \\ \frac{\partial q_2(\theta_2)}{\partial \theta_2} \end{bmatrix} \right|_{(\gamma,\theta_2)=(\gamma_o,\theta_{o2})}}_{=A_R} \underbrace{\left. \frac{\partial}{\partial \theta'} \begin{bmatrix} \gamma\left(\theta\right) \\ \theta_2 \end{bmatrix} \right|_{\theta=\theta_o}}_{=H_o}
\end{aligned}$$

Hence

$$V_{mGMM-QLIML}^{-1} = H_o' A_R' B_R^{-1} A_R H_o$$

B.4 Proof of Proposition 2.3.8

Note that quasi-likelihoods for reduced form model are

$$q_1\left(\gamma\left(\theta_1,\theta_2\right),\theta_2\right)$$

$$q_2\left(\theta_2\right)$$

We will show $V_{mGMM-QLIML}^{-1}-V_{GMM-QLIML}^{-1}\succeq 0$. First, note that

$$V_{mGMM-QLIML}^{-1}=H_o'A_R'B_R^{-1}A_RH_o$$

where

$$H_o=\left[\begin{array}{cc}\frac{\partial}{\partial\theta_1'}\gamma\left(\theta_o\right) & \frac{\partial}{\partial\theta_2'}\gamma\left(\theta_o\right)\\ 0 & I_{p_2}\end{array}\right]$$

$$A_R=\frac{\partial}{\partial\left(\gamma',\theta_2'\right)}E\left[\begin{array}{c}\frac{\partial q_1\left(\gamma,\theta_2\right)}{\partial\gamma}\\ \frac{\partial q_2\left(\theta_2\right)}{\partial\theta_2}\end{array}\right]\Bigg|_{\left(\gamma,\theta_2\right)=\left(\gamma_o,\theta_{o2}\right)}$$

$$B_R=V\left(\begin{array}{c}\frac{\partial}{\partial\gamma}q_{i1}\left(\gamma_o,\theta_{o2}\right)\\ \frac{\partial}{\partial\theta_2}q_{i2}\left(\theta_{o2}\right)\end{array}\right)$$

It will be shown that $V_{GMM-QLIML}^{-1}$ can be expressed in terms of H_o, A_R,B_R and an additional linear transformation. Suppose θ_{22} is not empty. With probability one, for some $p_{22}\times g$ matrix

$C_2(\theta)$

$$\begin{aligned}
\begin{bmatrix} \frac{\partial}{\partial\theta_1} q_{i1}(\theta) \\ \frac{\partial}{\partial\theta_{22}} q_{i1}(\theta) \\ \frac{\partial}{\partial\theta_2} q_{i2}(\theta_2) \end{bmatrix}
&= \begin{bmatrix} \frac{\partial\gamma}{\partial\theta_1}\frac{\partial}{\partial\gamma} q_{i1}(\gamma,\theta_2) \\ \frac{\partial\gamma}{\partial\theta_{22}}\frac{\partial}{\partial\gamma} q_{i1}(\gamma,\theta_2) + \frac{\partial}{\partial\theta_{22}} q_{i1}(\gamma,\theta_2) \\ \frac{\partial}{\partial\theta_2} q_{i2}(\theta_2) \end{bmatrix} \\
&= \begin{bmatrix} \frac{\partial\gamma}{\partial\theta_1}\frac{\partial}{\partial\gamma} q_{i1}(\gamma,\theta_2) \\ \left[\frac{\partial\gamma}{\partial\theta_{22}} + C_2(\theta)\right]\frac{\partial}{\partial\gamma} q_{i1}(\gamma,\theta_2) \\ \frac{\partial}{\partial\theta_2} q_{i2}(\theta_2) \end{bmatrix} \\
&= \underbrace{\begin{bmatrix} \frac{\partial\gamma}{\partial\theta_1} & 0 \\ \left[\frac{\partial\gamma}{\partial\theta_{22}} + C_2(\theta)\right] & 0 \\ 0 & I_{p_2} \end{bmatrix}}_{\equiv W(\theta):(p+p_{22})\times(g+p_2)} \begin{bmatrix} \frac{\partial}{\partial\gamma} q_{i1}(\gamma,\theta_2) \\ \frac{\partial}{\partial\theta_2} q_{i2}(\theta_2) \end{bmatrix}
\end{aligned}$$

By product rule of differentiation,

$$\begin{aligned}
&\frac{d}{d\theta'}\left(\underset{(p+p_{22})\times(g+p_2)}{W(\theta)} \underset{(g+p_2)\times 1}{E\begin{bmatrix} \frac{\partial}{\partial\gamma} q_{i1}(\gamma,\theta_2) \\ \frac{\partial}{\partial\theta_2} q_{i2}(\theta_2) \end{bmatrix}} \right) \\
&= \left(\underbrace{E\begin{bmatrix} \frac{\partial}{\partial\gamma'} q_{i1}(\gamma,\theta_2) & \frac{\partial}{\partial\theta_2'} q_{i2}(\theta_2) \end{bmatrix}}_{\text{vanish in expectation at true paramters}} \otimes I_{(p+p_{22})} \right) \frac{d}{d\theta'} vec\left[W(\theta)\right] \\
&+ W(\theta)\frac{d}{d\theta'}\left(E\begin{bmatrix} \frac{\partial}{\partial\gamma} q_{i1}(\gamma,\theta_2) \\ \frac{\partial}{\partial\theta_2} q_{i2}(\theta_2) \end{bmatrix} \right)
\end{aligned}$$

where

$$\begin{aligned}
&\frac{d}{d\theta'}\left(E\left[\begin{array}{c}\frac{\partial}{\partial\gamma}q_{i1}\left(\gamma,\theta_2\right)\\ \frac{\partial}{\partial\theta_2}q_{i2}\left(\theta_2\right)\end{array}\right]\right)\\
&=\left[\begin{array}{cc}\frac{1}{\partial\gamma'}\left(E\left[\frac{\partial}{\partial\gamma}q_{i1}\left(\gamma,\theta_2\right)\right]\right)\frac{\partial\gamma}{\partial\theta_1'} & \frac{1}{\partial\gamma'}\left(E\left[\frac{\partial}{\partial\gamma}q_{i1}\left(\gamma,\theta_2\right)\right]\right)\frac{\partial\gamma}{\partial\theta_2'}+\frac{1}{\partial\theta_2'}\left(E\left[\frac{\partial}{\partial\gamma}q_{i1}\left(\gamma,\theta_2\right)\right]\right)\\ 0 & \frac{1}{\partial\theta_2'}\left(E\left[\frac{\partial}{\partial\theta_2}q_{i2}\left(\theta_2\right)\right]\right)\end{array}\right]\\
&=\underbrace{\left[\begin{array}{cc}\frac{1}{\partial\gamma'}\left(E\left[\frac{\partial}{\partial\gamma}q_{i1}\left(\gamma,\theta_2\right)\right]\right) & \frac{1}{\partial\theta_2'}\left(E\left[\frac{\partial}{\partial\gamma}q_{i1}\left(\gamma,\theta_2\right)\right]\right)\\ 0 & \frac{1}{\partial\theta_2'}\left(E\left[\frac{\partial}{\partial\theta_2}q_{i2}\left(\theta_2\right)\right]\right)\end{array}\right]}_{=A_R\text{ when }(\gamma,\theta_2)=(\gamma_o,\theta_{o2})}\underbrace{\left[\begin{array}{cc}\frac{\partial\gamma}{\partial\theta_1'} & \frac{\partial\gamma}{\partial\theta_2'}\\ 0 & I_p\end{array}\right]}_{=H_o\text{ when }\theta=\theta_o}
\end{aligned}$$

Then we have

$$\frac{d}{d\theta'}\,E\left[\begin{array}{c}\frac{\partial}{\partial\theta_1}q_{i1}\left(\theta\right)\\ \frac{\partial}{\partial\theta_{22}}q_{i1}\left(\theta\right)\\ \frac{\partial}{\partial\theta_2}q_{i2}\left(\theta_2\right)\end{array}\right]\Bigg|_{\theta=\theta_o}=W_oA_RH_o$$

where $W_o=W\left(\theta_o\right)$. Also, it is easy to see

$$V\left(\begin{array}{c}\frac{\partial}{\partial\theta_1}q_{i1}\left(\theta_{o1},\theta_{o2}\right)\\ \frac{\partial}{\partial\theta_{22}}q_{i1}\left(\theta_{o1},\theta_{o2}\right)\\ \frac{\partial}{\partial\theta_2}q_{i2}\left(\theta_{o2}\right)\end{array}\right)=W_oB_MW_o'$$

Hence

$$V_{GMM-QLIML}^{-1}=H_o'A_R'W_o'\left(W_oB_RW_o'\right)^{-1}W_oA_RH_o$$

To see relative efficiency of MD-QLIML,

$$\begin{aligned}
&V_{mGMM-QLIML}^{-1}-V_{GMM-QLIML}^{-1}\\
&=H_o'A_R'B_R^{-1}A_RH_o-H_o'A_R'W_o'\left(W_oB_RW_o'\right)^{-1}W_oA_RH_o\\
&=H_o'A_R'\left(B_R^{-1}-W_o'\left(W_oB_RW_o'\right)^{-1}W_o\right)A_RH_o\\
&=H_o'A_R'B_R^{-\frac{1}{2}}\left(I-B_R^{\frac{1}{2}}W_o'\left(W_oB_R^{\frac{1}{2}\prime}B_R^{\frac{1}{2}}W_o'\right)^{-1}W_oB_R^{\frac{1}{2}\prime}\right)B_R^{-\frac{1}{2}\prime}A_RH_o\succeq 0
\end{aligned}$$

where

$$B_R = B_R^{\frac{1}{2}'} B_R^{\frac{1}{2}}$$

When $p_1 + p_{22} = g$ holds, W_o is invertible and we have

$$I - B_R^{\frac{1}{2}} W_o' \left(W_o B_R^{\frac{1}{2}'} B_R^{\frac{1}{2}} W_o' \right)^{-1} W_o B_R^{\frac{1}{2}'} = 0$$

In the case where θ_{22} is empty, the result can be shown with slight modification of above proof.

B.5 Proof of Proposition 2.3.9

(a)

(Well-definedness of reduced form model) The reduced form likelihood is

$$\begin{aligned} q_{i1}(\theta_1, \theta_2) &= l\left(y_{i1}, \mathbf{y}_2\alpha + \mathbf{z}_1\delta_1 + \mathbf{v}_2\eta, \lambda\right) \\ &= l\left(y_{i1}, \mathbf{z}_1\left(\boldsymbol{\delta}_{21}\alpha + \delta_1\right) + \mathbf{z}_2\boldsymbol{\delta}_{22}\alpha \ + \mathbf{v}_2\left(\alpha + \eta\right), \lambda\right) \\ &= l\left(y_{i1}, \mathbf{z}\gamma_1(\theta) + \mathbf{v}_2\gamma_2(\theta), \gamma_3(\theta)\right) \\ &= q_1\left(\gamma(\theta), \theta_2\right) \end{aligned}$$

Since q_{i1} depends on θ_2 only through $\boldsymbol{\delta}_2$, it suffices to show that each element of $\frac{\partial q_1}{\partial \boldsymbol{\delta}_2}$ can be expressed as a linear combination of $\frac{\partial q_1}{\partial \gamma_1}$. Note

$$\begin{aligned} \frac{\partial q_1(\theta_1, \theta_2)}{\partial \boldsymbol{\delta}_2} &= -s\left(y_{i1}, \mathbf{y}_2\alpha + \mathbf{z}_1\delta_1 + \mathbf{v}_2(\boldsymbol{\delta}_2)\eta, \lambda\right)\eta \otimes \mathbf{z}' \\ \frac{\partial q_1(\gamma, \theta_2)}{\partial \gamma_1} &= s\left(y_{i1}, \mathbf{z}\gamma_1 + \mathbf{v}_2(\boldsymbol{\delta}_2)\gamma_2, \gamma_3\right)\mathbf{z}' \\ &= s\left(y_{i1}, \mathbf{y}_2\alpha + \mathbf{z}_1\delta_1 + \mathbf{v}_2(\boldsymbol{\delta}_2)\eta, \lambda\right)\mathbf{z}' \end{aligned}$$

where $s(y_{i1}, \Upsilon, \theta_1) = \frac{\partial l(y_{i1}, \Upsilon, \theta_1)}{\partial \Upsilon}$. Hence the proof.

(b)

Suppose $k_2 = r$. [To show $V_{GMM-QLIML} = V_{QLIML} = V_{CF}$] The quasi-scores are

$$\frac{\partial q_1(\theta_1, \theta_2)}{\partial \theta_1} = \begin{bmatrix} s(y_{i1}, \mathbf{y}_2\alpha + \mathbf{z}_1\delta_1 + \mathbf{v}_2\eta, \lambda)\, \mathbf{y}_2' \\ s(y_{i1}, \mathbf{y}_2\alpha + \mathbf{z}_1\delta_1 + \mathbf{v}_2\eta, \lambda)\, \mathbf{z}_1' \\ s(y_{i1}, \mathbf{y}_2\alpha + \mathbf{z}_1\delta_1 + \mathbf{v}_2\eta, \lambda)\, \mathbf{v}_2' \\ \frac{\partial l(y_{i1}, \mathbf{y}_2\alpha + \mathbf{z}_1\delta_1 + \mathbf{v}_2\eta, \lambda)}{\partial \lambda} \end{bmatrix} \tag{A.2}$$

$$\frac{\partial q_1(\theta_1, \theta_2)}{\partial \theta_2} = \begin{bmatrix} -s(y_{i1}, \mathbf{y}_2\alpha + \mathbf{z}_1\delta_1 + \mathbf{v}_2\eta, \lambda)\, \eta \otimes \mathbf{z}' \\ 0_{\frac{r(r+1)}{2} \times 1} \end{bmatrix} \tag{A.3}$$

QLIML and CF rank conditions implies that $k_2 \times r$ matrix $\boldsymbol{\delta}_{o22}$ is required to be full column rank. Since $k_2 = r$, $\boldsymbol{\delta}_{o22}$ is an invertible matrix. Also, noting that

$$s(y_{i1}, \mathbf{y}_2\alpha + \mathbf{z}_1\delta_1 + \mathbf{v}_2\eta, \lambda)\, \mathbf{y}_2' = s(y_{i1}, \mathbf{y}_2\alpha + \mathbf{z}_1\delta_1 + \mathbf{v}_2\eta, \lambda)\, (\mathbf{z}_1\boldsymbol{\delta}_{21} + \mathbf{z}_2\boldsymbol{\delta}_{22} + \mathbf{v}_2)'$$

any moment function in $\frac{\partial q_1(\theta_1,\theta_2)}{\partial \theta_2}$ can be expressed as linear combination of $\frac{\partial q_1(\theta_1,\theta_2)}{\partial \theta_1}$. Thus θ_{22} is empty and the result follows by Proposition 2.3.6 (b). [To show $V_{MD-QLIML} = V_{GMM-QLIML}$] It suffices to show $p_1 + p_{22} = g$. Let $\lambda, \gamma_3 \in \mathbb{R}^l$. As shown above, $p_{22} = 0$. Then, $p_1 = r + k_1 + r + l$ and $g = k_1 + k_2 + r + l$. Since $k_2 = r$, the result follows.

(c)

Suppose $\eta_o \neq 0$. [To show $V_{MD-QLIML} = V_{GMM-QLIML}$] As in (b,2), $p_1 = r + k_1 + r + l$ and $g = k_1 + k_2 + r + l$. It suffices to show $p_1 + p_{22} = g$ or equivalently, $p_{22} = k_2 - r$. The case with $k_2 = r$ was shown in (b,2). Case $k_2 < r$ is ruled out by order condition. i.e. $\boldsymbol{\delta}_{o22}$ cannot be full column rank with $k_2 < r$. Suppose $k_2 > r$. By rank condition of reduced form model, we know linear independence of components in score

$$\frac{\partial q_1(\gamma(\theta),\theta_2)}{\partial\gamma} = \begin{bmatrix} s(y_{i1}, \mathbf{z}\gamma_1 + \mathbf{v}_2(\boldsymbol{\delta}_2)\gamma_2, \gamma_3)\mathbf{z}' \\ s(y_{i1}, \mathbf{z}\gamma_1 + \mathbf{v}_2(\boldsymbol{\delta}_2)\gamma_2, \gamma_3)\mathbf{v}_2' \\ \frac{\partial q_1}{\partial\gamma_3} \end{bmatrix}$$
$$= \begin{bmatrix} s(y_{i1}, \mathbf{y}_2\alpha + \mathbf{z}_1\delta_1 + \mathbf{v}_2(\boldsymbol{\delta}_2)\eta, \lambda)\mathbf{z}' \\ s(y_{i1}, \mathbf{y}_2\alpha + \mathbf{z}_1\delta_1 + \mathbf{v}_2(\boldsymbol{\delta}_2)\eta, \lambda)\mathbf{v}_2' \\ \frac{\partial q_1}{\partial\lambda} \end{bmatrix}$$

Then, due to explicit linear relationship $\mathbf{y}_2 = \mathbf{z}_1\boldsymbol{\delta}_{21} + \mathbf{z}_2\boldsymbol{\delta}_{22} + \mathbf{v}_2$, a maximal linearly independent set in $\left\{s\mathbf{y}_2, s\mathbf{z}_1, s\mathbf{z}_2, s\mathbf{v}_2, \frac{\partial q_1}{\partial\lambda}\right\}$ always contains $(k_1 + k_2 + r + l)$ elements. Hence, a maximal linearly independent set in $\left\{\frac{\partial q_1(\theta_o)}{\partial\theta_1}, \frac{\partial q_1(\theta_o)}{\partial\theta_2}\right\}$ contains $(k_1 + k_2 + r + l)$ elements whenever there exists at least one nonzero element in η_o. Since $\frac{\partial q_1(\theta_o)}{\partial\theta_1}$ contains $k_1 + 2r + l$ moment functions, it is implied that $p_{22} = k_2 - r$. [To show $V_{GMM-QLIML} \preceq V_{QLIML}, V_{CF}$] The result follows from Proposition 2.3.7

(d)

Suppose $\eta_o = 0$. In (A.3), we have $\frac{\partial q_1(\theta_1,\theta_2)}{\partial\theta_2} = 0_{p_2\times 1}$ and θ_{22} is empty. Then, $V_{GMM-QLIML} = V_{QLIML} = V_{CF}$ by Proposition 1.3.6 in Chapter 1.

(e)

Let M be the linear span generated by mGMM-QLIML moment functions

$$\begin{bmatrix} \frac{\partial q_1(\gamma(\theta_o),\theta_2)}{\partial\gamma} \\ \frac{\partial q_2(\theta_2)}{\partial\theta_2} \end{bmatrix} \tag{A.4}$$

When $\eta_o = 0$, θ_{22} is empty and $V_{GMM-QLIML} = V_{QLIML} = V_{CF}$ as shown in (d). Hence, GMM-QLIML moment functions are

$$\begin{bmatrix} \frac{\partial q_1(\theta_o)}{\partial\theta_1} \\ \frac{\partial q_2(\theta_{o2})}{\partial\theta_2} \end{bmatrix} = \underbrace{\begin{bmatrix} \frac{\partial\gamma(\theta_o)}{\partial\theta_1} & 0 \\ 0 & I_{p_2} \end{bmatrix}}_{(g+p_2)\times p} \underbrace{\begin{bmatrix} \frac{\partial q_1(\gamma(\theta_o),\theta_{o2})}{\partial\gamma} \\ \frac{\partial q_2(\theta_{o2})}{\partial\theta_2} \end{bmatrix}}_{mGMM-QLIML \text{ moments}}$$

where

$$\begin{bmatrix} \frac{\partial\gamma(\theta_o)}{\partial\theta_1} & 0 \\ 0 & I_{p_2} \end{bmatrix}$$

is assumed to be full column rank by Assumption 15. Hence, GMM-QLIML moment functions is a linearly independent set in M. Also, since $\frac{\partial q_1(\theta_o)}{\partial\theta_2} = 0$ when $\eta_o = 0$, clearly we have

$$\begin{bmatrix} \frac{\partial q_1(\theta_o)}{\partial\theta_1} \\ \frac{\partial q_2(\theta_{o2})}{\partial\theta_2} \end{bmatrix} = \begin{bmatrix} \frac{\partial q_1(\theta_o)}{\partial\theta_1} \\ \frac{\partial q_1(\theta_o)}{\partial\theta_2} + \frac{\partial q_2(\theta_{o2})}{\partial\theta_2} \end{bmatrix}$$

and it is implied that QLIML moment functions are also linearly independent in M. Since we are assuming $k_2 > r$, we have $g = k_1 + k_2 + r + l > r + k_1 + r + l = p_1$. Thus, the dimension of M is larger than number of GMM-QLIML(or QLIML) moment functions, p. Relative efficiency of mGMM-QLIML is obvious. To find a condition for asymptotic equivalence, consider QLIML moment functions

$$\begin{bmatrix} \frac{\partial q_1(\theta)}{\partial\theta_1} \\ \frac{\partial q_1(\theta)}{\partial\theta_2} + \frac{\partial q_2(\theta_2)}{\partial\theta_2} \end{bmatrix}$$

By replacement theorem (Thm 1.10, Friedberg et al, 2003), there exists $k_2 - r$ elements in (A.4) with which QLIML moment functions constitute a basis of M at true parameter values. Denote such $k_2 - r$ elements as

$$\frac{\partial q_1\left(\gamma\left(\theta\right), \theta_2\right)}{\partial\gamma^*}$$

where γ^* is $(k_2 - r) \times 1$. Then optimal GMM on

$$\begin{bmatrix} \frac{\partial q_1(\theta)}{\partial\theta_1} \\ \frac{\partial q_1(\theta)}{\partial\theta_2} + \frac{\partial q_2(\theta_2)}{\partial\theta_2} \\ \frac{\partial q_1(\gamma(\theta),\theta_2)}{\partial\gamma^*} \end{bmatrix} \tag{A.5}$$

is asympotically equivalent to mGMM-QLIML by similar reasoning in Lemma C.1. Equivalence condition follows by applying BQSW redundancy condition to (A.5). To see sufficiency of GIME's for reduced form model, note that

$$
\begin{aligned}
V\left(\frac{\partial q_1^o}{\partial\left(\gamma', \theta_2'\right)'}\right) &= \tau\ \frac{\partial}{\partial\left(\gamma', \theta_2'\right)} E\left[\frac{\partial q_1}{\partial\left(\gamma', \theta_2'\right)'}\right]\Bigg|_{\left(\gamma', \theta_2'\right)' = \left(\gamma_o', \theta_{o2}'\right)'} \\
V\left(\frac{\partial q_2^o}{\partial \theta_2}\right) &= \tau\ \frac{\partial}{\partial \theta_2'} E\left[\frac{\partial q_2^o}{\partial \theta_2}\right]\Bigg|_{\theta_2 = \theta_{o2}} \\
V\left(\begin{bmatrix} \frac{\partial q_1^o}{\partial \gamma} \\ \frac{\partial q_1^o}{\partial \theta_2} + \frac{\partial q_2^o}{\partial \theta_2} \end{bmatrix}\right) &= \tau\ \frac{\partial}{\partial\left(\gamma', \theta_2'\right)} E\begin{bmatrix} \frac{\partial q_1^o}{\partial \gamma \partial \gamma'} \\ \frac{\partial q_1^o}{\partial \theta_2 \partial\left(\gamma', \theta_2'\right)} + \frac{\partial q_2^o}{\partial \theta_2 \partial\left(\gamma', \theta_2'\right)} \end{bmatrix}\Bigg|_{\left(\gamma', \theta_2'\right)' = \left(\gamma_o', \theta_{o2}'\right)'}
\end{aligned}
$$

implies $cov\left(\frac{\partial q_1^o}{\partial\left(\gamma', \theta_2'\right)'}, \frac{\partial q_2^o}{\partial \theta_2}\right) = 0$ and GIME's for structural model. Then result follows by some algebra.

APPENDIX C

AN APPENDIX FOR CHAPTER 3

C.1 Appendix: Proofs

Many proof ideas and steps used in Theorem 3.4.7—3.4.15 are similar to Wang, Wu, Li (2012) and Sherwood and Wang (2016). In the following, C denotes a constant that does not depend on N. It is allowed to take different values in different places.

C.1.1 Proof of Theorem 3.3.1

The model structure (or admissible structure) S in consideration can be defined as following. Denote $\varepsilon_{it} = y_{it} - \mathbf{w}_{it}\boldsymbol{\beta} - g\left(\mathbf{x}_i, \mathbf{z}_i\right)$. Let $S = \{(\tilde{\boldsymbol{\beta}}, \tilde{g}, \tilde{F}_{\{\varepsilon_{it}, \mathbf{x}_{it}\}_{t=1}^T, \mathbf{z}_i}) | \tilde{\boldsymbol{\beta}} \in \mathbb{R}^{K_1+T-1}$, $\tilde{g} : \mathbb{R}^{K_1 T + K_2} \to \mathbb{R}$ is measurable, $\tilde{F}_{\{\varepsilon_{it}, \mathbf{x}_{it}\}_{t=1}^T, \mathbf{z}_i}$ is a distribution function on $\mathbb{R}^{K_1 T + K_2 + T}$ such that $Q_\tau\left(\varepsilon_{it} | \mathbf{x}_i, \mathbf{z}_i\right) = 0$ for each t and the support condition on $(\mathbf{w}_{it})_{t=1}^T$ given in the premise is satisfied.$\}$ Suppose $(\boldsymbol{\beta}^*, g^*, F^*_{\{\varepsilon_{it}, \mathbf{x}_{it}\}_{t=1}^T, \mathbf{z}_i}) \in S$. Then, for each $t = 2, \cdots, T$, we can write

$$Q_\tau\left[y_{it} | \mathbf{x}_i, \mathbf{z}_i ; \boldsymbol{\beta}^*, g^*, F^*_{\{\varepsilon_{it}, \mathbf{x}_{it}\}_{t=1}^T, \mathbf{z}_i}\right] = \mathbf{w}_{it}\boldsymbol{\beta}^* + g^*\left(\mathbf{x}_i, \mathbf{z}_i\right) \tag{A.1}$$

$$Q_\tau\left[y_{i(t-1)} | \mathbf{x}_i, \mathbf{z}_i ; \boldsymbol{\beta}^*, g^*, F^*_{\{\varepsilon_{it}, \mathbf{x}_{it}\}_{t=1}^T, \mathbf{z}_i}\right] = \mathbf{w}_{i(t-1)}\boldsymbol{\beta}^* + g^*\left(\mathbf{x}_i, \mathbf{z}_i\right) \tag{A.2}$$

Note that the conditional quantiles of y_{it} and $y_{i(t-1)}$ are unique. By taking difference across time periods, we have

$$Q_\tau\left[y_{it} | \mathbf{x}_i, \mathbf{z}_i ; \boldsymbol{\beta}^*, g^*, F^*_{\{\varepsilon_{it}, \mathbf{x}_{it}\}_{t=1}^T, \mathbf{z}_i}\right] - Q_\tau\left[y_{i(t-1)} | \mathbf{x}_i, \mathbf{z}_i ; \boldsymbol{\beta}^*, g^*, F^*_{\{\varepsilon_{it}, \mathbf{x}_{it}\}_{t=1}^T, \mathbf{z}_i}\right] \tag{A.3}$$

$$= \left(\mathbf{w}_{it} - \mathbf{w}_{i(t-1)}\right)\boldsymbol{\beta}^* \tag{A.4}$$

The full rank condition on (3.11) will be shown to imply point-identification of $\boldsymbol{\beta}^*$. First, note that there exists a square invertible submatrix of (3.11). Let such matrix be $\ddot{\mathbf{W}}$. By continuity of the matrix determinant, there exists a neighborhood around $\times_{j=1}^J (\tilde{\mathbf{x}}_{it}^{(j)}, \tilde{\mathbf{x}}_{i(t-1)}^{(j)})$ each of whose elements along with $(\dot{\mathbf{x}}_{it}^{(j)}, \dot{\mathbf{x}}_{i(t-1)}^{(j)})_{j=1}^{J}$ and time dummies constitute a perturbed version of $\ddot{\mathbf{W}}$ which is still invertible. Since $f_{\left(\tilde{\mathbf{x}}_{it}, \tilde{\mathbf{x}}_{i(t-1)}\right)|\left(\dot{\mathbf{x}}_{it}, \dot{\mathbf{x}}_{i(t-1)}\right)}(\tilde{\mathbf{x}}_{it}^{(j)}, \tilde{\mathbf{x}}_{i(t-1)}^{(j)}|\dot{\mathbf{x}}_{it}^{(j)}, \dot{\mathbf{x}}_{i(t-1)}^{(j)}) > 0 \; \forall j$ and it is continuously extendable, the probability of observing such collection of support points is positive. (equivalently, a change in $\boldsymbol{\beta}^*$ implies a nontrivial change in $F_{\{y_{it}, \mathbf{x}_{it}\}_{t=1}^{T}, \mathbf{z}_i}$) Hence the proof.

C.1.2 Proof of Theorem 3.4.3 and 3.4.4

Define $\boldsymbol{\theta}_A = (\boldsymbol{\beta}, \boldsymbol{\gamma}_A)$, $\check{\mathbf{w}}_{it} = \frac{1}{\sqrt{N}} \tilde{\mathbf{w}}_{it}^{A}$ and $\boldsymbol{\Psi}_\tau(\boldsymbol{e}) = \left(\Psi_\tau(\boldsymbol{e}_1)', \cdots, \Psi_\tau(\boldsymbol{e}_N)'\right)'$. Also, $\|A\| = \sqrt{\lambda_{\max}(A'A)}$ denotes spectral norm for matrix A, $\|\mathbf{v}\|$ denotes Euclidean norm for vector $\mathbf{v}$, and write $E_w(\cdot) = E(\cdot|\mathbf{x}_i, \mathbf{z}_i)$ and $P_w(\cdot) = P(\cdot|\mathbf{x}_i, \mathbf{z}_i)$.

Consider a following reparameterized objection function

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\rho_\tau(e_{it} - \check{\mathbf{w}}_{it}\boldsymbol{\delta}) = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\rho_\tau\left(y_{it} - \tilde{\mathbf{w}}_{it}^{A}\left(\theta_{oA} + \frac{1}{\sqrt{N}}\boldsymbol{\delta}\right)\right)$$

Let $\hat{\boldsymbol{\delta}}$ be the reparameterized oracle estimator

$$\hat{\boldsymbol{\delta}} = \arg\min_{\boldsymbol{\delta}} \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\rho_\tau(e_{it} - \check{\mathbf{w}}_{it}\boldsymbol{\delta})$$

where $\hat{\boldsymbol{\delta}} = \sqrt{N}\left(\hat{\boldsymbol{\theta}}_A - \boldsymbol{\theta}_{oA}\right)$ holds. Its Bahadur Representation can be written as

$$\tilde{\boldsymbol{\delta}} = \sqrt{N}\left(\tilde{\mathbf{W}}_A' B_N \tilde{\mathbf{W}}_A\right)^{-1}\tilde{\mathbf{W}}_A'\boldsymbol{\Psi}_\tau(\boldsymbol{\varepsilon})$$

Lemma C.1.1 If Assumption 1–4 hold, then$\left\|\tilde{\boldsymbol{\delta}}\right\| = O_p\left(\sqrt{q_N}\right)$. (ii) If Assumption 1-5 hold, then $G_N \Sigma_N^{-1/2}\tilde{\boldsymbol{\delta}} \xrightarrow{d} N(0, G)$.

Proof. Since

$$\begin{aligned}\sqrt{\lambda_{\max}\left(\frac{1}{N}\tilde{\mathbf{W}}_A' B_N \tilde{\mathbf{W}}_A\right)} &= \left\|\frac{1}{\sqrt{N}} B_N^{\frac{1}{2}}\tilde{\mathbf{W}}_A\right\| \\ &\le \left\|B_N^{\frac{1}{2}}\right\|\left\|\frac{1}{\sqrt{N}}\tilde{\mathbf{W}}_A\right\| = \sqrt{\lambda_{\max}(B_N)}\sqrt{\lambda_{\max}\left(\frac{1}{N}\tilde{\mathbf{W}}_A'\tilde{\mathbf{W}}_A\right)}\end{aligned}$$

$\lambda_{\max}\left(\frac{1}{N}\tilde{\mathbf{W}}'_A B_N \tilde{\mathbf{W}}_A\right)$ is also bounded above. Similarly, we can show $\lambda_{\min}\left(\frac{1}{N}\tilde{\mathbf{W}}'_A B_N \tilde{\mathbf{W}}_A\right)$ is bounded below by some positive constant. Then, note that

$$\left\|\tilde{\boldsymbol{\delta}}\right\| = \left\|\left(\frac{1}{N}\tilde{\mathbf{W}}'_A B_N \tilde{\mathbf{W}}_A\right)^{-1} \frac{1}{\sqrt{N}}\tilde{\mathbf{W}}'_A \boldsymbol{\Psi}_\tau(\boldsymbol{\varepsilon})\right\| \leq \left\|\left(\frac{1}{N}\tilde{\mathbf{W}}'_A B_N \tilde{\mathbf{W}}_A\right)^{-\frac{1}{2}}\right\|^2 \left\|\frac{1}{\sqrt{N}}\tilde{\mathbf{W}}'_A \boldsymbol{\Psi}_\tau(\boldsymbol{\varepsilon})\right\|$$

$$\leq C \left\|\frac{1}{\sqrt{N}}\tilde{\mathbf{W}}'_A \boldsymbol{\Psi}_\tau(\boldsymbol{\varepsilon})\right\| = O_p\left(\sqrt{q_N}\right)$$

(ii)

$$G_N \Sigma_N^{-1/2}\tilde{\boldsymbol{\delta}} = G_N \Sigma_N^{-1/2}\mathbf{K}_N^{-1} N^{-1/2}\tilde{\mathbf{W}}'_A \boldsymbol{\Psi}_\tau(\boldsymbol{\varepsilon}) = G_N \Sigma_N^{-1/2}\mathbf{K}_N^{-1} N^{-1/2}\sum_{i=1}^{N}\tilde{\mathbf{W}}_i^{A\prime}\Psi_\tau(\boldsymbol{\varepsilon}_i)$$

$$= G_N \Sigma_N^{-1/2}\mathbf{K}_N^{-1} N^{-1/2}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{\mathbf{w}}_{it}^{A\prime}\psi_\tau(\varepsilon_{it}) = \sum_{i=1}^{N}\sum_{t=1}^{T} D_{Nit}$$

where $D_{Nit} = G_N \Sigma_N^{-1/2}\mathbf{K}_N^{-1} N^{-1/2}\tilde{\mathbf{w}}_{it}^{A\prime}\psi_\tau(\varepsilon_{it})$. Note $E[D_{Nit}] = 0$ and $E\left[\sum_{t=1}^{T} D_{Nit}\right] = 0$. Then,

$$\sum_{i=1}^{N} E\left[\left(\sum_{t=1}^{T} D_{Nit}\right)\left(\sum_{t=1}^{T} D_{Nit}\right)'\right]$$

$$= E\left[G_N \Sigma_N^{-1/2}\mathbf{K}_N^{-1}\left[N^{-1}\sum_{i=1}^{N}\left(\sum_{t=1}^{T}\tilde{\mathbf{w}}_{it}^{A\prime}\psi_\tau(\varepsilon_{it})\right)\left(\sum_{t=1}^{T}\psi_\tau(\varepsilon_{it})\tilde{\mathbf{w}}_{it}^{A}\right)\right]\mathbf{K}_N^{-1}\Sigma_N^{-1/2}G'_N\right]$$

$$= E\left[G_N \Sigma_N^{-1/2}\mathbf{K}_N^{-1}\left[N^{-1}\sum_{i=1}^{N}\tilde{\mathbf{W}}_i^{A\prime}\Psi_\tau(\boldsymbol{\varepsilon}_i)\Psi_\tau(\boldsymbol{\varepsilon}_i)'\tilde{\mathbf{W}}_i^{A}\right]\mathbf{K}_N^{-1}\Sigma_N^{-1/2}G'_N\right]$$

$$= G_N E\left[\Sigma_N^{-1/2}\mathbf{K}_N^{-1}\mathbf{S}_N\mathbf{K}_N^{-1}\Sigma_N^{-1/2}\right]G'_N = G_N G'_N \to G$$

To check Lindeberg-Feller condition, fix $\varepsilon > 0$. By Assumption 3, 4, and 5,

$$\sum_{i=1}^{N} E\left[\left\|\sum_{t=1}^{T} D_{Nit}\right\|^2 1\left[\left\|\sum_{t=1}^{T} D_{Nit}\right\| > \varepsilon\right]\right] \le \frac{1}{\varepsilon^2}\sum_{i=1}^{N} E\left\|\sum_{t=1}^{T} D_{Nit}\right\|^4$$

$$\le \frac{1}{N^2\varepsilon^2}\sum_{i=1}^{N} E\left(\sum_{t=1}^{T}\sum_{t'=1}^{T}\left|\psi_\tau\left(\varepsilon_{it}\right)\psi_\tau\left(\varepsilon_{it'}\right)\tilde{\mathbf{w}}_{it}^{A}\mathbf{K}_N^{-1}\Sigma_N^{-1/2}G_N' G_N\Sigma_N^{-1/2}\mathbf{K}_N^{-1}\tilde{\mathbf{w}}_{it'}^{A\prime}\right|\right)^2$$

$$\le \frac{C}{N^2\varepsilon^2}\sum_{i=1}^{N} E\left(\sum_{t=1}^{T}\sum_{t'=1}^{T}\tilde{\mathbf{w}}_{it}^{A}\mathbf{K}_N^{-1}\Sigma_N^{-1/2}G_N' G_N\Sigma_N^{-1/2}\mathbf{K}_N^{-1}\tilde{\mathbf{w}}_{it'}^{A\prime}\right)^2$$

$$\le \frac{C}{N^2\varepsilon^2}\sum_{i=1}^{N} E\left\|G_N\Sigma_N^{-1/2}\mathbf{K}_N^{-1}\sum_{t'=1}^{T}\tilde{\mathbf{w}}_{it'}^{A\prime}\right\|^4 \le \frac{C}{N^2\varepsilon^2}\sum_{i=1}^{N} E\left\|G_N\right\|^4\left\|\Sigma_N^{-1/2}\,\mathbf{K}_N^{-1}\right\|^4\left\|\sum_{t'=1}^{T}\tilde{\mathbf{w}}_{it'}^{A\prime}\right\|^4$$

$$\le \frac{C}{N\varepsilon^2}\left(\frac{1}{N}\sum_{i=1}^{N} E\left\|\sum_{t=1}^{T}\tilde{\mathbf{w}}_{it}^{A}\right\|^4\right) = O_p\left(\frac{q_N^2}{N}\right) = o_p\left(1\right)$$

where the last inequality follows from $\lambda_{\max}\left(G_N' G_N\right) = \lambda_{\max}\left(G_N G_N'\right) \to c$ as $N \to \infty$. ■

Lemma C.1.2 (i) Assume Assumption 1–6. Then, for any finite constant M,

$$\sup_{\left\|\boldsymbol{\delta}-\tilde{\boldsymbol{\delta}}\right\| \le M}\left|\sum_{i=1}^{N} E_w\left[\tilde{Q}_i\left(\boldsymbol{\delta},\tilde{\boldsymbol{\delta}}\right)\right] - \frac{1}{2}\left[\boldsymbol{\delta}'\mathbf{K}_N\boldsymbol{\delta} - \tilde{\boldsymbol{\delta}}'\mathbf{K}_N\tilde{\boldsymbol{\delta}}\right]\left(1 + o\left(1\right)\right)\right| = o_p\left(1\right)$$

where $\tilde{Q}_i\left(\boldsymbol{\delta},\tilde{\boldsymbol{\delta}}\right) = \sum_{t=1}^{T}\left[\rho_\tau\left(e_{it} - \check{\mathbf{w}}_{it}\boldsymbol{\delta}\right) - \rho_\tau\left(e_{it} - \check{\mathbf{w}}_{it}\tilde{\boldsymbol{\delta}}\right)\right]$

Proof. Note that $\left\|\check{\mathbf{w}}_{it}\tilde{\boldsymbol{\delta}}\right\| = O_p\left(N^{-1/2}q_N\right)$ by Lemma C.1.1 (i), and $\left\|\check{\mathbf{w}}_{it}\boldsymbol{\delta}\right\| = O\left(N^{-1/2}q_N\right)$ by construction. Also, we have $\rho_\tau\left(e_{it} - \check{\mathbf{w}}_{it}\boldsymbol{\delta}\right) = \rho_\tau\left(\varepsilon_{it} - \check{\mathbf{w}}_{it}\boldsymbol{\delta}\right)$ by Assumption 6. Then, applying Knight's identity,

$$\sum_{i=1}^{N} E_w\left[\tilde{Q}_i\left(\boldsymbol{\delta},\tilde{\boldsymbol{\delta}}\right)\right] = \sum_{i=1}^{N}\sum_{t=1}^{T} E_w\left[\rho_\tau\left(\varepsilon_{it} - \check{\mathbf{w}}_{it}\boldsymbol{\delta}\right) - \rho_\tau\left(\varepsilon_{it} - \check{\mathbf{w}}_{it}\tilde{\boldsymbol{\delta}}\right)\right]$$

$$= \sum_{i=1}^{N}\sum_{t=1}^{T}\int_{\check{\mathbf{w}}_{it}\tilde{\boldsymbol{\delta}}}^{\check{\mathbf{w}}_{it}\boldsymbol{\delta}}\left(F_{it}\left(s\right) - F_{it}\left(0\right)\right)ds = \frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T} f_{it}\left(0\right)\left[\left(\check{\mathbf{w}}_{it}\boldsymbol{\delta}\right)^2 - \left(\check{\mathbf{w}}_{it}\tilde{\boldsymbol{\delta}}\right)^2\right]\left(1 + o_p\left(1\right)\right)$$

$$= \frac{1}{2}\left[\boldsymbol{\delta}'\mathbf{K}_N\boldsymbol{\delta} - \tilde{\boldsymbol{\delta}}'\mathbf{K}_N\tilde{\boldsymbol{\delta}}\right]\left(1 + o_p\left(1\right)\right)$$

where the third inequality is followed from $\left\|\check{\mathbf{w}}_{it}\tilde{\boldsymbol{\delta}}\right\| = o_p(1)$ and $\|\check{\mathbf{w}}_{it}\boldsymbol{\delta}\| = o(1)$. Hence the result. ■

Lemma C.1.3 Assume Assumption 1–6. Then, for any given positive constant M,

$$\sup_{\left\|\boldsymbol{\delta}-\tilde{\boldsymbol{\delta}}\right\|\leq M}\left|\sum_{i=1}^{N} A_i\left(\boldsymbol{\delta},\tilde{\boldsymbol{\delta}}\right)\right| = o_p(1)$$

where $A_i\left(\boldsymbol{\delta},\tilde{\boldsymbol{\delta}}\right) = \tilde{Q}_i\left(\boldsymbol{\delta},\tilde{\boldsymbol{\delta}}\right) - E\left[\tilde{Q}_i\left(\boldsymbol{\delta},\tilde{\boldsymbol{\delta}}\right)|\mathbf{x}_i,\mathbf{z}_i\right] + \sum_{t=1}^{T}\check{\mathbf{w}}_{it}\left(\boldsymbol{\delta}-\tilde{\boldsymbol{\delta}}\right)\psi\left(e_{it}\right)$.

Proof. By Assumption 4, $\max_{i,t}\|\check{\mathbf{w}}_{it}\| \leq \alpha_1\sqrt{\frac{q_N}{N}}$ for some positive constant α_1. (F_{n1} in Sherwood and Wang (2016) has probability one here.) It suffices to show for $\forall\varepsilon > 0$,

$$P\left(\sup_{\left\|\boldsymbol{\delta}-\tilde{\boldsymbol{\delta}}\right\|\leq M}\left|\sum_{i=1}^{N} A_i\left(\boldsymbol{\delta},\tilde{\boldsymbol{\delta}}\right)\right| > \varepsilon\right) \to 0$$

Let

$$\bar{\Delta} = \left\{\boldsymbol{\delta}\in\mathbb{R}^{q_N} : \left\|\boldsymbol{\delta}-\tilde{\boldsymbol{\delta}}\right\|\leq M\right\}$$

We can partition $\bar{\Delta}$ into disjoint sets $\bar{\Delta}_1,\cdots\bar{\Delta}_{D_N}$ such that the diameter of each set does not exceed $m_0^* = \frac{\varepsilon}{10T\alpha_1\sqrt{Nq_N}}$ and the cardinality of partition satisfies $D_N \leq \left(\frac{C\sqrt{Nq_N}}{2\varepsilon}\right)^{q_N}$ (For example, by similar argument used in Lemma 5.2 of Vershynin, 2011). Pick arbitrary $\boldsymbol{\delta}_d \in D_d$ for $1 \leq d \leq D_N$. Then

$$\begin{aligned}
&P\left(\sup_{\left\|\boldsymbol{\delta}-\tilde{\boldsymbol{\delta}}\right\|\leq M}\left|\sum_{i=1}^{N} A_i\left(\boldsymbol{\delta},\tilde{\boldsymbol{\delta}}\right)\right| > \varepsilon\right)\\
&\leq \sum_{d=1}^{D_N} P\left(\left|\sum_{i=1}^{N} A_i\left(\boldsymbol{\delta}_d,\tilde{\boldsymbol{\delta}}\right)\right| + \sup_{\boldsymbol{\delta}\in\bar{\Delta}_d}\left|\sum_{i=1}^{N}\left[A_i\left(\boldsymbol{\delta},\tilde{\boldsymbol{\delta}}\right) - A_i\left(\boldsymbol{\delta}_d,\tilde{\boldsymbol{\delta}}\right)\right]\right| > \varepsilon\right)
\end{aligned}$$

Since $u1\left[u<0\right]=\frac{1}{2}u-\frac{1}{2}\left|u\right|$, we have

$$\begin{aligned}
&\tilde{Q}_i\left(\boldsymbol{\delta},\tilde{\boldsymbol{\delta}}\right)-\tilde{Q}_i\left(\boldsymbol{\delta}_d,\tilde{\boldsymbol{\delta}}\right)\\
&=\sum_{t=1}^{T}\left[\rho_\tau\left(e_{it}-\check{\mathbf{w}}_{it}\boldsymbol{\delta}\right)-\rho_\tau\left(e_{it}-\check{\mathbf{w}}_{it}\tilde{\boldsymbol{\delta}}\right)\right]-\sum_{t=1}^{T}\left[\rho_\tau\left(e_{it}-\check{\mathbf{w}}_{it}\boldsymbol{\delta}_d\right)-\rho_\tau\left(e_{it}-\check{\mathbf{w}}_{it}\tilde{\boldsymbol{\delta}}\right)\right]\\
&=\sum_{t=1}^{T}\left[\left(\tau-1\left[e_{it}-\check{\mathbf{w}}_{it}\boldsymbol{\delta}<\mathbf{0}\right]\right)\left(e_{it}-\check{\mathbf{w}}_{it}\boldsymbol{\delta}\right)-\left(\tau-1\left[e_{it}-\check{\mathbf{w}}_{it}\boldsymbol{\delta}_d<\mathbf{0}\right]\right)\left(e_{it}-\check{\mathbf{w}}_{it}\boldsymbol{\delta}_d\right)\right]\\
&=\sum_{t=1}^{T}\left[\tau\check{\mathbf{w}}_{it}\left(\boldsymbol{\delta}_d-\boldsymbol{\delta}\right)-\left(e_{it}-\check{\mathbf{w}}_{it}\boldsymbol{\delta}\right)1\left[e_{it}-\check{\mathbf{w}}_{it}\boldsymbol{\delta}<\mathbf{0}\right]+\left(e_{it}-\check{\mathbf{w}}_{it}\boldsymbol{\delta}_d\right)1\left[e_{it}-\check{\mathbf{w}}_{it}\boldsymbol{\delta}_d<\mathbf{0}\right]\right]\\
&=\sum_{t=1}^{T}\left[\left(\tau-\frac{1}{2}\right)\check{\mathbf{w}}_{it}\left(\boldsymbol{\delta}_d-\boldsymbol{\delta}\right)+\frac{1}{2}\left|\left(e_{it}-\check{\mathbf{w}}_{it}\boldsymbol{\delta}\right)\right|-\frac{1}{2}\left|\left(e_{it}-\check{\mathbf{w}}_{it}\boldsymbol{\delta}_d\right)\right|\right]\\
&\leq 2T\max_{i,t}\left\|\check{\mathbf{w}}_{it}\right\|\sup_{\boldsymbol{\delta}\in\bar{\Delta}_d}\left[\left\|\boldsymbol{\delta}-\boldsymbol{\delta}_d\right\|\right]
\end{aligned}$$

Thus

$$\begin{aligned}
&\sup_{\boldsymbol{\delta}\in\bar{\Delta}_d}\left|\sum_{i=1}^{N}\left[A_i\left(\boldsymbol{\delta},\tilde{\boldsymbol{\delta}}\right)-A_i\left(\boldsymbol{\delta}_d,\tilde{\boldsymbol{\delta}}\right)\right]\right|\\
&=\sup_{\boldsymbol{\delta}\in\bar{\Delta}_d}\left|\sum_{i=1}^{N}\{\tilde{Q}_i\left(\boldsymbol{\delta},\tilde{\boldsymbol{\delta}}\right)-E_w\left[\tilde{Q}_i\left(\boldsymbol{\delta},\tilde{\boldsymbol{\delta}}\right)\right]+\sum_{t=1}^{T}\check{\mathbf{w}}_{it}\left(\boldsymbol{\delta}-\tilde{\boldsymbol{\delta}}\right)\psi\left(e_{it}\right)\right.\\
&\left.-\tilde{Q}_i\left(\boldsymbol{\delta}_d,\tilde{\boldsymbol{\delta}}\right)+E_w\left[\tilde{Q}_i\left(\boldsymbol{\delta}_d,\tilde{\boldsymbol{\delta}}\right)\right]-\sum_{t=1}^{T}\check{\mathbf{w}}_{it}\left(\boldsymbol{\delta}_d-\tilde{\boldsymbol{\delta}}\right)\psi\left(e_{it}\right)\}\right|\\
&\leq 5NT\max_{i,t}\left\|\check{\mathbf{w}}_{it}\right\|\sup_{\boldsymbol{\delta}\in\bar{\Delta}_d}\left[\left\|\boldsymbol{\delta}-\boldsymbol{\delta}_d\right\|\right]\\
&\leq 5NT\alpha_1\sqrt{\frac{q_N}{N}}m_0=\frac{\varepsilon}{2}
\end{aligned}$$

Therefore, now it suffices to show $\sum_{d=1}^{D_N}P_w\left(\left|\sum_{i=1}^{N}A_i(\boldsymbol{\delta}_d,\tilde{\boldsymbol{\delta}})\right|>\frac{\varepsilon}{2}\right)$ has a vanishing upper bound that does not depend on $(\mathbf{x}_i,\mathbf{z}_i)$. Berstein's inequality is used. To evaluate maximum, using

$\rho\left(u\right) = \left(\tau - \frac{1}{2}\right) u + \frac{1}{2}\left|u\right|$, we can write

$$
\begin{aligned}
&\max_{i,d} A_i\left(\boldsymbol{\delta}_d, \tilde{\boldsymbol{\delta}}\right) \\
&= \max_{i,d} \tilde{Q}_i\left(\boldsymbol{\delta}_d, \tilde{\boldsymbol{\delta}}\right) - E_w[\tilde{Q}_i\left(\boldsymbol{\delta}_d, \tilde{\boldsymbol{\delta}}\right)] + \sum_{t=1}^{T} \check{\mathbf{w}}_{it}(\boldsymbol{\delta}_d - \tilde{\boldsymbol{\delta}})\psi\left(e_{it}\right) \\
&= \max_{i,d} \sum_{t=1}^{T} [\rho_\tau\left(e_{it} - \check{\mathbf{w}}_{it}\boldsymbol{\delta}_d\right) - \rho_\tau(e_{it} - \check{\mathbf{w}}_{it}\tilde{\boldsymbol{\delta}})] - E_w[\sum_{t=1}^{T}\left[\rho_\tau\left(e_{it} - \check{\mathbf{w}}_{it}\boldsymbol{\delta}_d\right) - \rho_\tau(e_{it} - \check{\mathbf{w}}_{it}\tilde{\boldsymbol{\delta}})\right]] \\
&+ \sum_{t=1}^{T} \check{\mathbf{w}}_{it}\left(\boldsymbol{\delta}_d - \tilde{\boldsymbol{\delta}}\right)\psi\left(e_{it}\right) \\
&= \max_{i,d} \sum_{t=1}^{T} \left[\frac{1}{2}\left[|(e_{it} - \check{\mathbf{w}}_{it}\boldsymbol{\delta}_d)| - \left|(e_{it} - \check{\mathbf{w}}_{it}\tilde{\boldsymbol{\delta}})\right|\right] + \left(\tau - \frac{1}{2}\right)\check{\mathbf{w}}_{it}(\tilde{\boldsymbol{\delta}} - \boldsymbol{\delta}_d)\right] \\
&- E_w \sum_{t=1}^{T} \left[\frac{1}{2}\left[|(e_{it} - \check{\mathbf{w}}_{it}\boldsymbol{\delta}_d)| - \left|(e_{it} - \check{\mathbf{w}}_{it}\tilde{\boldsymbol{\delta}})\right|\right] + \left(\tau - \frac{1}{2}\right)\check{\mathbf{w}}_{it}(\tilde{\boldsymbol{\delta}} - \boldsymbol{\delta}_d)\right] \\
&+ \sum_{t=1}^{T} \check{\mathbf{w}}_{it}(\boldsymbol{\delta}_d - \tilde{\boldsymbol{\delta}})\psi\left(e_{it}\right) \\
&\leq 3T \max_{i,t}\left\|\check{\mathbf{w}}_{it}\right\| \max_{d}\left\|\boldsymbol{\delta} - \boldsymbol{\delta}_d\right\| \leq C\sqrt{\frac{q_N}{N}}
\end{aligned}
$$

To evaluate variance, applying Knight's identity, we have

$$
\begin{aligned}
&\tilde{Q}_i\left(\boldsymbol{\delta}_d, \tilde{\boldsymbol{\delta}}\right) + \sum_{t=1}^{T} \check{\mathbf{w}}_{it}\left(\boldsymbol{\delta}_d - \tilde{\boldsymbol{\delta}}\right)\psi\left(e_{it}\right) \\
&= \sum_{t=1}^{T} \rho_\tau\left(e_{it} - \check{\mathbf{w}}_{it}\boldsymbol{\delta}_d\right) - \sum_{t=1}^{T} \rho_\tau\left(e_{it} - \check{\mathbf{w}}_{it}\tilde{\boldsymbol{\delta}}\right) + \sum_{t=1}^{T} \check{\mathbf{w}}_{it}\left(\boldsymbol{\delta}_d - \tilde{\boldsymbol{\delta}}\right)\psi\left(e_{it}\right) \\
&= -\sum_{t=1}^{T} \check{\mathbf{w}}_{it}\boldsymbol{\delta}_d\psi_\tau\left(e_{it}\right) + \sum_{t=1}^{T} \int_{0}^{\check{\mathbf{w}}_{it}\boldsymbol{\delta}_d} \left(1\left[\left(e_{it} < t\right)\right] - 1\left[e_{it} < 0\right]\right) dt \\
&+ \sum_{t=1}^{T} \check{\mathbf{w}}_{it}\tilde{\boldsymbol{\delta}}\psi_\tau\left(e_{it}\right) - \sum_{t=1}^{T} \int_{0}^{\check{\mathbf{w}}_{it}\tilde{\boldsymbol{\delta}}} \left(1\left[\left(e_{it} < t\right)\right] - 1\left[e_{it} < 0\right]\right) dt + \sum_{t=1}^{T} \check{\mathbf{w}}_{it}\left(\boldsymbol{\delta}_d - \tilde{\boldsymbol{\delta}}\right)\psi\left(e_{it}\right)
\end{aligned}
$$

$$
= \sum_{t=1}^{T} \int_{\check{\mathbf{w}}_{it}\tilde{\boldsymbol{\delta}}}^{\check{\mathbf{w}}_{it}\boldsymbol{\delta}_d} \left(1\left[\left(e_{it} < t\right)\right] - 1\left[e_{it} < 0\right]\right) dt
$$

Thus,

$$
\begin{aligned}
& A_i(\boldsymbol{\delta}_d, \tilde{\boldsymbol{\delta}}) \\
& = \sum_{t=1}^{T} \int_{\check{\mathbf{w}}_{it}\tilde{\boldsymbol{\delta}}}^{\check{\mathbf{w}}_{it}\boldsymbol{\delta}_d} \left(1\left[(e_{it} < t)\right] - 1\left[e_{it} < 0\right]\right) dt \\
& - E_w\left[\sum_{t=1}^{T} \int_{\check{\mathbf{w}}_{it}\tilde{\boldsymbol{\delta}}}^{\check{\mathbf{w}}_{it}\boldsymbol{\delta}_d} \left(1\left[(e_{it} < t)\right] - 1\left[e_{it} < 0\right]\right) dt\right] + E_w\left[\sum_{t=1}^{T} \check{\mathbf{w}}_{it}\left(\boldsymbol{\delta}_d - \tilde{\boldsymbol{\delta}}\right)\psi\left(e_{it}\right)\right]
\end{aligned}
$$

And it implies

$$
\begin{aligned}
& \sum_{i=1}^{N} Var\left(A_i\left(\boldsymbol{\delta}_d, \tilde{\boldsymbol{\delta}}\right)|\mathbf{x}_i, \mathbf{z}_i\right) \leq \sum_{i=1}^{N} E_w\left[\left(\sum_{t=1}^{T} \int_{\check{\mathbf{w}}_{it}\tilde{\boldsymbol{\delta}}}^{\check{\mathbf{w}}_{it}\boldsymbol{\delta}_d} \left(1\left[(e_{it} < t)\right] - 1\left[e_{it} < 0\right]\right) dt\right)^2\right] \\
& \leq \sum_{i=1}^{N} E_w\left[\max_{i,t}\left|\check{\mathbf{w}}_{it}\left(\boldsymbol{\delta}_d - \tilde{\boldsymbol{\delta}}\right)\right| \sum_{t=1}^{T} \int_{\check{\mathbf{w}}_{it}\tilde{\boldsymbol{\delta}}}^{\check{\mathbf{w}}_{it}\boldsymbol{\delta}_d} \left(F_{it}(t) - F_{it}(0)\right) dt\right] \\
& \leq C\sqrt{\frac{q_N}{N}} \sum_{i=1}^{N}\sum_{t=1}^{T} f_{it}(0)\left(\left(\check{\mathbf{w}}_{it}\boldsymbol{\delta}_d\right)^2 - \left(\check{\mathbf{w}}_{it}\tilde{\boldsymbol{\delta}}\right)^2\right)(1 + o(1)) \leq C\sqrt{\frac{q_N}{N}}(1 + o(1))
\end{aligned}
$$

by similar argument as in Lemma C.1.2. Then, by Bernstein's inequality and Assumption 5,

$$
\begin{aligned}
& \sum_{d=1}^{D_N} P_w\left(\left|\sum_{i=1}^{N} A_i\left(\boldsymbol{\delta}_d, \tilde{\boldsymbol{\delta}}\right)\right| > \frac{\varepsilon}{2}\right) \leq \sum_{d=1}^{D_N} exp\left(\frac{-\varepsilon^2/4}{C\sqrt{\frac{q_N}{N}} + \varepsilon C\sqrt{\frac{q_N}{N}}}\right) \leq \sum_{d=1}^{D_N} exp\left(-C\sqrt{\frac{N}{q_N}}\right) \\
& \leq C\left(C\sqrt{Nq_N}\right)^{q_N} exp\left(-C\sqrt{\frac{N}{q_N}}\right) \leq Cexp\left(C\left(q_N \log N - \sqrt{\frac{N}{q_N}}\right)\right) \to 0
\end{aligned}
$$

■

Lemma C.1.4 (Asymptotic Equivalence with Bahadur Representation) Assume Assumption 1–6. Then, we have $\left\|\tilde{\boldsymbol{\delta}} - \hat{\boldsymbol{\delta}}\right\| = o_p(1)$.

Proof. It suffices to show that for any positive constants M,

$$
P\left(\inf_{\left\|\tilde{\boldsymbol{\delta}} - \boldsymbol{\delta}\right\| \geq M} \sum_{i=1}^{N} \tilde{Q}_i\left(\boldsymbol{\delta}, \tilde{\boldsymbol{\delta}}\right) > 0\right) \to 1
$$

since, we have $\sum_{i=1}^{N} \tilde{Q}_i \left(\hat{\boldsymbol{\delta}}, \tilde{\boldsymbol{\delta}}\right) \leq 0.$ By Lemma C.1.3,

$$\sup_{\left\|\boldsymbol{\delta}-\tilde{\boldsymbol{\delta}}\right\| \leq M} \left| \sum_{i=1}^{N} \left[\tilde{Q}_i \left(\boldsymbol{\delta}, \tilde{\boldsymbol{\delta}}\right) - E_w \left[\tilde{Q}_i \left(\boldsymbol{\delta}, \tilde{\boldsymbol{\delta}}\right)\right] + \sum_{t=1}^{T} \check{\mathbf{w}}_{it} \left(\boldsymbol{\delta} - \tilde{\boldsymbol{\delta}}\right) \psi\left(e_{it}\right) \right] \right| = o_p(1)$$

Then by Lemma C.1.2,

$$\begin{aligned} \sup_{\left\|\boldsymbol{\delta}-\tilde{\boldsymbol{\delta}}\right\| \leq M} & \left| \sum_{i=1}^{N} \tilde{Q}_i \left(\boldsymbol{\delta}, \tilde{\boldsymbol{\delta}}\right) - \frac{1}{2} \left[\boldsymbol{\delta}'\mathbf{K}_N \boldsymbol{\delta} - \tilde{\boldsymbol{\delta}}'\mathbf{K}_N \tilde{\boldsymbol{\delta}}\right] \left(1 + o_p(1)\right) \right. \\ & \left. \quad + \sum_{i=1}^{N} \sum_{t=1}^{T} \check{\mathbf{w}}_{it} \left(\boldsymbol{\delta} - \tilde{\boldsymbol{\delta}}\right) \psi\left(e_{it}\right) \right| \\ & = o_p(1) \end{aligned} \tag{A.5}$$

And since

$$\begin{aligned} \tilde{\boldsymbol{\delta}} &= \left(\frac{1}{N} \tilde{\mathbf{W}}_A' B_N \tilde{\mathbf{W}}_A\right)^{-1} \frac{1}{\sqrt{N}} \tilde{\mathbf{W}}_A' \boldsymbol{\Psi}_\tau\left(\boldsymbol{e}\right) \\ &= \mathbf{K}_N^{-1} \frac{1}{\sqrt{N}} \tilde{\mathbf{W}}_A' \boldsymbol{\Psi}_\tau\left(\boldsymbol{e}\right) \end{aligned}$$

we have

$$\begin{aligned} \sum_{i=1}^{N} \sum_{t=1}^{T} \check{\mathbf{w}}_{it} \left(\boldsymbol{\delta} - \tilde{\boldsymbol{\delta}}\right) \psi\left(e_{it}\right) &= \left(\boldsymbol{\delta} - \tilde{\boldsymbol{\delta}}\right)' \sum_{i=1}^{N} \sum_{t=1}^{T} \check{\mathbf{w}}_{it}' \psi\left(e_{it}\right) \\ &= \left(\boldsymbol{\delta} - \tilde{\boldsymbol{\delta}}\right)' \frac{1}{\sqrt{N}} \tilde{\mathbf{W}}_A' \boldsymbol{\Psi}_\tau\left(\boldsymbol{e}\right) \\ &= \left(\boldsymbol{\delta} - \tilde{\boldsymbol{\delta}}\right)' \mathbf{K}_N \tilde{\boldsymbol{\delta}} \end{aligned} \tag{A.6}$$

Combining (A.5) and (A.6), we have

$$\sup_{\left\|\boldsymbol{\delta}-\tilde{\boldsymbol{\delta}}\right\| \leq M} \left| \sum_{i=1}^{N} \tilde{Q}_i \left(\boldsymbol{\delta}, \tilde{\boldsymbol{\delta}}\right) - \frac{1}{2} \left[\boldsymbol{\delta}'\mathbf{K}_N \boldsymbol{\delta} - \tilde{\boldsymbol{\delta}}'\mathbf{K}_N \tilde{\boldsymbol{\delta}}\right] + \left(\boldsymbol{\delta} - \tilde{\boldsymbol{\delta}}\right)' \mathbf{K}_N \tilde{\boldsymbol{\delta}} \right| = o_p(1)$$

which implies

$$\sup_{\left\|\boldsymbol{\delta}-\tilde{\boldsymbol{\delta}}\right\| \leq M} \left| \sum_{i=1}^{N} \tilde{Q}_i \left(\boldsymbol{\delta}, \tilde{\boldsymbol{\delta}}\right) - \frac{1}{2} \left(\boldsymbol{\delta} - \tilde{\boldsymbol{\delta}}\right)' \mathbf{K}_N \left(\boldsymbol{\delta} - \tilde{\boldsymbol{\delta}}\right) \right| = o_p(1)$$

By assumption 4, for any $\left\| \boldsymbol{\delta} - \tilde{\boldsymbol{\delta}} \right\| > M$,

$$\frac{1}{2}\left(\boldsymbol{\delta} - \tilde{\boldsymbol{\delta}}\right)' \mathbf{K}_N \left(\boldsymbol{\delta} - \tilde{\boldsymbol{\delta}}\right) > CM$$

for some positive C. ■

C.1.3 Proof of Theorem 3.4.5 and 3.4.6

Consider partition of $\tilde{\mathbf{w}}_{it}^{A}$ into a single variable (that is not penalized) and the rest of variables, $(\tilde{w}_{it}^{b}, \tilde{\mathbf{w}}_{it}^{a})$. Define their stacked versions as $\tilde{\mathbf{W}}_{A}^{a} = \left(\tilde{\mathbf{w}}_{11}^{a\prime}, \cdots, \tilde{\mathbf{w}}_{1T}^{a\prime}, \cdots, \tilde{\mathbf{w}}_{NT}^{a\prime}\right)'$ and $W_A^b = (\tilde{w}_{11}^{b}, \cdots, \tilde{w}_{1T}^{b}, \cdots, \tilde{w}_{NT}^{b})$. Let $\boldsymbol{\delta}_{ab} = (\boldsymbol{\delta}_a, \delta_b)$ denote another reparametrization of θ_A such that

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\rho_\tau\left(y_{it} - \tilde{\mathbf{w}}_{it}^{A}\boldsymbol{\theta}_A\right) = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\rho_\tau\left(\varepsilon_{it} + r_i - \ddot{\mathbf{w}}_{it}^{a}\boldsymbol{\delta}_a - \ddot{w}_{it}^{b}\delta_b\right)$$

where $\ddot{\mathbf{w}}_{it}^{a} = \frac{1}{\sqrt{N}}\left[\tilde{\mathbf{w}}_{it}^{a} - \ddot{w}_{it}^{b}\left(W_A^{b\prime}B_N W_A^{b}\right)^{-1} W_A^{b\prime}B_N\tilde{\mathbf{W}}_A^{a}\right]$ and $\ddot{w}_{it}^{b} = \left[W_A^{b\prime}B_N W_A^{b\prime}\right]^{-1/2}\tilde{w}_{it}^{b}$. Then, we have $\hat{\boldsymbol{\delta}}_a = \sqrt{N}(\hat{\boldsymbol{\theta}}_A^{a} - \boldsymbol{\theta}_{oA}^{a})$ and

$$\hat{\delta}_b = \left[W_A^{b\prime}B_N W_A^{b\prime}\right]^{1/2}\left(\hat{\theta}_A^{b} - \theta_{oA}^{b}\right) + \left[W_A^{b\prime}B_N W_A^{b\prime}\right]^{-1/2} W_A^{b\prime}B_N\tilde{\mathbf{W}}_A^{a}\left(\hat{\boldsymbol{\theta}}_A^{a} - \boldsymbol{\theta}_{oA}^{a}\right)$$

where $\left(\boldsymbol{\theta}_A^{a}, \theta_A^{b}\right)$ is defined similarly. Also, note that $\hat{\boldsymbol{\delta}}_a$ is a subvector of $\hat{\boldsymbol{\delta}}$.

Lemma C.1.5 Assume Assumption 1–5, 7 and 8. Then, for any finite constant M_1 and M_2,

$$\sup_{\left\|\boldsymbol{\delta}_a - \tilde{\boldsymbol{\delta}}_a\right\| \le M_1, \|\delta_b\| \le M_2\sqrt{q_N}}\left|\sum_{i=1}^{N} E_w\left[\tilde{Q}_i\left(\boldsymbol{\delta}_a, \tilde{\boldsymbol{\delta}}_a, \delta_b\right)\right] - \frac{1}{2}\left[\boldsymbol{\delta}_a'\ddot{\mathbf{K}}_N\boldsymbol{\delta}_a - \tilde{\boldsymbol{\delta}}_a'\,\ddot{\mathbf{K}}_N\tilde{\boldsymbol{\delta}}_a\right](1 + o(1))\right|$$

is $o_p(1)$ where

$$\begin{aligned}
\tilde{Q}_i\left(\boldsymbol{\delta}_a, \tilde{\boldsymbol{\delta}}_a, \delta_b\right) &= \sum_{t=1}^{T}[\rho_\tau(\varepsilon_{it} + r_i - \ddot{\mathbf{w}}_{it}^{a}\boldsymbol{\delta}_a - \ddot{w}_{it}^{b}\delta_b) - \rho_\tau(\varepsilon_{it} + r_i - \ddot{\mathbf{w}}_{it}^{a}\tilde{\boldsymbol{\delta}}_a - \ddot{w}_{it}^{b}\delta_b)] \\
\ddot{\mathbf{W}}_a &= \left(\ddot{\mathbf{w}}_{11}^{a\prime}, \cdots, \ddot{\mathbf{w}}_{iT}^{a\prime}, \cdots, \ddot{\mathbf{w}}_{NT}^{a\prime}\right)' \\
\ddot{\mathbf{K}}_N &= \ddot{\mathbf{W}}_a' B_N \ddot{\mathbf{W}}_a
\end{aligned}$$

Proof. Note that $\left\|\check{\mathbf{w}}_{it}\tilde{\boldsymbol{\delta}}\right\| = O_p\left(N^{-1/2}q_N\right)$ is implied by Lemma C.1.1 (i), and $\|\check{\mathbf{w}}_{it}\boldsymbol{\delta}\| = O\left(N^{-1/2}q_N\right)$ by construction. Similarly, we have $\left\|\ddot{\mathbf{w}}_{it}^{a}\tilde{\boldsymbol{\delta}}_a\right\| = O_p\left(N^{-1/2}q_N\right)$ and $\|\ddot{\mathbf{w}}_{it}^{a}\boldsymbol{\delta}_a\| = O_p\left(N^{-1/2}q_N\right)$. Then, applying Knight's identity,

$$\begin{aligned}
&\sum_{i=1}^{N} E_w\left[\tilde{Q}_i\left(\boldsymbol{\delta}_a, \tilde{\boldsymbol{\delta}}_a, \delta_b\right)\right] \\
&= \sum_{i=1}^{N}\sum_{t=1}^{T} E_w[\rho_\tau(\varepsilon_{it} + r_i - \ddot{\mathbf{w}}_{it}^{a}\boldsymbol{\delta}_a - \ddot{w}_b\delta_b) - \rho_\tau(\varepsilon_{it} + r_i - \ddot{\mathbf{w}}_{it}^{a}\tilde{\boldsymbol{\delta}}_a - \ddot{w}_{it}^{b}\delta_b)] \\
&= \sum_{i=1}^{N}\sum_{t=1}^{T} \int_{\ddot{\mathbf{w}}_{it}^{a}\tilde{\boldsymbol{\delta}}_a + \ddot{w}_b\delta_b - r_i}^{\ddot{\mathbf{w}}_{it}^{a}\boldsymbol{\delta}_a + \ddot{w}_b\delta_b - r_i} \left(F_{it}\left(s\right) - F_{it}\left(0\right)\right) ds \\
&= \frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T} f_{it}\left(0\right)\left[\left(\ddot{\mathbf{w}}_{it}^{a}\boldsymbol{\delta}_a + \ddot{w}_{it}^{b}\delta_b - r_i\right)^2 - \left(\ddot{\mathbf{w}}_{it}^{a}\tilde{\boldsymbol{\delta}}_a + \ddot{w}_{it}^{b}\delta_b - r_i\right)^2\right]\left(1 + o_p\left(1\right)\right) \\
&= \frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T} f_{it}\left(0\right)\left[(\ddot{\mathbf{w}}_{it}^{a}\boldsymbol{\delta}_a)^2 - (\ddot{\mathbf{w}}_{it}^{a}\tilde{\boldsymbol{\delta}}_a)^2 + 2(\ddot{w}_{it}^{b}\delta_b - r_i)\ddot{\mathbf{w}}_{it}^{a}(\boldsymbol{\delta}_a - \tilde{\boldsymbol{\delta}}_a)\right]\left(1 + o_p\left(1\right)\right) \\
&= \frac{1}{2}\left[\boldsymbol{\delta}_a'\ddot{\mathbf{K}}_N\boldsymbol{\delta}_a - \tilde{\boldsymbol{\delta}}_a'\,\ddot{\mathbf{K}}_N\tilde{\boldsymbol{\delta}}_a\right]\left(1 + o_p\left(1\right)\right) - (\boldsymbol{\delta}_a - \tilde{\boldsymbol{\delta}}_a)'\sum_{i=1}^{N}\sum_{t=1}^{T} f_{it}\left(0\right) r_i\ddot{\mathbf{w}}_{it}^{a\prime}
\end{aligned}$$

Note that

$$\begin{aligned}
&\sum_{i=1}^{N}\sum_{t=1}^{T} f_{it}\left(0\right) r_i\ddot{\mathbf{w}}_{it}^{a}(\boldsymbol{\delta}_a - \tilde{\boldsymbol{\delta}}_a) \\
&= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{t=1}^{T} f_{it}\left(0\right) r_i\left[\tilde{\mathbf{w}}_{it}^{a} - \left(W_A^{b\prime}B_N W_A^{b}\right)^{-1} W_A^{b\prime}B_N\tilde{\mathbf{W}}_A^{a}\right](\boldsymbol{\delta}_a - \tilde{\boldsymbol{\delta}}_a)
\end{aligned}$$

and that $\frac{1}{\sqrt{N}}[\tilde{\mathbf{w}}_{it}^{a} - \left(W_A^{b\prime}B_N W_A^{b}\right)^{-1} W_A^{b\prime}B_N\tilde{\mathbf{W}}_A^{a}] = \frac{1}{\sqrt{N}}\ddot{\Delta}_{it}^{a} + o_p\left(1\right)$ where $\ddot{\Delta}_{it}^{a}$ denotes the population projection error. $\ddot{\Delta}_{it}^{a}$ is well-defined by Assumption 5. To show

$$\sup_{\left\|\boldsymbol{\delta}_a - \tilde{\boldsymbol{\delta}}_a\right\| \le M_1, \|\delta_b\| \le M_2\sqrt{q_N}} \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{t=1}^{T} f_{it}\left(0\right) r_i\ddot{\mathbf{w}}_{it}^{a}(\boldsymbol{\delta}_a - \tilde{\boldsymbol{\delta}}_a) = o_p\left(1\right)$$

, note that by Markov inequality, we have

$$P\left(\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{t=1}^{T} f_{it}\left(0\right) r_i\ddot{\Delta}_{it}^{a}(\boldsymbol{\delta}_a - \tilde{\boldsymbol{\delta}}_a)\right| \ge \varepsilon\right) \le \frac{E\left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{t=1}^{T} f_{it}\left(0\right) r_i\ddot{\Delta}_{it}^{a}(\boldsymbol{\delta}_a - \tilde{\boldsymbol{\delta}}_a)\right]^2}{\varepsilon^2}$$

where

$$
\begin{aligned}
&E\left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{t=1}^{T} f_{it}\left(0\right) r_i \ddot{\mathbf{w}}_{it}^{a}(\boldsymbol{\delta}_a-\tilde{\boldsymbol{\delta}}_a)\right]^2 \\
&= V\left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{t=1}^{T} f_{it}\left(0\right) r_i \ddot{\mathbf{w}}_{it}^{a}(\boldsymbol{\delta}_a-\tilde{\boldsymbol{\delta}}_a)\right]+\left(E\left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{t=1}^{T} f_{it}\left(0\right) r_i \ddot{\mathbf{w}}_{it}^{a}(\boldsymbol{\delta}_a-\tilde{\boldsymbol{\delta}}_a)\right]\right)^2
\end{aligned}
$$

The first part:

$$
\begin{aligned}
V\left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{t=1}^{T} f_{it}\left(0\right) r_i \ddot{\mathbf{w}}_{it}^{a}(\boldsymbol{\delta}_a-\tilde{\boldsymbol{\delta}}_a)\right] &= V\left[\sum_{t=1}^{T} f_{it}\left(0\right) r_i \ddot{\mathbf{w}}_{it}^{a}(\boldsymbol{\delta}_a-\tilde{\boldsymbol{\delta}}_a)\right] \\
&\leq C\sum_{t=1}^{T} V\left[f_{it}\left(0\right) r_i \ddot{\mathbf{w}}_{it}^{a}(\boldsymbol{\delta}_a-\tilde{\boldsymbol{\delta}}_a)\right] \\
&\leq C\sum_{t=1}^{T} E\left[f_{it}\left(0\right) r_i \ddot{\mathbf{w}}_{it}^{a}(\boldsymbol{\delta}_a-\tilde{\boldsymbol{\delta}}_a)\right]^2 \\
&\leq C\sum_{t=1}^{T} E\left[f_{it}\left(0\right) \ddot{\mathbf{w}}_{it}^{a}(\boldsymbol{\delta}_a-\tilde{\boldsymbol{\delta}}_a)\right]^2 \left(\sup\left|r_i\right|\right)^2 \\
&\rightarrow 0,
\end{aligned}
$$

The second part:

$$
\begin{aligned}
\left|E\left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{t=1}^{T} f_{it}\left(0\right) r_i \ddot{\mathbf{w}}_{it}^{a}(\boldsymbol{\delta}_a-\tilde{\boldsymbol{\delta}}_a)\right]\right| &= \left|\sqrt{N}\sum_{t=1}^{T} E\left[f_{it}\left(0\right) r_i \ddot{\mathbf{w}}_{it}^{a}(\boldsymbol{\delta}_a-\tilde{\boldsymbol{\delta}}_a)\right]\right| \\
&\leq \sqrt{N}\sum_{t=1}^{T} E\left|f_{it}\left(0\right) r_i \ddot{\mathbf{w}}_{it}^{a}(\boldsymbol{\delta}_a-\tilde{\boldsymbol{\delta}}_a)\right| \\
&\leq \sqrt{N}\sup\left|r_i\right|\sum_{t=1}^{T} E\left|f_{it}\left(0\right) \ddot{\mathbf{w}}_{it}^{a}(\boldsymbol{\delta}_a-\tilde{\boldsymbol{\delta}}_a)\right|
\end{aligned}
$$

Hence the result. ■

Lemma C.1.6 Assume Assumption 1–5, 7 and 8. Then, for any positive constant L,

$$
q_N^{-1}\sup_{\|\boldsymbol{\delta}_{ab}\|\leq L}\left|\sum_{i=1}^{N} D_i\left(\boldsymbol{\delta}_{ab},\sqrt{q_N}\right)\right| = o_p\left(1\right)
$$

where

$$Q_i\left(q_N\right) = \sum_{t=1}^{T} \rho_\tau(\varepsilon_{it} + r_i - q_N \ddot{\mathbf{w}}_{it}^{a} \boldsymbol{\delta}_a - q_N \ddot{w}_{it}^{b} \delta_b)$$

$$D_i(\delta, q_N) = Q_i\left(q_N\right) - Q_i\left(0\right) - E_w\left[Q_i\left(q_N\right) - Q_i\left(0\right)\right]$$
$$+ q_N\left(\ddot{\mathbf{w}}_{it}^{a}\boldsymbol{\delta}_a + \ddot{w}_b\delta_b\right)\psi_\tau\left(\varepsilon_{it}\right)$$

Proof. It suffices to show for all $\varepsilon > 0$,

$$P\left(q_N^{-1} \sup_{\|\boldsymbol{\delta}_{ab}\|\leq 1} \left|\sum_{i=1}^{N} D_i(\boldsymbol{\delta}_{ab}, L\sqrt{q_N})\right| > \varepsilon\right) \to 0$$

First, consider a constant α_1 such that

$$\max\left\|\left(\ddot{\mathbf{w}}_{it}^{a}, \ddot{w}_{it}^{b}\right)\right\| \leq \alpha_1 N^{-1/2} q_N^{1/2}.$$

Partition $\mathcal{B} = \{\boldsymbol{\delta} : \|\boldsymbol{\delta}\| \leq 1\}$ into M_N disjoint sets $B_1, \cdots, B_{M_N}$ with diameter less than $m_0 = \frac{\varepsilon}{4\alpha_1 L\sqrt{N}}$ where $M_N \leq C\left(\frac{C\sqrt{N}}{\varepsilon}\right)^{q_N+1}$. Let $\boldsymbol{d}_m = (\boldsymbol{d}_m^a, d_m^b) \in B_m$ for $1 \leq m \leq M_N$. Then

$$\begin{aligned}
&P\left(q_N^{-1} \sup_{\boldsymbol{\delta}_{ab}\in\mathcal{B}} \left|\sum_{i=1}^{N} D_i(\boldsymbol{\delta}_{ab}, L\sqrt{q_N})\right| > \varepsilon\right) \\
&\leq \sum_{m=1}^{M_N} P\left(q_N^{-1} \sup_{\boldsymbol{\delta}_{ab}\in B_m} \left|\sum_{i=1}^{N} D_i(\boldsymbol{\delta}_{ab}, L\sqrt{q_N})\right| > \varepsilon\right) \\
&\leq \sum_{m=1}^{M_N} P\left(\begin{array}{c} \left|\sum_{i=1}^{N} D_i(\boldsymbol{d}_m, L\sqrt{q_N})\right| \\ + \sup\limits_{\boldsymbol{\delta}_{ab}\in B_m} \left|\sum_{i=1}^{N} D_i(\boldsymbol{\delta}_{ab}, L\sqrt{q_N}) - \sum_{i=1}^{N} D_i(\boldsymbol{d}_m, L\sqrt{q_N})\right| \end{array} > \varepsilon q_N\right)
\end{aligned}$$

Since $u1\left[U<0\right]=\frac{1}{2}u-\frac{1}{2}\left|u\right|$, we can write

$$
\begin{aligned}
&\sup_{\boldsymbol{\delta}_{ab}\in B_m}\left|\sum_{i=1}^{N}D_i(\boldsymbol{\delta}_{ab},L\sqrt{q_N})-\sum_{i=1}^{N}D_i(\boldsymbol{d}_m,L\sqrt{q_N})\right|\\
&=\sup_{\boldsymbol{\delta}_{ab}\in B_m}\left|\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{1}{2}\left[\left|\varepsilon_{it}-\ddot{\mathbf{w}}_{it}^{a}\boldsymbol{\delta}_a L\sqrt{q_N}-\ddot{w}_{it}^{b}\delta_b L\sqrt{q_N}+r_i\right|-\left|\varepsilon_{it}+r_i\right|\right]\right.\\
&-\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{1}{2}E_w\left[\left|\varepsilon_{it}-\ddot{\mathbf{w}}_{it}^{a}\boldsymbol{\delta}_a L\sqrt{q_N}-\ddot{w}_{it}^{b}\delta_b L\sqrt{q_N}+r_i\right|-\left|\varepsilon_{it}+r_i\right|\right]\\
&+\sum_{i=1}^{N}\sum_{t=1}^{T}L\sqrt{q_N}\left(\ddot{\mathbf{w}}_{it}^{a}\boldsymbol{\delta}_a+\ddot{w}_{it}^{b}\delta_b\right)\psi_\tau\left(\varepsilon_{it}\right)\\
&-\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{1}{2}\left[\left|\varepsilon_{it}-\ddot{\mathbf{w}}_{it}^{a}\boldsymbol{d}_m^{a}L\sqrt{q_N}-\ddot{w}_{it}^{b}d_m^{b}L\sqrt{q_N}+r_i\right|-\left|\varepsilon_{it}+r_i\right|\right]\\
&+\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{1}{2}E_w\left[\left|\varepsilon_{it}-\ddot{\mathbf{w}}_{it}^{a}\boldsymbol{d}_m^{a}L\sqrt{q_N}-\ddot{w}_{it}^{b}d_m^{b}L\sqrt{q_N}+r_i\right|-\left|\varepsilon_{it}+r_i\right|\right]\\
&\left.-\sum_{i=1}^{N}\sum_{t=1}^{T}L\sqrt{q_N}\left(\ddot{\mathbf{w}}_{it}^{a}\boldsymbol{d}_m^{a}+\ddot{w}_{it}^{b}d_m^{b}\right)\psi_\tau\left(\varepsilon_{it}\right)\right|\\
&\leq 2NLm_0q_N^{-1/2}\max\left\|\left(\ddot{\mathbf{w}}_{it}^{a},\ddot{w}_{it}^{b}\right)\right\|\\
&\leq 2\alpha_1NLm_0q_N^{-1/2}\sqrt{q_N/N}=2\alpha_1L\sqrt{N}m_0=\varepsilon/2
\end{aligned}
$$

Now, it suffices to show

$$
\sum_{m=1}^{M_N}P\left(\left|\sum_{i=1}^{N}D_i(\boldsymbol{d}_m,L\sqrt{q_N})\right|>\varepsilon q_N/2\right)\to 0
$$

Bernstein's inequality will be used. To evaluate the maximum, note

$$
\begin{aligned}
&\max_i\left|D_i(\boldsymbol{d}_m,L\sqrt{q_N})\right|\\
&\leq\max_i\left|\left|\varepsilon_{it}-\ddot{\mathbf{w}}_{it}^{a}\boldsymbol{\delta}_a L\sqrt{q_N}-\ddot{w}_{it}^{b}\delta_b L\sqrt{q_N}+r_i\right|-\left|\varepsilon_{it}+r_i\right|\right|\\
&+\max_i\left|L\sqrt{q_N}\left(\ddot{\mathbf{w}}_{it}^{a}\boldsymbol{\delta}_a L\sqrt{q_N}+\ddot{w}_{it}^{b}\delta_b L\sqrt{q_N}\right)\psi_\tau\left(\varepsilon_{it}\right)\right|\\
&\leq 2L\sqrt{q_N}\max\left\|\left(\ddot{\mathbf{w}}_{it}^{a},\ddot{w}_{it}^{b}\right)\right\|\leq Cq_NN^{-1/2}
\end{aligned}
$$

To evaluate an upper bound of the variance, first note that, by Knight's identity,

$$\begin{aligned}
& Q_i\left(L\sqrt{q_N}\right) - Q_i\left(0\right) + L\sqrt{q_N}\left(\ddot{\mathbf{w}}_{it}^{a}\boldsymbol{\delta}_a + \ddot{w}_{it}^{b}\delta_b\right)\psi_\tau\left(\varepsilon_{it}\right) \\
& = L\sqrt{q_N}\left(\ddot{\mathbf{w}}_{it}^{a}\boldsymbol{\delta}_a + \ddot{w}_{it}^{b}\delta_b\right)\left[I\left(\varepsilon_{it} + r_i < 0\right) - I\left(\varepsilon_{it} < 0\right)\right] \\
& + \int_0^{L\sqrt{q_N}\left(\ddot{\mathbf{w}}_{it}^{a}\boldsymbol{\delta}_a + \ddot{w}_{it}^{b}\delta_b\right)}\left[I\left(\varepsilon_{it} + r_i < s\right) - I\left(\varepsilon_{it} - r_i < 0\right)\right] ds \\
& = V_{i1} + V_{i2}
\end{aligned}$$

Then, the second order conditional moments can be bounded as

$$\begin{aligned}
\sum_{i=1}^{N} E_w\left[V_{i1}^2\right] &= \sum_{i=1}^{N} E_w\left[q_N L^2\left(\ddot{\mathbf{w}}_{it}^{a}\boldsymbol{\delta}_a + \ddot{w}_{it}^{b}\delta_b\right)^2\left|I\left(\varepsilon_{it} + r_i < 0\right) - I\left(\varepsilon_{it} < 0\right)\right|\right] \\
&\leq 2L^2 q_N \sum_{i=1}^{N} E_w\left[\left(\max\left\|\left(\ddot{\mathbf{w}}_{it}^{a}, \ddot{w}_{it}^{b}\right)\right\|\right)^2 I\left(0 \leq \left|\varepsilon_{it}\right| \leq \left|r_i\right|\right)\right] \\
&\leq C q_N^2 N^{-1}\sum_{i=1}^{N}\int_{-\left|r_i\right|}^{\left|r_i\right|} f_{it}\left(s\right) ds \leq C q_N^2\sqrt{\frac{q_N}{N}}
\end{aligned}$$

and

$$\begin{aligned}
& \sum_{i=1}^{N} E_w\left[V_{i2}^2\right] \\
& \leq C q_N N^{-1/2}\sum_{i=1}^{N}\int_0^{\sqrt{q_N}L\left(\ddot{\mathbf{w}}_{it}^{a}\boldsymbol{\delta}_a + \ddot{w}_{it}^{b}\delta_b\right)}\left[F_{it}\left(s - r_i\right) - F_{it}\left(r_i\right)\right] ds \\
& \leq C q_N N^{-1/2}\sum_{i=1}^{N}\int_0^{\sqrt{q_N}L\left(\ddot{\mathbf{w}}_{it}^{a}\boldsymbol{\delta}_a + \ddot{w}_{it}^{b}\delta_b\right)}\left(f_{it}\left(0\right) + o\left(1\right)\right)\left(s + O\left(s^2\right)\right) ds
\end{aligned}$$

$$\begin{aligned}
& \leq C q_N^2 N^{-1/2}\left[\boldsymbol{\delta}_a'\sum_{i=1}^{N} f_{it}\left(0\right)\ddot{\mathbf{w}}_{it}^{a\prime}\ddot{\mathbf{w}}_{it}^{a}\boldsymbol{\delta}_a + \sum_{i=1}^{N} f_{it}\left(0\right)\left(\ddot{w}_{it}^{b}\delta_b\right)^2\right]\left(1 + o\left(1\right)\right) \\
& \leq C q_N^2 N^{-1/2}\left(1 + o\left(1\right)\right)
\end{aligned}$$

Since bounds do not depend on $\tilde{w}$,we have

$$\sum_{i=1}^{N} Var_s\left(D_i(\boldsymbol{d}_m, L\sqrt{q_N})\right) \leq C q_N^2\sqrt{\frac{q_N}{N}}$$

Then, Bernstein's inequality implies

$$
\begin{aligned}
&\sum_{m=1}^{M_N} P_s\left(\left|\sum_{i=1}^{N} D_i(\boldsymbol{d}_m, L\sqrt{q_N/N})\right| > q_N\varepsilon/2\right)\\
&\leq 2\sum_{m=1}^{M_N} exp\left(\frac{-q_N^2\varepsilon^2/4}{Cq_N^2\sqrt{\frac{q_N}{N}} + C\varepsilon q_N N^{-1/2}}\right)\\
&\leq 2\sum_{m=1}^{M_N} exp\left(-C\sqrt{\frac{N}{q_N}}\right) = 2M_N exp\left(-C\sqrt{\frac{N}{q_N}}\right)\\
&\leq Cexp\left(C\left(q_N+1\right)\log N - C\sqrt{\frac{N}{q_N}}\right) \to 0
\end{aligned}
$$

■

Lemma C.1.7 Assume Assumption 1–5, 7 and 8. Then, $\forall\eta > 0$, there exists an $L > 0$ such that

$$
P\left(\inf_{\|\boldsymbol{\delta}_{ab}\|=L} q_N^{-1}\sum_{i=1}^{N}(Q_i\left(\sqrt{q_N}\right) - Q_i\left(0\right)) > 0\right) \geq 1-\eta
$$

Proof. Consider

$$
\begin{aligned}
q_N^{-1}\sum_{i=1}^{N}(Q_i\left(\sqrt{q_N}\right) - Q_i\left(0\right)) &= q_N^{-1}\sum_{i=1}^{N} D_i(\boldsymbol{\delta}_{ab}, \sqrt{q_N}) + q_N^{-1}\sum_{i=1}^{N} E_w\left[Q_i\left(\sqrt{q_N}\right) - Q_i\left(0\right)\right]\\
&\quad - q_N^{-1/2}\sum_{i=1}^{N}\left(\ddot{\mathbf{w}}_{it}^{a}\boldsymbol{\delta}_a + \ddot{w}_{it}^{b}\delta_b\right)\psi_\tau\left(\varepsilon_{it}\right)\\
&= G_{N1} + G_{N2} + G_{N3}
\end{aligned}
$$

By Lemma C.1.6, we have $\sup_{\|\boldsymbol{\delta}_{ab}\|\leq L} |G_{N1}| = o_p\left(1\right)$. Also, note that $E\left[G_{N3}\right] = 0$ and that

$$
E\left[G_{N3}^2\right] \leq Cq_N^{-1}E\left[\boldsymbol{\delta}_a'\ddot{\mathbf{W}}_a'\ddot{\mathbf{W}}_a\boldsymbol{\delta}_a + \delta_b^2\ddot{W}_b'\ddot{W}_b\right] = O\left(q_N^{-1}\left\|\boldsymbol{\delta}_{ab}\right\|^2\right)
$$

Thus, $G_{N3} = O_p\left(q_N^{-1/2}\left\|\boldsymbol{\delta}_{ab}\right\|\right)$. By applying Knight's identity,

$$\begin{aligned}
G_{N2} &= \frac{1}{q_N}\sum_{i=1}^{N}\sum_{t=1}^{T} E_w\left[\int_{-r_i}^{\sqrt{q_N}\left(\ddot{\mathbf{w}}_{it}^{a}\boldsymbol{\delta}_a+\ddot{w}_{it}^{b}\delta_b\right)-r_i}\left[I\left(\varepsilon_{it}<s\right)-I\left(\varepsilon_{it}<0\right)\right]ds\right]\\
&= \frac{1}{q_N}\sum_{i=1}^{N}\sum_{t=1}^{T}\int_{-r_i}^{\sqrt{q_N}\left(\ddot{\mathbf{w}}_{it}^{a}\boldsymbol{\delta}_a+\ddot{w}_{it}^{b}\delta_b\right)-r_i} f_{it}\left(0\right)s\,ds\left(1+o\left(1\right)\right)\\
&= \frac{1}{q_N}\sum_{i=1}^{N}\sum_{t=1}^{T} f_{it}\left(0\right)\left[\frac{1}{2}q_N\left(\ddot{\mathbf{w}}_{it}^{a}\boldsymbol{\delta}_a+\ddot{w}_{it}^{b}\delta_b\right)^2 - r_i\sqrt{q_N}\left(\ddot{\mathbf{w}}_{it}^{a}\boldsymbol{\delta}_a+\ddot{w}_{it}^{b}\delta_b\right)\right]\left(1+o\left(1\right)\right)\\
&= C\boldsymbol{\delta}_a'\left(\sum_{i=1}^{N}\sum_{t=1}^{T} f_{it}\left(0\right)\ddot{\mathbf{w}}_{it}^{a\prime}\ddot{\mathbf{w}}_{it}^{a}\right)\boldsymbol{\delta}_a\left(1+o\left(1\right)\right)\\
&+ C\delta_b^2\sum_{i=1}^{N}\sum_{t=1}^{T} f_{it}\left(0\right)\left(1+o\left(1\right)\right)\\
&- q_N^{-1/2}\sum_{i=1}^{N}\sum_{t=1}^{T} f_{it}\left(0\right)r_i\left(\ddot{\mathbf{w}}_{it}^{a}\boldsymbol{\delta}_a+\ddot{w}_{it}^{b}\delta_b\right)\\
&= C\boldsymbol{\delta}_a'\ddot{\mathbf{W}}_a'B_N\ddot{\mathbf{W}}_a\boldsymbol{\delta}_a\left(1+o\left(1\right)\right)+C\delta_b^2\left(1+o\left(1\right)\right)\\
&- q_N^{-1/2}\sum_{t=1}^{T}\sum_{i=1}^{N} f_{it}\left(0\right)r_i\left(\ddot{\mathbf{w}}_{it}^{a}\boldsymbol{\delta}_a+\ddot{w}_{it}^{b}\delta_b\right)
\end{aligned}$$

Note that there exists a constant M such that $C\boldsymbol{\delta}_a'\ddot{\mathbf{W}}_a'B_N\ddot{\mathbf{W}}_a\boldsymbol{\delta}_a\left(1+o\left(1\right)\right)+C\delta_b^2\left(1+o\left(1\right)\right)\geq M\left\|\boldsymbol{\delta}_{ab}\right\|^2$. Let $R_N = \left(r_1, r_1, \cdots, r_N\right)\in\mathbb{R}^{TN}$. Then we have $\left\|R_N\right\| = O\left(\sqrt{q_N}\right)$. By Cauchy-Schwarz inequality,

$$\begin{aligned}
q_N^{-1/2}\sum_{i=1}^{N}\sum_{t=1}^{T} f_{it}\left(0\right)r_i\ddot{\mathbf{w}}_{it}^{a}\boldsymbol{\delta}_a &= q_N^{-1/2}\boldsymbol{\delta}_a'\ddot{\mathbf{W}}_a'B_NR_N\\
&\leq q_N^{-1/2}\left\|\boldsymbol{\delta}_a'\ddot{\mathbf{W}}_a'\right\|\left\|B_NR_N\right\|\\
&= O_p\left(q_N^{-1/2}q_N^{1/2}\right)\left\|\boldsymbol{\delta}_a\right\| = O_p\left(\left\|\boldsymbol{\delta}_{ab}\right\|\right)
\end{aligned}$$

Similarly,

$$\begin{aligned}
q_N^{-1/2}\sum_{i=1}^{N}\sum_{t=1}^{T} f_{it}\left(0\right)r_i\ddot{w}_{it}^{b}\delta_b &= q_N^{-1/2}\delta_b\ddot{W}_b'B_NR_N\\
&\leq q_N^{-1/2}\left\|\delta_b\ddot{W}_b'B_N^{1/2}\right\|\left\|B_N^{1/2}R_N\right\| = O_p\left(\left\|\boldsymbol{\delta}_{ab}\right\|\right)
\end{aligned}$$

Therefore, for L sufficiently large, $q_N^{-1}\sum_{i=1}^{N}(Q_i\left(\sqrt{q_N}\right) - Q_i\,(0)) > 0$ has asymptotically a lower bound cL^2. ■

Lemma C.1.8 Assume Assumption 1–5, 7 and 8. Then, for any given positive constant M_1 and M_2,

$$\sup_{\left\|\boldsymbol{\delta}_a - \tilde{\boldsymbol{\delta}}_a\right\| \le M_1, \|\delta_b\| \le M_2\sqrt{q_N}} \left|\sum_{i=1}^{N} A_i\,(\boldsymbol{\delta}_a, \tilde{\boldsymbol{\delta}}_a, \delta_b)\right| = o_p\,(1)$$

where $A_i\,(\boldsymbol{\delta}_a, \tilde{\boldsymbol{\delta}}_a, \delta_b) = \tilde{Q}_i\,(\boldsymbol{\delta}_a, \tilde{\boldsymbol{\delta}}_a, \delta_b) - E\left[\tilde{Q}_i\,(\boldsymbol{\delta}_a, \tilde{\boldsymbol{\delta}}_a, \delta_b)|\mathbf{x}_i, \mathbf{z}_i\right] + \sum_{t=1}^{T} \ddot{\mathbf{w}}_{it}^{a}\left(\boldsymbol{\delta}_a - \tilde{\boldsymbol{\delta}}_a\right)\psi_\tau\,(\varepsilon_{it})\,.$

Proof. By Assumption 4, $\max_{i,t}\|\check{\mathbf{w}}_{it}\| \le \alpha_1\sqrt{\frac{q_N}{N}}$ for some positive constant α_1. By Assumption 8 (i), $\max_{i}\|r_i\| \le \alpha_2\sqrt{\frac{q_N}{N}}$ for some positive constant α_2. It suffices to show for $\forall\varepsilon > 0$,

$$P\left(\sup_{\left\|\boldsymbol{\delta}_a - \tilde{\boldsymbol{\delta}}_a\right\| \le M_1, \|\delta_b\| \le M_2\sqrt{q_N}} \left|\sum_{i=1}^{N} A_i\,(\boldsymbol{\delta}_a, \tilde{\boldsymbol{\delta}}_a, \delta_b)\right| > \varepsilon\right) \to 0$$

Let

$$\bar{\Delta}^a = \left\{\boldsymbol{\delta} \in \mathbb{R}^{q_N} : \left\|\boldsymbol{\delta}_a - \tilde{\boldsymbol{\delta}}_a\right\| \le M_1\right\}$$
$$\bar{\Delta}^b = \left\{\delta_b \in \mathbb{R} : \|\delta_b\| \le M_2\sqrt{q_N}\right\}$$

We can partition $\bar{\Delta}^a$ ($\bar{\Delta}^b$) into disjoint sets $\bar{\Delta}^a_1, \cdots \bar{\Delta}^a_{D^a_N}$ ($\bar{\Delta}^b_1, \cdots \bar{\Delta}^b_{D^b_N}$) such that the diameter of each set does not exceed $m_0^* = \frac{\varepsilon}{10T\alpha_1\sqrt{Nq_N}}$ and the cardinality of partition satisfies $D^a_N \le \left(\frac{C\sqrt{Nq_N}}{2\varepsilon}\right)^{q_N}$ and $D^b_N \le \left(\frac{C\sqrt{N}q_N}{2\varepsilon}\right)^{q_N}$ (For example, by similar argument used in Lemma 5.2 of Vershynin, 2011). Pick arbitrary $\boldsymbol{\delta}^k_a \in \bar{\Delta}^a_k$ for $1 \le k \le D^a_N$ and $\boldsymbol{\delta}^l_b \in \bar{\Delta}^b_l$ for $1 \le l \le D^b_N$. Then

$$P\left(\sup_{\left\|\boldsymbol{\delta}_a - \tilde{\boldsymbol{\delta}}_a\right\| \le M_1, \|\delta_b\| \le M_2\sqrt{q_N}} \left|\sum_{i=1}^{N} A_i\,(\boldsymbol{\delta}_a, \tilde{\boldsymbol{\delta}}_a, \delta_b)\right| > \varepsilon\right)$$
$$\le \sum_{k=1}^{D^a_N}\sum_{l=1}^{D^b_N} P\left(\left|\sum_{i=1}^{N} A_i\left(\boldsymbol{\delta}^k_a, \tilde{\boldsymbol{\delta}}, \delta^l_b\right)\right| + \sup_{\boldsymbol{\delta}_a \in \bar{\Delta}^a_k, \delta^l_b \in \bar{\Delta}^b_l}\left|\sum_{i=1}^{N}\left[A_i\,(\boldsymbol{\delta}_a, \tilde{\boldsymbol{\delta}}_a, \delta_b) - A_i\left(\boldsymbol{\delta}^k_a, \tilde{\boldsymbol{\delta}}, \delta^l_b\right)\right]\right| > \varepsilon\right)$$

Since $u1\left[u<0\right]=\frac{1}{2}u-\frac{1}{2}\left|u\right|$, we have

$$
\begin{aligned}
&\tilde{Q}_i\left(\boldsymbol{\delta}_a,\tilde{\boldsymbol{\delta}}_a,\delta_b\right)-\tilde{Q}_i\left(\boldsymbol{\delta}_a^k,\tilde{\boldsymbol{\delta}}_a,\delta_b^l\right)\\
&=\sum_{t=1}^{T}\left[\rho_\tau\left(e_{it}-\ddot{\mathbf{w}}_{it}^a\boldsymbol{\delta}_a-\ddot{w}_b\delta_b\right)-\rho_\tau\left(e_{it}-\ddot{\mathbf{w}}_{it}^a\tilde{\boldsymbol{\delta}}_a-\ddot{w}_b\delta_b\right)\right]\\
&-\sum_{t=1}^{T}\left[\rho_\tau\left(e_{it}-\ddot{\mathbf{w}}_{it}^a\boldsymbol{\delta}_a^k-\ddot{w}_b\delta_b^l\right)-\rho_\tau\left(e_{it}-\ddot{\mathbf{w}}_{it}^a\tilde{\boldsymbol{\delta}}_a-\ddot{w}_b\delta_b^l\right)\right]\\
&=\sum_{t=1}^{T}[\left(\tau-\frac{1}{2}\right)\ddot{\mathbf{w}}_{it}^a\left(\boldsymbol{\delta}_a^k-\boldsymbol{\delta}_a\right)+\frac{1}{2}\left(\left|\varepsilon_{it}+r_i-\ddot{\mathbf{w}}_{it}^a\boldsymbol{\delta}_a-\ddot{w}_b\delta_b\right|-\left|\varepsilon_{it}+r_i-\ddot{\mathbf{w}}_{it}^a\boldsymbol{\delta}_a^k-\ddot{w}_b\delta_b^l\right|\right)\\
&-\frac{1}{2}\left(\left|\varepsilon_{it}+r_i-\ddot{\mathbf{w}}_{it}^a\tilde{\boldsymbol{\delta}}_a-\ddot{w}_b\delta_b\right|-\left|\varepsilon_{it}+r_i-\ddot{\mathbf{w}}_{it}^a\tilde{\boldsymbol{\delta}}_a-\ddot{w}_b\delta_b^l\right|\right)]\\
&\leq 2T\max_{i,t}\left\|\ddot{\mathbf{w}}_{it}^a\right\|\sup_{\boldsymbol{\delta}_a^k\in\bar{\Delta}_k^a,\delta_b^l\in\bar{\Delta}_l^b}\left[\left\|\boldsymbol{\delta}_a-\boldsymbol{\delta}_a^k\right\|+\left\|\delta_b-\delta_b^l\right\|\right]
\end{aligned}
$$

Thus

$$
\begin{aligned}
&\sup_{\boldsymbol{\delta}_a\in\bar{\Delta}_k^a,\delta_b^l\in\bar{\Delta}_l^b}\left|\sum_{i=1}^{N}\left[A_i\left(\boldsymbol{\delta}_a,\tilde{\boldsymbol{\delta}}_a,\delta_b\right)-A_i\left(\boldsymbol{\delta}_a^k,\tilde{\boldsymbol{\delta}},\delta_b^l\right)\right]\right|\\
&=\sup_{\boldsymbol{\delta}\in\bar{\Delta}_d}\left|\sum_{i=1}^{N}\{\tilde{Q}_i\left(\boldsymbol{\delta}_a,\tilde{\boldsymbol{\delta}}_a,\delta_b\right)-E_w\left[\tilde{Q}_i\left(\boldsymbol{\delta}_a,\tilde{\boldsymbol{\delta}}_a,\delta_b\right)\right]+\sum_{t=1}^{T}\ddot{\mathbf{w}}_{it}^a\left(\boldsymbol{\delta}_a-\tilde{\boldsymbol{\delta}}_a\right)\psi_\tau\left(\varepsilon_{it}\right)\right.\\
&\left.-\tilde{Q}_i(\boldsymbol{\delta}_d^k,\tilde{\boldsymbol{\delta}}_a,\delta_b^l)+E_w\left[\tilde{Q}_i(\boldsymbol{\delta}_a^k,\tilde{\boldsymbol{\delta}}_a,\delta_b^l)\right]-\sum_{t=1}^{T}\ddot{\mathbf{w}}_{it}^a\left(\boldsymbol{\delta}_a^k-\tilde{\boldsymbol{\delta}}_a\right)\psi_\tau\left(\varepsilon_{it}\right)\right|
\end{aligned}
$$

$$
\begin{aligned}
&=\sup_{\boldsymbol{\delta}\in\bar{\Delta}_d}\left|\sum_{i=1}^{N}\{\tilde{Q}_i(\boldsymbol{\delta}_a,\tilde{\boldsymbol{\delta}}_a,\delta_b)-\tilde{Q}_i(\boldsymbol{\delta}_d^k,\tilde{\boldsymbol{\delta}}_a,\delta_b^l)-E_w\left[\tilde{Q}_i(\boldsymbol{\delta}_a,\tilde{\boldsymbol{\delta}}_a,\delta_b)-\tilde{Q}_i(\boldsymbol{\delta}_a^k,\tilde{\boldsymbol{\delta}}_a,\delta_b^l)\right]\right.\\
&\left.+\sum_{t=1}^{T}\check{\mathbf{w}}_{it}\left(\boldsymbol{\delta}_a-\boldsymbol{\delta}_a^k\right)\psi_\tau\left(\varepsilon_{it}\right)\}\right|\\
&\leq 5NT\max_{i,t}\left\|\ddot{\mathbf{w}}_{it}^a\right\|\sup_{\boldsymbol{\delta}_a^k\in\bar{\Delta}_k^a,\boldsymbol{\delta}_b^l\in\bar{\Delta}_l^b}\left[\left\|\boldsymbol{\delta}_a-\boldsymbol{\delta}_a^k\right\|+\left\|\delta_b-\delta_b^l\right\|\right]\\
&\leq 5NT\alpha_1\sqrt{\frac{q_N}{N}}m_0=\frac{\varepsilon}{2}
\end{aligned}
$$

Therefore, now it suffices to show $\sum_{k=1}^{D_N^a}\sum_{l=1}^{D_N^b} P_w\left(\left|\sum_{i=1}^N A_i\left(\boldsymbol{\delta}_a^k,\tilde{\boldsymbol{\delta}},\delta_b^l\right)\right| > \frac{\varepsilon}{2}\right)$ has a vanishing upper bound that does not depend on $(\mathbf{x}_i,\mathbf{z}_i)$. Berstein's inequality is used. To evaluate maximum, using $\rho\left(u\right)=\left(\tau-\frac{1}{2}\right)u+\frac{1}{2}\left|u\right|$, we can write

$$
\begin{aligned}
&\max_{i,k,l} A_i\left(\boldsymbol{\delta}_a^k,\tilde{\boldsymbol{\delta}},\delta_b^l\right)\\
&=\max_{i,k,l}[\tilde{Q}_i(\boldsymbol{\delta}_a^k,\tilde{\boldsymbol{\delta}}_a,\delta_b^l)-E\left[\tilde{Q}_i(\boldsymbol{\delta}_a^k,\tilde{\boldsymbol{\delta}}_a,\delta_b^l)|\mathbf{x}_i,\mathbf{z}_i\right]+\sum_{t=1}^T\ddot{\mathbf{w}}_{it}^a\left(\boldsymbol{\delta}_a^k-\tilde{\boldsymbol{\delta}}_a\right)\psi_\tau\left(\varepsilon_{it}\right)]\\
&=\max_{i,k,l}\sum_{t=1}^T\left[\rho_\tau\left(e_{it}-\ddot{\mathbf{w}}_{it}^a\boldsymbol{\delta}_a^k-\ddot{w}_b\delta_b^l\right)-\rho_\tau\left(e_{it}-\ddot{\mathbf{w}}_{it}^a\tilde{\boldsymbol{\delta}}_a-\ddot{w}_b\delta_b^l\right)\right]\\
&-E_w[\sum_{t=1}^T\left[\rho_\tau\left(e_{it}-\ddot{\mathbf{w}}_{it}^a\boldsymbol{\delta}_a^k-\ddot{w}_b\delta_b^l\right)-\rho_\tau\left(e_{it}-\ddot{\mathbf{w}}_{it}^a\tilde{\boldsymbol{\delta}}_a-\ddot{w}_b\delta_b\right)\right]]+\sum_{t=1}^T\ddot{\mathbf{w}}_{it}^a\left(\boldsymbol{\delta}_a-\tilde{\boldsymbol{\delta}}_a\right)\psi_\tau\left(\varepsilon_{it}\right)\\
&=\max_{i,k,l}\sum_{t=1}^T\frac{1}{2}[\left|e_{it}-\ddot{\mathbf{w}}_{it}^a\boldsymbol{\delta}_a^k-\ddot{w}_b\delta_b^l\right|-\left|e_{it}-\ddot{\mathbf{w}}_{it}^a\tilde{\boldsymbol{\delta}}_a-\ddot{w}_b\delta_b^l\right|+\left(\tau-\frac{1}{2}\right)\ddot{\mathbf{w}}_{it}^a\left(\tilde{\boldsymbol{\delta}}_a-\boldsymbol{\delta}_a^k\right)]\\
&-E_w\sum_{t=1}^T\frac{1}{2}[\left|e_{it}-\ddot{\mathbf{w}}_{it}^a\boldsymbol{\delta}_a^k-\ddot{w}_b\delta_b^l\right|-\left|e_{it}-\ddot{\mathbf{w}}_{it}^a\tilde{\boldsymbol{\delta}}_a-\ddot{w}_b\delta_b^l\right|+\left(\tau-\frac{1}{2}\right)\ddot{\mathbf{w}}_{it}^a\left(\tilde{\boldsymbol{\delta}}_a-\boldsymbol{\delta}_a^k\right)]\\
&+\sum_{t=1}^T\ddot{\mathbf{w}}_{it}^a\left(\boldsymbol{\delta}_a-\tilde{\boldsymbol{\delta}}_a\right)\psi_\tau\left(\varepsilon_{it}\right)\\
&\leq 3T\max_{i,t}\left\|\ddot{\mathbf{w}}_{it}^a\right\|\max_k\left\|\tilde{\boldsymbol{\delta}}_a-\boldsymbol{\delta}_a^k\right\|\leq C\sqrt{\frac{q_N}{N}}
\end{aligned}
$$

To evaluate variance, applying Knight's identity, we have

$$\begin{aligned}
&\tilde{Q}_i(\boldsymbol{\delta}_a^k, \tilde{\boldsymbol{\delta}}_a, \delta_b^l) + \sum_{t=1}^{T} \ddot{\mathbf{w}}_{it}^a \left(\boldsymbol{\delta}_a^k - \tilde{\boldsymbol{\delta}}_a\right) \psi_\tau\left(\varepsilon_{it}\right) \\
&= \sum_{t=1}^{T} \left[\rho_\tau\left(e_{it} - \ddot{\mathbf{w}}_{it}^a \boldsymbol{\delta}_a^k - \ddot{w}_b \delta_b^l\right) - \rho_\tau\left(e_{it} - \ddot{\mathbf{w}}_{it}^a \tilde{\boldsymbol{\delta}}_a - \ddot{w}_b \delta_b^l\right)\right] + \sum_{t=1}^{T} \ddot{\mathbf{w}}_{it}^a \left(\boldsymbol{\delta}_a^k - \tilde{\boldsymbol{\delta}}_a\right) \psi_\tau\left(\varepsilon_{it}\right) \\
&= -\sum_{t=1}^{T}(\ddot{\mathbf{w}}_{it}^a \boldsymbol{\delta}_a^k + \ddot{w}_b \delta_b^l - r_i)\psi_\tau\left(\varepsilon_{it}\right) + \sum_{t=1}^{T} \int_0^{\ddot{\mathbf{w}}_{it}^a \boldsymbol{\delta}_a^k + \ddot{w}_b \delta_b^l - r_i} \left[I\left(\varepsilon_{it} < t\right) - I(\varepsilon_{it} < 0)\right] dt \\
&+ \sum_{t=1}^{T}(\ddot{\mathbf{w}}_{it}^a \tilde{\boldsymbol{\delta}}_a + \ddot{w}_b \delta_b^l - r_i)\psi_\tau\left(\varepsilon_{it}\right) - \sum_{t=1}^{T} \int_0^{\ddot{\mathbf{w}}_{it}^a \tilde{\boldsymbol{\delta}}_a + \ddot{w}_b \delta_b^l - r_i} \left[I\left(\varepsilon_{it} < t\right) - I(\varepsilon_{it} < 0)\right] dt \\
&+ \sum_{t=1}^{T} \ddot{\mathbf{w}}_{it}^a \left(\boldsymbol{\delta}_a^k - \tilde{\boldsymbol{\delta}}_a\right) \psi_\tau\left(\varepsilon_{it}\right) \\
&= \sum_{t=1}^{T} \int_{\ddot{\mathbf{w}}_{it}^a \tilde{\boldsymbol{\delta}}_a + \ddot{w}_b \delta_b^l - r_i}^{\ddot{\mathbf{w}}_{it}^a \boldsymbol{\delta}_a^k + \ddot{w}_b \delta_b^l - r_i} \left[I\left(\varepsilon_{it} < t\right) - I(\varepsilon_{it} < 0)\right] dt
\end{aligned}$$

Thus,

$$\begin{aligned}
&A_i\left(\boldsymbol{\delta}_a^k, \tilde{\boldsymbol{\delta}}_a, \delta_b^l\right) \\
&= \tilde{Q}_i\left(\boldsymbol{\delta}_a^k, \tilde{\boldsymbol{\delta}}_a, \delta_b^l\right) - E\left[\tilde{Q}_i\left(\boldsymbol{\delta}_a^k, \tilde{\boldsymbol{\delta}}_a, \delta_b^l\right) | \mathbf{x}_i, \mathbf{z}_i\right] + \sum_{t=1}^{T} \ddot{\mathbf{w}}_{it}^a \left(\boldsymbol{\delta}_a^k - \tilde{\boldsymbol{\delta}}_a\right) \psi_\tau\left(\varepsilon_{it}\right) \\
&= \sum_{t=1}^{T} \int_{\ddot{\mathbf{w}}_{it}^a \tilde{\boldsymbol{\delta}}_a + \ddot{w}_b \delta_b^l - r_i}^{\ddot{\mathbf{w}}_{it}^a \boldsymbol{\delta}_a^k + \ddot{w}_b \delta_b^l - r_i} \left[I\left(\varepsilon_{it} < t\right) - I(\varepsilon_{it} < 0)\right] dt - \sum_{t=1}^{T} \ddot{\mathbf{w}}_{it}^a \left(\boldsymbol{\delta}_a^k - \tilde{\boldsymbol{\delta}}_a\right) \psi_\tau\left(\varepsilon_{it}\right) \\
&- E_w\left[\sum_{t=1}^{T} \int_{\ddot{\mathbf{w}}_{it}^a \tilde{\boldsymbol{\delta}}_a + \ddot{w}_b \delta_b^l - r_i}^{\ddot{\mathbf{w}}_{it}^a \boldsymbol{\delta}_a^k + \ddot{w}_b \delta_b^l - r_i} \left[I\left(\varepsilon_{it} < t\right) - I(\varepsilon_{it} < 0)\right] dt - \sum_{t=1}^{T} \ddot{\mathbf{w}}_{it}^a \left(\boldsymbol{\delta}_a^k - \tilde{\boldsymbol{\delta}}_a\right) \psi_\tau\left(\varepsilon_{it}\right)\right] \\
&+ \sum_{t=1}^{T} \ddot{\mathbf{w}}_{it}^a \left(\boldsymbol{\delta}_a^k - \tilde{\boldsymbol{\delta}}_a\right) \psi_\tau\left(\varepsilon_{it}\right)
\end{aligned}$$

$$\begin{aligned}
&= \sum_{t=1}^{T} \int_{\ddot{\mathbf{w}}_{it}^a \tilde{\boldsymbol{\delta}}_a + \ddot{w}_b \delta_b^l - r_i}^{\ddot{\mathbf{w}}_{it}^a \boldsymbol{\delta}_a^k + \ddot{w}_b \delta_b^l - r_i} \left[I\left(\varepsilon_{it} < t\right) - I(\varepsilon_{it} < 0)\right] dt \\
&- E_w\left[\sum_{t=1}^{T} \int_{\ddot{\mathbf{w}}_{it}^a \tilde{\boldsymbol{\delta}}_a + \ddot{w}_b \delta_b^l - r_i}^{\ddot{\mathbf{w}}_{it}^a \boldsymbol{\delta}_a^k + \ddot{w}_b \delta_b^l - r_i} \left[I\left(\varepsilon_{it} < t\right) - I(\varepsilon_{it} < 0)\right] dt\right] + E_w\left[\sum_{t=1}^{T} \ddot{\mathbf{w}}_{it}^a \left(\boldsymbol{\delta}_a^k - \tilde{\boldsymbol{\delta}}_a\right) \psi_\tau\left(\varepsilon_{it}\right)\right]
\end{aligned}$$

And it implies

$$
\begin{aligned}
&\sum_{i=1}^{N} Var\left(A_i\left(\boldsymbol{\delta}_a^k, \tilde{\boldsymbol{\delta}}_a, \delta_b^l\right) |\mathbf{x}_i, \mathbf{z}_i\right) \\
&\leq \sum_{i=1}^{N} E_w\left[\left(\sum_{t=1}^{T} \int_{\ddot{\mathbf{w}}_{it}^a \tilde{\boldsymbol{\delta}}_a + \ddot{w}_b \delta_b^l - r_i}^{\ddot{\mathbf{w}}_{it}^a \boldsymbol{\delta}_a^k + \ddot{w}_b \delta_b^l - r_i} [I\left(\varepsilon_{it} < t\right) - I(\varepsilon_{it} < 0)]dt\right)^2\right] \\
&\leq \sum_{i=1}^{N} T \max_{i,t} \left|\ddot{\mathbf{w}}_{it}^a\left(\boldsymbol{\delta}_a^k - \tilde{\boldsymbol{\delta}}_a\right)\right| \sum_{t=1}^{T} \int_{\ddot{\mathbf{w}}_{it}^a \tilde{\boldsymbol{\delta}}_a + \ddot{w}_b \delta_b^l - r_i}^{\ddot{\mathbf{w}}_{it}^a \boldsymbol{\delta}_a^k + \ddot{w}_b \delta_b^l - r_i} [F_{it}\left(t\right) - F_{it}\left(0\right)]dt \\
&\leq C\sqrt{\frac{q_N}{N}} \sum_{i=1}^{N} \sum_{t=1}^{T} f_{it}\left(0\right)\left(\left(\ddot{\mathbf{w}}_{it}^a \boldsymbol{\delta}_a^k + \ddot{w}_b \delta_b^l - r_i\right)^2 - \left(\ddot{\mathbf{w}}_{it}^a \tilde{\boldsymbol{\delta}}_a + \ddot{w}_b \delta_b^l - r_i\right)^2\right)\left(1 + o\left(1\right)\right) \\
&\leq C\sqrt{\frac{q_N}{N}}\left(1 + o\left(1\right)\right)
\end{aligned}
$$

by similar argument as in Lemma C.1.2. Then, by Bernstein's inequality and Assumption 5,

$$
\begin{aligned}
&\sum_{k=1}^{D_N^a} \sum_{l=1}^{D_N^b} P_w\left(\left|\sum_{i=1}^{N} A_i\left(\boldsymbol{\delta}_a^k, \tilde{\boldsymbol{\delta}}, \delta_b^l\right)\right| > \frac{\varepsilon}{2}\right) \\
&\leq \sum_{k=1}^{D_N^a} \sum_{l=1}^{D_N^b} exp\left(\frac{-\varepsilon^2/4}{C\sqrt{\frac{q_N}{N}} + \varepsilon C\sqrt{\frac{q_N}{N}}}\right) \leq \sum_{k=1}^{D_N^a} \sum_{l=1}^{D_N^b} exp\left(-C\sqrt{\frac{N}{q_N}}\right) \\
&\leq C\left(C\sqrt{Nq_N}\right)^{q_N}\left(C\sqrt{N}q_N\right)^{q_N} exp\left(-C\sqrt{\frac{N}{q_N}}\right) \leq C exp\left(C\left(q_N \log N - \sqrt{\frac{N}{q_N}}\right)\right) \to 0
\end{aligned}
$$

■

Lemma C.1.9 (Asymptotic Equivalence with Bahadur Representation) Assume Assumption 1–5, 7 and 8. Then, we have $\left\|\tilde{\boldsymbol{\delta}}_a - \hat{\boldsymbol{\delta}}_a\right\| = o_p\left(1\right).$

Proof. Note that $\sum_{i=1}^{N}[Q_i\left(\sqrt{q_N}\check{\boldsymbol{\delta}}_a, \sqrt{q_N}\check{\delta}_b\right) - Q_i\left(0, 0\right)] \leq 0$ where $\left(\sqrt{q_N}\check{\boldsymbol{\delta}}_a, \sqrt{q_N}\check{\delta}_b\right)$ coincides with oracle estimator $\hat{\boldsymbol{\delta}}_{ab}$. Then, Lemma C.1.7 implies $\left\|\hat{\boldsymbol{\delta}}_{ab}\right\| = O_p\left(\sqrt{q_N}\right)$. Now, it suffices to show that for any positive constants M_1 and M_2,

$$
P\left(\inf_{\left\|\tilde{\boldsymbol{\delta}}_a - \hat{\boldsymbol{\delta}}_a\right\| \geq M_1, \|\delta_b\| \leq M_2\sqrt{q_N}} \sum_{i=1}^{N} \tilde{Q}_i\left(\boldsymbol{\delta}_a, \tilde{\boldsymbol{\delta}}_a, \delta_b\right) > 0\right) \to 1
$$

since, we have $\sum_{i=1}^{N} \tilde{Q}_i \left(\hat{\boldsymbol{\delta}}_a, \tilde{\boldsymbol{\delta}}, \hat{\delta}_b\right) \leq 0$. Let $\mathcal{B} = \left\{\boldsymbol{\delta}_{ab} : \left\|\boldsymbol{\delta}_a - \tilde{\boldsymbol{\delta}}_a\right\| \leq M_1, \|\delta_b\| \leq M_2\sqrt{q_N}\right\}$. By Lemma C.1.8,

$$\sup_{\boldsymbol{\delta}_{ab} \in \mathcal{B}} \left| \sum_{i=1}^{N} \left[\tilde{Q}_i (\boldsymbol{\delta}_a, \tilde{\boldsymbol{\delta}}_a, \delta_b) - E\left[\tilde{Q}_i (\boldsymbol{\delta}_a, \tilde{\boldsymbol{\delta}}_a, \delta_b) | \mathbf{x}_i, \mathbf{z}_i\right] + \sum_{t=1}^{T} \ddot{\mathbf{w}}_{it}^{a} \left(\boldsymbol{\delta}_a - \tilde{\boldsymbol{\delta}}_a\right) \psi_\tau\left(\varepsilon_{it}\right) \right] \right| = o_p\left(1\right)$$

Then by Lemma C.1.5,

$$\begin{aligned} &\sup_{\boldsymbol{\delta}_{ab} \in \mathcal{B}} \left| \sum_{i=1}^{N} \left[\tilde{Q}_i (\boldsymbol{\delta}_a, \tilde{\boldsymbol{\delta}}_a, \delta_b) - \frac{1}{2}\left[\boldsymbol{\delta}_a' \ddot{\mathbf{K}}_N \boldsymbol{\delta}_a - \tilde{\boldsymbol{\delta}}_a' \ddot{\mathbf{K}}_N \tilde{\boldsymbol{\delta}}_a\right] (1 + o\left(1\right)) \right.\right. && \text{(A.7)} \\ &\left.\left. + \sum_{t=1}^{T} \ddot{\mathbf{w}}_{it}^{a} \left(\boldsymbol{\delta}_a - \tilde{\boldsymbol{\delta}}_a\right) \psi_\tau\left(\varepsilon_{it}\right) \right] \right| \\ &= o_p\left(1\right) \end{aligned}$$

And since

$$\begin{aligned} \tilde{\boldsymbol{\delta}}_a &= \left(\ddot{\mathbf{W}}_a' B_N \ddot{\mathbf{W}}_a\right)^{-1} \ddot{\mathbf{W}}_a' \boldsymbol{\Psi}_\tau\left(\boldsymbol{\varepsilon}\right) \\ &= \ddot{\mathbf{K}}_N^{-1} \ddot{\mathbf{W}}_a' \boldsymbol{\Psi}_\tau\left(\boldsymbol{\varepsilon}\right) \end{aligned}$$

we have

$$\begin{aligned} \sum_{i=1}^{N}\sum_{t=1}^{T} \ddot{\mathbf{w}}_{it}^{a} \left(\boldsymbol{\delta}_a - \tilde{\boldsymbol{\delta}}_a\right) \psi_\tau\left(\varepsilon_{it}\right) &= \left(\boldsymbol{\delta}_a - \tilde{\boldsymbol{\delta}}_a\right)' \sum_{i=1}^{N}\sum_{t=1}^{T} \ddot{\mathbf{w}}_{it}^{a\prime} \psi\left(e_{it}\right) \\ &= \left(\boldsymbol{\delta}_a - \tilde{\boldsymbol{\delta}}_a\right)' \ddot{\mathbf{W}}_a' \boldsymbol{\Psi}_\tau\left(\boldsymbol{\varepsilon}\right) && \text{(A.8)} \\ &= \left(\boldsymbol{\delta}_a - \tilde{\boldsymbol{\delta}}_a\right)' \ddot{\mathbf{K}}_N \tilde{\boldsymbol{\delta}}_a && \text{(A.9)} \end{aligned}$$

Combining (A.7) and (A.9), we have

$$\sup_{\boldsymbol{\delta}_{ab} \in \mathcal{B}} \left| \sum_{i=1}^{N} \left[\tilde{Q}_i (\boldsymbol{\delta}_a, \tilde{\boldsymbol{\delta}}_a, \delta_b) - \frac{1}{2}\left[\boldsymbol{\delta}_a' \ddot{\mathbf{K}}_N \boldsymbol{\delta}_a - \tilde{\boldsymbol{\delta}}_a' \ddot{\mathbf{K}}_N \tilde{\boldsymbol{\delta}}_a\right] + \left(\boldsymbol{\delta}_a - \tilde{\boldsymbol{\delta}}_a\right)' \ddot{\mathbf{K}}_N \tilde{\boldsymbol{\delta}}_a \right] \right| = o_p\left(1\right)$$

which implies

$$\sup_{\boldsymbol{\delta}_{ab} \in \mathcal{B}} \left| \sum_{i=1}^{N} \left[\tilde{Q}_i (\boldsymbol{\delta}_a, \tilde{\boldsymbol{\delta}}_a, \delta_b) - \frac{1}{2}\left(\boldsymbol{\delta}_a - \tilde{\boldsymbol{\delta}}_a\right)' \ddot{\mathbf{K}}_N \left(\boldsymbol{\delta}_a - \tilde{\boldsymbol{\delta}}_a\right) \right] \right| = o_p\left(1\right)$$

By assumption 4, for any $\left\|\left(\boldsymbol{\delta}_a - \tilde{\boldsymbol{\delta}}_a\right)\right\| > M$,

$$\frac{1}{2}\left(\boldsymbol{\delta}_a - \tilde{\boldsymbol{\delta}}_a\right)' \ddot{\mathbf{K}}_N \left(\boldsymbol{\delta}_a - \tilde{\boldsymbol{\delta}}_a\right) > CM$$

for some positive C. ■

Since the partition $(\tilde{w}_{it}^b, \tilde{\mathbf{w}}_{it}^a)$ is arbitrary, Lemma C.1.9 implies $\left\|\tilde{\boldsymbol{\delta}} - \hat{\boldsymbol{\delta}}\right\| = o_p(1)$.

1) convergence rate of $\hat{\boldsymbol{\beta}}$: Consider the Bahadur representation $\tilde{\boldsymbol{\delta}}$ in Lemma C.1.1. Let $\tilde{\boldsymbol{\delta}}_1$ be the subvector of the first K_4 components in $\tilde{\boldsymbol{\delta}}$. Then, by paritioned matrix formula, we can write

$$\begin{aligned}
\tilde{\boldsymbol{\delta}}_1 &= \sqrt{N}\left(\mathbf{W}' B_N \mathbf{W} - \mathbf{W}' B_N \boldsymbol{\Pi}_A \left(\boldsymbol{\Pi}_A' B_N \boldsymbol{\Pi}_A\right)^{-1} \boldsymbol{\Pi}_A' B_N \mathbf{W}\right)^{-1} \\
&\quad \times \left[\mathbf{W}' \boldsymbol{\Psi}_\tau(\varepsilon) - \mathbf{W}' B_N \boldsymbol{\Pi}_A \left(\boldsymbol{\Pi}_A' B_N \boldsymbol{\Pi}_A\right)^{-1} \boldsymbol{\Pi}_A' \boldsymbol{\Psi}_\tau(\varepsilon)\right] \\
&= \underbrace{\left(\frac{1}{N}\mathbf{W}' B_N \mathbf{W} - \frac{1}{N}\mathbf{W}' B_N \boldsymbol{\Pi}_A \left(\frac{1}{N}\boldsymbol{\Pi}_A' B_N \boldsymbol{\Pi}_A\right)^{-1} \frac{1}{N}\boldsymbol{\Pi}_A' B_N \mathbf{W}\right)^{-1}}_{(a)} \\
&\quad \times \left[\frac{1}{\sqrt{N}}\mathbf{W}' \boldsymbol{\Psi}_\tau(\varepsilon) - \frac{1}{N}\mathbf{W}' B_N \boldsymbol{\Pi}_A \left(\frac{1}{N}\boldsymbol{\Pi}_A' B_N \boldsymbol{\Pi}_A\right)^{-1} \frac{1}{\sqrt{N}}\boldsymbol{\Pi}_A' \boldsymbol{\Psi}_\tau(\varepsilon)\right]
\end{aligned}$$

Since part (a) is upper-left submatrix of $\left(\frac{1}{N}\tilde{\mathbf{W}}' B_N \tilde{\mathbf{W}}\right)^{-1}$, its maximum eigenvalue is bouned above by an argument used in Lemma C.1.1. Then, by Lemma C.1.10,

$$\begin{aligned}
\left\|\tilde{\boldsymbol{\delta}}_1\right\| &\le \left\|\left(\frac{1}{N}\mathbf{W}' B_N \mathbf{W} - \frac{1}{N}\mathbf{W}' B_N \boldsymbol{\Pi}_A \left(\frac{1}{N}\boldsymbol{\Pi}_A' B_N \boldsymbol{\Pi}_A\right)^{-1} \frac{1}{N}\boldsymbol{\Pi}_A' B_N \mathbf{W}\right)^{-\frac{1}{2}}\right\|^2 \\
&\quad \times \left\|\frac{1}{\sqrt{N}}\mathbf{W}' \boldsymbol{\psi}_\tau(\varepsilon) - \frac{1}{N}\mathbf{W}' B_N \boldsymbol{\Pi}_A \left(\frac{1}{N}\boldsymbol{\Pi}_A' B_N \boldsymbol{\Pi}_A\right)^{-1} \frac{1}{\sqrt{N}}\boldsymbol{\Pi}_A' \boldsymbol{\Psi}_\tau(\varepsilon)\right\| \\
&\le C\left\|\frac{1}{\sqrt{N}}\underbrace{\mathbf{W}'\left[\boldsymbol{I}_{NT} - B_N \boldsymbol{\Pi}_A \left(\boldsymbol{\Pi}_A' B_N \boldsymbol{\Pi}_A\right)^{-1} \boldsymbol{\Pi}_A'\right]}_{=\mathbf{W}^{*\prime}}\boldsymbol{\Psi}_\tau(\varepsilon)\right\| \\
&= C\left\|\left(\frac{1}{\sqrt{N}}\boldsymbol{\Delta} + o_p(1)\right)\boldsymbol{\Psi}_\tau(\varepsilon)\right\| = O_p(1)
\end{aligned}$$

Thus $\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_o\right\| = O_p\left(N^{-1/2}\right)$. ■

2) convergence rate of $\hat{g}$

By Lemma C.1.1 and C.1.9,

$$\|\hat{\boldsymbol{\gamma}}_A - \boldsymbol{\gamma}_{oA}\| \leq \left\| \begin{pmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_o \\ \hat{\boldsymbol{\gamma}}_A - \boldsymbol{\gamma}_{oA} \end{pmatrix} \right\| = \left\| \frac{1}{\sqrt{N}} \hat{\boldsymbol{\delta}} \right\| = O_p\left(\sqrt{\frac{q_N}{N}}\right)$$

And, by Assumption 8, we have

$$\begin{aligned} \frac{1}{N}\sum_{i=1}^{N} [\hat{g}(\mathbf{x}_i, \mathbf{z}_i) - g_o(\mathbf{x}_i, \mathbf{z}_i)]^2 &= \frac{1}{N}\sum_{i=1}^{N} (\boldsymbol{\pi}_{iA}\hat{\boldsymbol{\gamma}}_A - g_o(\mathbf{x}_i, \mathbf{z}_i))^2 \\ &\leq \frac{1}{N}\sum_{i=1}^{N} (\boldsymbol{\pi}_{iA}(\hat{\boldsymbol{\gamma}}_A - \boldsymbol{\gamma}_{oA}))^2 + \frac{1}{N}\sum_{i=1}^{N} (\boldsymbol{\pi}_{iA}\boldsymbol{\gamma}_{oA} - g_o(\mathbf{x}_i, \mathbf{z}_i))^2 \\ &\leq \frac{1}{N}\sum (\boldsymbol{\pi}_{iA}(\hat{\boldsymbol{\gamma}}_A - \boldsymbol{\gamma}_{oA}))^2 + O\left(\frac{q_N}{N}\right) \end{aligned}$$

Now, it suffices to show $\sum(\boldsymbol{\pi}_{iA}(\hat{\boldsymbol{\gamma}}_A - \boldsymbol{\gamma}_{oA}))^2 = O_p(q_N)$. First note that

$$\begin{aligned} \sum(\boldsymbol{\pi}_{iA}(\hat{\boldsymbol{\gamma}}_A - \boldsymbol{\gamma}_{oA}))^2 &= (\boldsymbol{\Pi}_A(\hat{\boldsymbol{\gamma}}_A - \boldsymbol{\gamma}_{oA}))'(\boldsymbol{\Pi}_A(\hat{\boldsymbol{\gamma}}_A - \boldsymbol{\gamma}_{oA})) \\ &= N^{\frac{1}{2}}(\hat{\boldsymbol{\gamma}}_A - \boldsymbol{\gamma}_{oA})'\left(N^{-1}\boldsymbol{\Pi}'_A\boldsymbol{\Pi}_A\right)^{\frac{1}{2}}\left(N^{-1}\boldsymbol{\Pi}'_A\boldsymbol{\Pi}_A\right)^{\frac{1}{2}} N^{\frac{1}{2}}(\hat{\boldsymbol{\gamma}}_A - \boldsymbol{\gamma}_{oA}) \end{aligned}$$

Then, since we have

$$\begin{aligned} \left\|\left(N^{-1}\boldsymbol{\Pi}'_A\boldsymbol{\Pi}_A\right)^{\frac{1}{2}} N^{\frac{1}{2}}(\hat{\boldsymbol{\gamma}}_A - \boldsymbol{\gamma}_{oA})\right\| &\leq \left\|\left(N^{-1}\boldsymbol{\Pi}'_A\boldsymbol{\Pi}_A\right)^{\frac{1}{2}}\right\| \left\|N^{\frac{1}{2}}(\hat{\boldsymbol{\gamma}}_A - \boldsymbol{\gamma}_{oA})\right\| \\ &\leq \lambda_{\max}\left(N^{-1}\boldsymbol{\Pi}'_A\boldsymbol{\Pi}_A\right)\left\|N^{\frac{1}{2}}(\hat{\boldsymbol{\gamma}}_A - \boldsymbol{\gamma}_{oA})\right\| \\ &= O_p\left(\sqrt{q_N}\right) \end{aligned}$$

The result follows. ■

Lemma C.1.10 Assume Assumption 1–5, 7 and 8. Then (i) $N^{-\frac{1}{2}}\mathbf{W}^* = N^{-\frac{1}{2}}\boldsymbol{\Delta} + o_p(1)$ (ii) $N^{-1}\mathbf{W}^{*\prime}B_N\mathbf{W}^* = \mathbf{K}_N^* + o_p(1)$ where $\mathbf{W}^* = \left[\boldsymbol{I}_{NT} - \boldsymbol{\Pi}_A\left(\boldsymbol{\Pi}'_A B_N \boldsymbol{\Pi}_A\right)^{-1}\boldsymbol{\Pi}'_A B_N\right]\mathbf{W}$ and $\mathbf{K}_N^* = \frac{1}{N}\boldsymbol{\Delta}' B_N \boldsymbol{\Delta}$.

Proof. (i) Let $P = \boldsymbol{\Pi}_A\left(\boldsymbol{\Pi}'_A B_N \boldsymbol{\Pi}_A\right)^{-1}\boldsymbol{\Pi}'_A B_N$. We can write

$$N^{-\frac{1}{2}}\mathbf{W}^* = N^{-\frac{1}{2}}[\mathbf{W} - P\mathbf{W}] = N^{-\frac{1}{2}}\boldsymbol{\Delta} + N^{-\frac{1}{2}}(H - P\mathbf{W})$$

Note that

$$\begin{aligned}\left\| N^{-\frac{1}{2}} (H - P\mathbf{W}) \right\|^2 &= N^{-1}\lambda_{\max}\left((H - P\mathbf{W})' (H - P\mathbf{W})\right) \\ &\leq N^{-1} tr\left((H - P\mathbf{W})' (H - P\mathbf{W})\right) \\ &= N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{k=1}^{K_1+T-1}\left[h_k^*(\mathbf{x}_i, \mathbf{z}_i) - \hat{h}_k(\mathbf{x}_i, \mathbf{z}_i)\right]^2 = o_p(1)\end{aligned}$$

where the last equality follows from Assumption 8.

(ii) Note

$$N^{-1}\mathbf{W}^{*\prime}B_N\mathbf{W}^* = \left(N^{-\frac{1}{2}}\mathbf{\Delta} + o_p(1)\right)' B_N \left(N^{-\frac{1}{2}}\mathbf{\Delta} + o_p(1)\right) = N^{-1}\mathbf{\Delta}' B_N \mathbf{\Delta} + o_p(1)$$

■

Lemma C.1.11 Assume Assumption 1–5, 7 and 8. Then, $\Sigma_N^{*-1/2}\tilde{\boldsymbol{\delta}}_1 \xrightarrow{d} N(0, I)$

Proof. We can write

$$\begin{aligned}\tilde{\boldsymbol{\delta}}_1 &= \sqrt{N}\left(\mathbf{W}^{*\prime}B_N\mathbf{W}^*\right)^{-1}\mathbf{W}^{*\prime}\mathbf{\Psi}_\tau(\varepsilon) \\ &= \left(\mathbf{K}_N^* + o_p(1)\right)^{-1}\left(N^{-\frac{1}{2}}\mathbf{\Delta} + N^{-\frac{1}{2}}(H - P\mathbf{W})\right)' \mathbf{\Psi}_\tau(\varepsilon)\end{aligned}$$

Note that a similar argument used in Lemma C.1.10 yields

$$\begin{aligned}\left(N^{-\frac{1}{2}}\mathbf{\Delta} + N^{-\frac{1}{2}}(H - P\mathbf{W})\right)' \mathbf{\Psi}_\tau(\varepsilon) &= N^{-\frac{1}{2}}\mathbf{\Delta}'\mathbf{\Psi}_\tau(\varepsilon) + N^{-\frac{1}{2}}(H - P\mathbf{W})' \mathbf{\Psi}_\tau(\varepsilon) \\ &= N^{-\frac{1}{2}}\mathbf{\Delta}'\mathbf{\Psi}_\tau(\varepsilon) + o_p(1)\end{aligned}$$

so that we have

$$\tilde{\boldsymbol{\delta}}_1 = \left(\mathbf{K}_N^*\right)^{-1} N^{-\frac{1}{2}}\mathbf{\Delta}'\mathbf{\Psi}_\tau(\varepsilon) + o_p(1)$$

Define $D_{Nit}^* = \Sigma_N^{*-1/2}\mathbf{K}_N^{*-1}N^{-\frac{1}{2}}\mathbf{\Delta}_{it}'\psi_\tau(\varepsilon_{it})$. Then,

$$\sum_{i=1}^{N}\sum_{t=1}^{T} D_{Nit}^* = \Sigma_N^{*-1/2}\mathbf{K}_N^{*-1}N^{-\frac{1}{2}}\mathbf{\Delta}'\mathbf{\Psi}_\tau(\varepsilon)$$

Similarly as in Lemma C.1.1, note $E\left[D^*_{Nit}\right] = 0$ and $E\left[\sum_{t=1}^T D^*_{Nit}\right] = 0$. Also,

$$\begin{aligned}
&\sum_{i=1}^N E\left[\left(\sum_{t=1}^T D^*_{Nit}\right)\left(\sum_{t=1}^T D^*_{Nit}\right)'\right] \\
&= \sum_{i=1}^N E\left[\left(\sum_{t=1}^T \Sigma_N^{*-1/2}\mathbf{K}_N^{*-1}N^{-\frac{1}{2}}\boldsymbol{\Delta}'_{it}\psi_\tau\left(\varepsilon_{it}\right)\right)\left(\sum_{t=1}^T \Sigma_N^{*-1/2}\mathbf{K}_N^{*-1}N^{-\frac{1}{2}}\boldsymbol{\Delta}'_{it}\psi_\tau\left(\varepsilon_{it}\right)\right)'\right] \\
&= E\left[\Sigma_N^{*-1/2}\mathbf{K}_N^{*-1}\left[N^{-1}\sum_{i=1}^N\left(\sum_{t=1}^T \boldsymbol{\Delta}'_{it}\psi_\tau\left(\varepsilon_{it}\right)\right)\left(\sum_{t=1}^T \psi_\tau\left(\varepsilon_{it}\right)\boldsymbol{\Delta}_{it}\right)\right]\mathbf{K}_N^{*-1}\Sigma_N^{*-1/2}\right] \\
&= E\left[\Sigma_N^{*-1/2}\mathbf{K}_N^{*-1}\mathbf{S}_N^*\mathbf{K}_N^{*-1}\Sigma_N^{*-1/2}\right] = I
\end{aligned}$$

To check Lindeberg-Feller condition, fix $\varepsilon > 0$.

$$\begin{aligned}
&\sum_{i=1}^N E\left[\left\|\sum_{t=1}^T D^*_{Nit}\right\|^2 1_{\left(\left\|\sum_{t=1}^T D^*_{Nit}\right\|>\varepsilon\right)}\right] \leq \frac{1}{\varepsilon^2}\sum_{i=1}^N E\left\|\sum_{t=1}^T D^*_{Nit}\right\|^4 \\
&\leq \frac{M}{\varepsilon^2 N^2}\sum_{i=1}^N E\left(\sum_{t=1}^T \boldsymbol{\Delta}_{it}\mathbf{K}_N^{*-1}\Sigma_N^{*-1/2}\Sigma_N^{*-1/2}\mathbf{K}_N^{*-1}\sum_{t=1}^T \boldsymbol{\Delta}'_{it}\right)^2 \\
&\leq \frac{M}{\varepsilon^2 N^2}\sum_{i=1}^N E\left\|\Sigma_N^{*-1/2}\mathbf{K}_N^{*-1}\sum_{t=1}^T \boldsymbol{\Delta}'_{it}\right\|^4 \\
&\leq \frac{C}{\varepsilon^2 N^2}\sum_{i=1}^N E\left\|\sum_{t=1}^T \boldsymbol{\Delta}'_{it}\right\|^4 = O_p\left(\frac{1}{N}\right) = o_p\left(1\right)
\end{aligned}$$

■

Proof of Theorem 3.4.8

The penalized objective function in (3.33) can be expressed as a difference of two convex functions $k\left(\boldsymbol{\beta},\boldsymbol{\gamma}\right)$ and $l\left(\boldsymbol{\beta},\boldsymbol{\gamma}\right)$

$$\frac{1}{N}\sum_{i=1}^N\sum_{t=1}^T \rho_\tau\left(y_{it} - \mathbf{w}_{it}\boldsymbol{\beta} - \boldsymbol{\pi}\left(\mathbf{x}_i,\mathbf{z}_i\right)\boldsymbol{\gamma}\right) + \sum_{j=1}^{p_N} p_\lambda\left(\left|\gamma_j\right|\right) = k\left(\boldsymbol{\beta},\boldsymbol{\gamma}\right) - l\left(\boldsymbol{\beta},\boldsymbol{\gamma}\right)$$

where

$$k(\boldsymbol{\beta},\boldsymbol{\gamma}) = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\rho_\tau\left(y_{it} - \mathbf{w}_{it}\boldsymbol{\beta} - \boldsymbol{\pi}(\mathbf{x}_i,\mathbf{z}_i)\boldsymbol{\gamma}\right) + \lambda\sum_{j=1}^{p_N}\left|\gamma_j\right|$$

$$l(\boldsymbol{\beta},\boldsymbol{\gamma}) = \sum_{j=1}^{p_N} L\left(\boldsymbol{\gamma}_j\right)$$

for some $L(\cdot)$. The function $L\left(\boldsymbol{\gamma}_j\right)$ for SCAD and MCP is defined as

$$L\left(\gamma_j\right) = \frac{\gamma_j^2 + 2\lambda\left|\gamma_j\right| + \lambda^2}{2(a-1)} 1\left[\lambda \le \left|\gamma_j\right| \le a\lambda\right] + \left(\lambda\left|\gamma_j\right| - \frac{(a+1)\lambda^2}{2}\right) 1\left[\left|\gamma_j\right| > a\lambda\right]$$

with $a > 2$, and

$$L\left(\gamma_j\right) = \frac{\gamma_j^2}{2a} 1\left[0 \le \left|\gamma_j\right| < a\lambda\right] + \left(\lambda\left|\gamma_j\right| - \frac{a\lambda^2}{2}\right) 1\left[\left|\gamma_j\right| > a\lambda\right]$$

with $(a > 1)$, respectively. Subdifferential of f at η_o is defined to be a set

$$\partial f(\eta_o) = \left\{t : f(\eta) \ge f(\eta_o) + (\eta - \eta_o)' t,\ \forall\eta\right\}$$

Then ∂l, the subdifferential of l, is merely a regular derivative

$$\partial l(\boldsymbol{\beta},\boldsymbol{\gamma}) = \left\{\left(\mu_1, \cdots, \mu_{K_4+p_N}\right) : \mu_j = \left\{\begin{array}{cc} 0 & \text{for } 1 \le j \le K_4 \\ \frac{\partial l(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial\gamma_{j-K_4}} & \text{for } K_4 + 1 \le j \le K_4 + p_N \end{array}\right\}\right\}$$

where

$$\frac{\partial l(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial\gamma_j} = \left\{\begin{array}{cc} 0 & 0 \le \left|\gamma_j\right| < \lambda \\ \frac{\gamma_j - \lambda sgn\left(\gamma_j\right)}{a-1} & \lambda \le \left|\gamma_j\right| \le a\lambda \\ \lambda sgn\left(\gamma_j\right) & \left|\gamma_j\right| > a\lambda \end{array}\right.$$

for SCAD, and

$$\frac{\partial l(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial\gamma_j} = \left\{\begin{array}{cc} \frac{\gamma_j}{a} & 0 \le \left|\gamma_j\right| < a\lambda \\ \lambda sgn\left(\gamma_j\right) & \left|\gamma_j\right| \ge a\lambda \end{array}\right.$$

for MCP. Before we derive the subdifferential of k, first consider the subgradient of the unpenalized objective function

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\rho_\tau\left(y_{it} - \mathbf{w}_{it}\boldsymbol{\beta} - \boldsymbol{\pi}(\mathbf{x}_i,\mathbf{z}_i)\boldsymbol{\gamma}\right).$$

For $1 \leq j \leq K_4$,

$$\begin{aligned} s_j\left(\boldsymbol{\beta}, \boldsymbol{\gamma}\right) = & -\tau \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{w}_{itj} 1\left[y_{it} - \mathbf{w}_{it}\boldsymbol{\beta} - \boldsymbol{\pi}\left(\mathbf{x}_i, \mathbf{z}_i\right)\boldsymbol{\gamma} > 0\right] \\ & - \left(\tau - 1\right) \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{w}_{itj} 1\left[y_{it} - \mathbf{w}_{it}\boldsymbol{\beta} - \boldsymbol{\pi}\left(\mathbf{x}_i, \mathbf{z}_i\right)\boldsymbol{\gamma} < 0\right] - \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} w_{itj} a_{it} \end{aligned}$$

and for $K_4 + 1 \leq j \leq K_4 + p_N$,

$$\begin{aligned} s_j\left(\boldsymbol{\beta}, \boldsymbol{\gamma}\right) = & -\tau \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{w}_{itj} 1\left[y_{it} - \mathbf{w}_{it}\boldsymbol{\beta} - \boldsymbol{\pi}\left(\mathbf{x}_i, \mathbf{z}_i\right)\boldsymbol{\gamma} > 0\right] \\ & - \left(\tau - 1\right) \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{w}_{itj} 1\left[y_{it} - \mathbf{w}_{it}\boldsymbol{\beta} - \boldsymbol{\pi}\left(\mathbf{x}_i, \mathbf{z}_i\right)\boldsymbol{\gamma} < 0\right] - \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{w}_{itj} a_{it} \end{aligned}$$

where $a_{it} = 0$ if $y_{it} - \mathbf{w}_{it}\boldsymbol{\beta} - \boldsymbol{\pi}\left(\mathbf{x}_i, \mathbf{z}_i\right)\boldsymbol{\gamma} \neq 0$, $a_{it} \in [\tau - 1, \tau]$ otherwise. The subgradient of k conincides with $s\left(\boldsymbol{\beta}, \boldsymbol{\gamma}\right)$ with additional term l_j introduced for $K_4 + 1 \leq j \leq K_4 + p_N$.

$$\partial k\left(\boldsymbol{\beta}, \boldsymbol{\gamma}\right) = \left\{\left(\kappa_1, \cdots, \kappa_{K_4 + p_N}\right) : \kappa_j = s_j\left(\boldsymbol{\beta}, \boldsymbol{\gamma}\right) \text{ if } 1 \leq j \leq K_4,\ s_j\left(\boldsymbol{\beta}, \boldsymbol{\gamma}\right) + l_j \text{ otherwise}\right\}$$

where $l_j = sgn\left(\gamma_{j-K_4}\right)$ if $\gamma_{j-K_4} \neq 0$ and $l_j \in [-1, 1]$ otherwise. Let $\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}\right)$ be the oracle estimator. Define $\mathcal{K}$ be a collection of vector $\kappa = \left(\kappa_1, \cdots, \kappa_{K_4 + p_N}\right)$ such that

$$\kappa_j = \begin{cases} 0 & \text{if } j = 1, \cdots, K_4 \\ \lambda sgn\left(\hat{\gamma}_{j-K_4}\right) & \text{if } j = K_4 + 1, \cdots, K_4 + q_N \\ s_j\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}\right) + \lambda l_j & \text{if } j = K_4 + q_N + 1, \cdots, K_4 + p_N \end{cases}$$

Then, Lemma C.1.12 and Lemma C.1.13 [Lemma C.1.16] below deliver the result. First note that, by Lemma C.1.13 [Lemma C.1.16], it is implied that $\lim_{N \to \infty} P\left(\mathcal{K} \subset \partial k\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}\right)\right) = 1$ since $l_j = sgn\left(\hat{\gamma}_{j-K_4}\right)$ for $j = K_4 + 1, \cdots, K_4 + q_N$. Now, consider a point $\left(\boldsymbol{\beta}, \boldsymbol{\gamma}\right) \in B\left(\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}\right), \frac{\lambda}{2}\right)$. By Lemma C.1.12, it suffices to show that there exists $\boldsymbol{\kappa}^* \in \mathcal{K}$ such that $P\left(\boldsymbol{\kappa}^* \in \partial l\left(\boldsymbol{\beta}, \boldsymbol{\gamma}\right)\right) \to 1$ i.e.

$$\lim_{N \to \infty} P\left(\kappa_j^* = \frac{\partial l\left(\boldsymbol{\beta}, \boldsymbol{\gamma}\right)}{\partial \beta_j},\ j = 1, \cdots, K_4\right) = 1 \tag{A.10}$$

$$\lim_{N \to \infty} P\left(\kappa_j^* = \frac{\partial l\left(\boldsymbol{\beta}, \boldsymbol{\gamma}\right)}{\partial \gamma_{j-K_4}},\ j = K_4 + 1, \cdots, K_4 + p_N\right) = 1 \tag{A.11}$$

Since $\frac{\partial l(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial \beta_j} = 0$ for $j = 1, \cdots, K_4$, (A.10) holds by definition of $\mathcal{K}$. To show (A.11): For both SCAD and MCP penalty function, $\frac{\partial l(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial \gamma_{j-K_4}} = \lambda sgn\left(\gamma_{j-K_4}\right)$ for $\left|\gamma_{j-K_4}\right| \geq a\lambda$ (the case with $\left|\gamma_{j-K_4}\right| = a\lambda$ can be easily checked.). By Lemma C.1.13 [Lemma C.1.16], it is implied that

$$\min_{1\leq j\leq q_N} \left|\gamma_j\right| \geq \min_{1\leq j\leq q_N} \left|\hat{\gamma}_j\right| - \max_{1\leq j\leq q_N} \left|\hat{\gamma}_j - \gamma_j\right| \geq \left(a + \frac{1}{2}\right)\lambda - \frac{\lambda}{2} = a\lambda$$

with probability approaching one. Thus, $\lim_{N\to\infty} P\left(\frac{\partial l(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial \gamma_{j-K_4}} = \lambda sgn\left(\gamma_{j-K_4}\right)\right) = 1$ for $j = K_4 + 1, \cdots, K_4 + p_N$. For $K_4 + 1 \leq j \leq K_4 + q_N$, it suffices to show

$$\lim_{N\to\infty} P\left(\lambda sgn\left(\hat{\gamma}_{j-K_4}\right) = \lambda sgn\left(\gamma_{j-K_4}\right)\right) = 1$$

since $\kappa_j^* = \lambda sgn\left(\hat{\gamma}_{j-K_4}\right)$ for $K_4 + 1 \leq j \leq K_4 + q_N$. From the fact that $\left\|\hat{\gamma}_j - \gamma_{o_j}\right\| = O_p\left(N^{-1/2}q_N^{1/2}\right) = o(\lambda)$ for $1 \leq j \leq q_N$ where $\gamma_{o_j} > 0$, and that $\left\|\hat{\gamma}_j - \gamma_j\right\| < \frac{\lambda}{2}$, it is implied that $\hat{\gamma}_{j-K_4}$ and γ_{j-K_4} have the same sign for $K_4 + 1 \leq j \leq K_4 + q_N$ with probability tending to one. For $K_4 + q_N + 1 \leq j \leq K_4 + p_N$, we have $\hat{\gamma}_{j-K_4} = 0$ by the definition of oracle estimator. Then,

$$\left|\gamma_{j-K_4}\right| \leq \left|\gamma_{j-K_4} - \hat{\gamma}_{j-K_4}\right| + \left|\hat{\gamma}_{j-K_4}\right| = \left|\gamma_{j-K_4} - \hat{\gamma}_{j-K_4}\right| < \frac{\lambda}{2}.$$

For $\left|\gamma_j\right| < \lambda$, $\frac{\partial l(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial \gamma_j} = 0$ for SCAD and $\frac{\partial l(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial \gamma_j} = \frac{\gamma_j}{a}$ for MCP, which implies $\left|\frac{\partial l(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial \gamma_{j-K_4}}\right| \leq \lambda$, $K_4 + q_N + 1 \leq j \leq K_4 + p_N$, for both penalty functions. Also, by Lemma C.1.13 [Lemma C.1.16], it is implied that $\left|s_j\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}\right)\right| \leq \frac{\lambda}{2}$ with probability tending to one for $K_4 + q_N + 1 \leq j \leq K_4 + p_N$. Therefore, there exists $l_j^* \in [-1, 1]$ such that

$$\lim_{N\to\infty} P\left(s_j\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}\right) + \lambda l_j^* = \frac{\partial l\left(\boldsymbol{\beta}, \boldsymbol{\gamma}\right)}{\partial \gamma_{j-K_4}},\ j = K_4 + q_N + 1, \cdots, K_4 + p_N\right) = 1.$$

Take $\kappa_j^* = s_j\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}\right) + \lambda l_j^*$ with such l_j^*. Then the result follows. ■

Lemma C.1.12 (Tao and An, 1997) Consider the function $k(\eta) - l(\eta)$ where both k and l are convex with subdifferential functions $\partial k(\eta)$ and $\partial l(\eta)$. Let η^* be a point that has neighborhood U such that $\partial l(\eta) \cap \partial k(\eta^*) \neq \phi, \forall \eta \in U \cap dom(k)$. Then η^* is a local minimizer of $k(\eta) - l(\eta)$.

Lemma C.1.13 Assume Assumption 1–6 and 9. Suppose $\lambda = o\left(N^{-(1-C_4)/2}\right)$, $N^{-1/2}q_N^{1/2} = o(\lambda)$, and $\log(p_N) = o\left(N\lambda^2\right)$. Let $\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}\right)$ be the oracle estimator. Then, there exists a_{it}^* with $a_{it}^* = 0$ if $y_{it} - \mathbf{w}_{it}\boldsymbol{\beta} - \boldsymbol{\pi}(\mathbf{x}_i, \mathbf{z}_i)\boldsymbol{\gamma} \neq 0$ and $a_{it}^* \in [1-\tau, \tau]$ otherwise, such that, for $s_j(\boldsymbol{\beta}, \boldsymbol{\gamma})$ with $a_{it} = a_{it}^*$, with probability approaching one

$$s_j\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}\right) = 0,\ j = 1, \cdots, K_4 + q_N \tag{A.12}$$

$$\left|\hat{\gamma}_j\right| \geq \left(a + \frac{1}{2}\right)\lambda,\ j = 1, \cdots, q_N \tag{A.13}$$

$$\left|s_j\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}\right)\right| \leq c\lambda,\ \forall c > 0,\ j = K_4 + q_N + 1, \cdots, K_4 + p_N \tag{A.14}$$

Proof. Define $\mathcal{D} = \left\{(i,t) : y_{it} - \mathbf{w}_{it}\hat{\boldsymbol{\beta}} - \boldsymbol{\pi}_A(\mathbf{x}_i, \mathbf{z}_i)\hat{\boldsymbol{\gamma}}_A = 0\right\}$. To show (A.12), note that, with probability tending to one, $(y_{it}, \tilde{\mathbf{w}}_{it})$ is in general position i.e. $|\mathcal{D}| = K_4 + q_N$ since y_{it} has a continuous density (2.2.2., Koenker, 2005). Thus, there exists $\left\{a_{it}^*\right\}$ with $(K_4 + q_N)$ nonzero elements such that (A.12) holds. (Alternatively, optimality of $\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}_A\right)$ implies $0_{K_4+p_N} \in \partial\left(\sum_i \sum_t \rho\left(y_{it} - \mathbf{w}_{it}\boldsymbol{\beta} - \boldsymbol{\pi}_A(\mathbf{x}_i, \mathbf{z}_i)\boldsymbol{\gamma}_A\right)\right)$ at $\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}\right)$ so that such $\left\{a_{it}^*\right\}$ exist.) To show (A.13), note that

$$\min_{1\leq j\leq q_N}\left|\hat{\gamma}_j\right| \geq \min_{1\leq j\leq q_N}\left|\gamma_{oj}\right| - \max_{1\leq j\leq q_N}\left|\hat{\gamma}_j - \gamma_{oj}\right|$$

By Assumption 9, $\min\limits_{1\leq j\leq q_N}\left|\gamma_{oj}\right| \geq C_5 N^{-(1-C_4)/2}$. Then, by Lemma C.1.1 and Lemma C.1.4, it is implied that $\max\limits_{1\leq j\leq q_N}\left|\hat{\gamma}_j - \gamma_{oj}\right| = O_p\left(\sqrt{\frac{q_N}{N}}\right) = o_p\left(N^{-(1-C_4)/2}\right)$. The result follows from $\lambda = o\left(N^{-(1-C_4)/2}\right)$. To show (A.14), define $J_3 \equiv \{j : K_4 + q_N + 1 \leq j \leq K_4 + p_N\}$. Then, for $j \in J_3$, by definition of $s_j\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}\right)$,

$$\begin{aligned} & s_j\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}\right) \\ &= \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{w}_{itj}\left(1\left[y_{it} - \mathbf{w}_{it}\hat{\boldsymbol{\beta}} - \boldsymbol{\pi}_A(\mathbf{x}_i, \mathbf{z}_i)\hat{\boldsymbol{\gamma}}_A \leq 0\right] - \tau\right) - \frac{1}{N}\sum_{(i,t)\in\mathcal{D}}\tilde{w}_{itj}\left(a_{it}^* + (1-\tau)\right) \end{aligned}$$

where a_{it}^* satisfies the given condition. Thus, $\frac{1}{N}\sum_{(i,t)\in\mathcal{D}}\tilde{w}_{itj}\left(a_{it}^* + (1-\tau)\right) = O_p\left(\frac{q_N}{N}\right) =$

$o_p(\lambda)$ since $|\mathcal{D}| = K_4 + q_N$ with probability tending to one. It will be shown that

$$\lim_{N\to\infty} P\left(\max_{j\in J_3}\left|\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{w}_{itj}\left(1\left[y_{it}-\mathbf{w}_{it}\hat{\boldsymbol{\beta}}-\boldsymbol{\pi}_A(\mathbf{x}_i,\mathbf{z}_i)\hat{\boldsymbol{\gamma}}_A\le 0\right]-\tau\right)\right| > c\lambda\right) = 0$$

Define

$$\begin{aligned}
I_{it}(\boldsymbol{\theta}) &= 1\left[y_{it}-\tilde{\mathbf{w}}_{it}^{A}\boldsymbol{\theta}\le 0\right]\\
P_{it}(\boldsymbol{\theta}) &= P\left(y_{it}-\tilde{\mathbf{w}}_{it}^{A}\boldsymbol{\theta}\le 0|\tilde{\mathbf{w}}_{it}^{A}\right)\\
H_{it}(\boldsymbol{\theta}) &= I_{it}(\boldsymbol{\theta})-I_{it}(\boldsymbol{\theta}_o)-P_{it}(\boldsymbol{\theta})+P_{it}(\boldsymbol{\theta}_o)
\end{aligned}$$

Then, note

$$P\left(\max_{j\in J_3}\left|\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{w}_{itj}\left(1\left[y_{it}-\mathbf{w}_{it}\hat{\boldsymbol{\beta}}-\boldsymbol{\pi}_A(\mathbf{x}_i,\mathbf{z}_i)\hat{\boldsymbol{\gamma}}_A\le 0\right]-\tau\right)\right| > c\lambda\right)$$

$$\begin{aligned}
&P\left(\max_{j\in J_3}\left|\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{w}_{itj}\left(1\left[y_{it}-\mathbf{w}_{it}\hat{\boldsymbol{\beta}}-\boldsymbol{\pi}_A(\mathbf{x}_i,\mathbf{z}_i)\hat{\boldsymbol{\gamma}}_A\le 0\right]-\tau\right)\right| > c\lambda\right)\\
&\le P\left(\max_{j\in J_3}\left|\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{w}_{itj}\left(I_{it}\left(\hat{\boldsymbol{\theta}}_A\right)-I_{it}(\boldsymbol{\theta}_{oA})\right)\right| > c\frac{\lambda}{2}\right)\\
&+P\left(\max_{j\in J_3}\left|\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{w}_{itj}\left(I_{it}(\boldsymbol{\theta}_{oA})-\tau\right)\right| > c\frac{\lambda}{2}\right)\\
&\le P\left(\max_{j\in J_3}\sup_{\|\boldsymbol{\theta}_A-\boldsymbol{\theta}_{oA}\|\le C\sqrt{\frac{q_N}{N}}}\left|\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{w}_{itj}\left(I_{it}(\boldsymbol{\theta}_A)-I_{it}(\boldsymbol{\theta}_{oA})\right)\right| > c\frac{\lambda}{2}\right)+o_p(1)\\
&\le P\left(\max_{j\in J_3}\sup_{\|\boldsymbol{\theta}_A-\boldsymbol{\theta}_{oA}\|\le C\sqrt{\frac{q_N}{N}}}\left|\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{w}_{itj}(H_{it}(\boldsymbol{\theta}_A))\right| > c\frac{\lambda}{4}\right)\\
&+P\left(\max_{j\in J_3}\sup_{\|\boldsymbol{\theta}_A-\boldsymbol{\theta}_{oA}\|\le C\sqrt{\frac{q_N}{N}}}\left|\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{w}_{itj}(P_{it}(\boldsymbol{\theta}_A)-P_{it}(\boldsymbol{\theta}_{oA}))\right| > c\frac{\lambda}{4}\right)+o_p(1)
\end{aligned}$$

where the second inequality follows by Lemma C.1.14. Here, note that

$$\begin{aligned}
&\max_{j\in J_3}\sup_{\|\boldsymbol{\theta}_A-\boldsymbol{\theta}_{oA}\|\le C\sqrt{\frac{q_N}{N}}}\left|\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{w}_{itj}\left(P_{it}\left(\boldsymbol{\theta}_A\right)-P_{it}\left(\boldsymbol{\theta}_{oA}\right)\right|\right.\\
&=\max_{j\in J_3}\sup_{\|\boldsymbol{\theta}_A-\boldsymbol{\theta}_{oA}\|\le C\sqrt{\frac{q_N}{N}}}\left|\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{w}_{itj}\left(F_{it}\left(\tilde{\mathbf{w}}_{it}^{A}\left(\boldsymbol{\theta}_A-\boldsymbol{\theta}_{oA}\right)\right)-F_{it}\left(0\right)\right|\right.\\
&\le C\sup_{\|\boldsymbol{\theta}_A-\boldsymbol{\theta}_{oA}\|\le C\sqrt{\frac{q_N}{N}}}\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\left|\tilde{\mathbf{w}}_{it}^{A}\left(\boldsymbol{\theta}_A-\boldsymbol{\theta}_{oA}\right)\right|\\
&\le C\sup_{\|\boldsymbol{\theta}_A-\boldsymbol{\theta}_{oA}\|\le C\sqrt{\frac{q_N}{N}}}\sqrt{\lambda_{\max}\left(\frac{1}{N}\tilde{\mathbf{W}}_A\tilde{\mathbf{W}}'_A\right)}\left\|\boldsymbol{\theta}_A-\boldsymbol{\theta}_{oA}\right\|\le C\sqrt{\frac{q_N}{N}}=o\left(\lambda\right)
\end{aligned}$$

Thus, now it suffices to show

$$P\left(\max_{j\in J_3}\sup_{\boldsymbol{\theta}_A\in L_N}\left|\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{w}_{itj}\left(I_{it}\left(\boldsymbol{\theta}_A\right)-I_{it}\left(\boldsymbol{\theta}_{oA}\right)-P_{it}\left(\boldsymbol{\theta}_A\right)+P_{it}\left(\boldsymbol{\theta}_{oA}\right)\right)\right|>c\lambda/4\right)$$

tends to 0 as N goes to infinity where $L_N=\left\{\boldsymbol{\theta}_A:\left\|\boldsymbol{\theta}_A-\boldsymbol{\theta}_{oA}\right\|\le C\sqrt{\frac{q_N}{N}}\right\}$. It is implied by Lemma C.1.15. ■

Lemma C.1.14 Assume Assumption 1–6. Suppose $\log\left(p_N\right)=o\left(N\lambda^2\right)$ and $N\lambda^2\to\infty$. Then

$$P\left(\max_{j\in J_3}\left|\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{w}_{itj}\left(I_{it}\left(\boldsymbol{\theta}_{oA}\right)-\tau\right)\right|>c\frac{\lambda}{2}\right)\to 0.$$

Proof. The argument is similar to Wang, Wu, Li (2012). Note that, for some constant C , the random variable $\frac{1}{C}\sum_{t=1}^{T}\tilde{w}_{itj}I_{it}\left(\boldsymbol{\theta}_{oA}\right)$ is independent across i, bounded by the interval $[0,1]$. Then, by Hoeffding's inequality, it is implied

$$P\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{w}_{itj}\left(I_{it}\left(\boldsymbol{\theta}_{oA}\right)-\tau\right)>c\frac{\lambda}{2}\right)\le exp\left(-CN\lambda^2\right)$$

so that

$$
\begin{aligned}
&P\left(\max_{j\in J_3}\left|\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{w}_{itj}\left(I_{it}\left(\boldsymbol{\theta}_{oA}\right)-\tau\right)\right|>c\frac{\lambda}{2}\right)\\
&\leq 2p_N exp\left(-CN\lambda^2\right)\\
&=2exp\left(\log p_N-CN\lambda^2\right)\to 0
\end{aligned}
$$

■

Lemma C.1.15 Assume Assumption 1–6 and 9. Suppose $\lambda = o\left(N^{-(1-C_4)/2}\right)$, $N^{-1/2}q_N^{1/2} = o\left(\lambda\right)$, and $\log\left(p_N\right) = o\left(N\lambda^2\right)$. Then, for any given positive constant C,

$$
\lim_{N\to\infty} P\left(\max_{j\in J_3}\sup_{\boldsymbol{\theta}_A\in\mathcal{B}}\left|\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{w}_{itj}(I_{it}\left(\boldsymbol{\theta}_A\right)-I_{it}\left(\boldsymbol{\theta}_{oA}\right)-P_{it}\left(\boldsymbol{\theta}_A\right)+P_{it}\left(\boldsymbol{\theta}_{oA}\right))\right|>N\lambda\right)=0
$$

where $\mathcal{B} = \left\{\boldsymbol{\theta}_A : \|\boldsymbol{\theta}_A-\boldsymbol{\theta}_{oA}\|\leq C\sqrt{\frac{q_N}{N}}\right\}$.

Proof. $\mathcal{B}$ can be covered with a net of balls with radius $C\sqrt{\frac{q_N}{N^5}}$ with cardinality $\mathcal{N}_1 \equiv |\mathcal{B}| \leq CN^{2q_N}$. Denote the $\mathcal{N}_1$ balls centered at $\boldsymbol{t}_m$ by $b\left(\boldsymbol{t}_m\right)$ for $m = 1,\cdots,\mathcal{N}_1$.

$$
\begin{aligned}
&P\left(\sup_{\|\boldsymbol{\theta}_A-\boldsymbol{\theta}_{oA}\|\leq C\sqrt{\frac{q_N}{N}}}\left|\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{w}_{itj}(I_{it}\left(\boldsymbol{\theta}_A\right)-I_{it}\left(\boldsymbol{\theta}_{oA}\right)-P_{it}\left(\boldsymbol{\theta}_A\right)+P_{it}\left(\boldsymbol{\theta}_{oA}\right))\right|>N\lambda\right)\\
&\leq\sum_{m=1}^{\mathcal{N}_1}P\left(\left|\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{w}_{itj}(I_{it}\left(\boldsymbol{t}_m\right)-I_{it}\left(\boldsymbol{\theta}_{oA}\right)-P_{it}\left(\boldsymbol{t}_m\right)+P_{it}\left(\boldsymbol{\theta}_{oA}\right))\right|>N\lambda/2\right)\\
&+\sum_{m=1}^{\mathcal{N}_1}P\left(\sup_{\left\|\tilde{\boldsymbol{\theta}}_A-\boldsymbol{t}_m\right\|\leq C\sqrt{\frac{q_N}{N^5}}}\begin{array}{c}\left|\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{w}_{itj}(I_{it}\left(\tilde{\boldsymbol{\theta}}_A\right)-I_{it}\left(\boldsymbol{t}_m\right)\right.\\ \left.-P_{it}\left(\tilde{\boldsymbol{\theta}}_A\right)+P_{it}\left(\boldsymbol{t}_m\right))\right|\end{array}>N\lambda/2\right)\\
&\equiv I_{Nj1}+I_{Nj2}.
\end{aligned}
$$

To evaluate I_{Nj1}, define $v_{ij} = \sum_{t=1}^{T}\tilde{w}_{itj}(I_{it}\left(\boldsymbol{t}_m\right)-I_{it}\left(\boldsymbol{\theta}_{oA}\right)-P_{it}\left(\boldsymbol{t}_m\right)+P_{it}\left(\boldsymbol{\theta}_{oA}\right))$, which are bounded, independent mean zero random variables. First, note that $I_{it}\left(\boldsymbol{t}_m\right)-I_{it}\left(\boldsymbol{\theta}_{oA}\right)$ is nonzero only if ($I_{it}(\boldsymbol{t}_m) = 1$ and $I_{it}(\boldsymbol{\theta}_{oA}) = 0$) or ($I_{it}(\boldsymbol{t}_m) = 0$ and $I_{it}(\boldsymbol{\theta}_{oA}) = 1$), which implies $|e_{it}| < \left|\tilde{\mathbf{w}}_{it}^{A}(\boldsymbol{t}_m - \boldsymbol{\theta}_{oA})\right|$ a.s. where $e_{it} = y_{it} - \tilde{\mathbf{w}}_{it}^{A}\boldsymbol{\theta}_{oA}$. Then,

$$\begin{aligned}
V\left(v_{ij}\right) &\leq C\sum_{t=1}^{T} V\left(\tilde{w}_{itj}(I_{it}(\boldsymbol{t}_m) - I_{it}(\boldsymbol{\theta}_{oA}) - P_{it}(\boldsymbol{t}_m) + P_{it}(\boldsymbol{\theta}_{oA}))\right) \\
&\leq C\sum_{t=1}^{T} E\left(\tilde{w}_{itj}^{2}\left[I_{it}(\boldsymbol{t}_m) - I_{it}(\boldsymbol{\theta}_{oA})\right]^2\right) \\
&\leq C\sum_{t=1}^{T} P\left(\left[I_{it}(\boldsymbol{t}_m) - I_{it}(\boldsymbol{\theta}_{oA})\right]^2 = 1\right) \\
&\leq C\sum_{t=1}^{T} P\left(|e_{it}| < \left|\tilde{\mathbf{w}}_{it}^{A}(\boldsymbol{t}_m - \boldsymbol{\theta}_{oA})\right|\right)
\end{aligned}$$

$$\begin{aligned}
&\leq C\sum_{t=1}^{T} E\left(\left[F_{it}\left(\left|\tilde{\mathbf{w}}_{it}^{A}(\boldsymbol{t}_m - \boldsymbol{\theta}_{oA})\right|\right) - F_{it}\left(-\left|\tilde{\mathbf{w}}_{it}^{A}(\boldsymbol{t}_m - \boldsymbol{\theta}_{oA})\right|\right)\right]\right) \\
&\leq CE\left(\sum_{t=1}^{T}\left|\tilde{\mathbf{w}}_{it}^{A}(\boldsymbol{t}_m - \boldsymbol{\theta}_{oA})\right|\right)
\end{aligned}$$

Therefore,

$$\begin{aligned}
\sum_{i=1}^{N} V\left(v_{ij}\right) &\leq CE\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\left|\tilde{\mathbf{w}}_{it}^{A}(\boldsymbol{t}_m - \boldsymbol{\theta}_{oA})\right|\right) \leq CNE\left(\sqrt{\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{\mathbf{w}}_{it}^{A}(\boldsymbol{t}_m - \boldsymbol{\theta}_{oA})\right)^2}\right) \\
&\leq CNE\left(\sqrt{\lambda_{\max}\left(\frac{1}{N}\tilde{W}_A\tilde{W}_A'\right)}\right)\|\boldsymbol{t}_m - \boldsymbol{\theta}_{oA}\| \leq C\sqrt{Nq_N}
\end{aligned}$$

Applying Bernstein's inequality,

$$I_{Nj1} \leq \mathcal{N}_1 exp\left(-\frac{N^2\lambda^2/8}{C\sqrt{Nq_N} + CN\lambda}\right) \leq \mathcal{N}_1 exp\left(-CN\lambda\right) \leq exp\left(Cq_N\log(N) - CN\lambda\right)$$

To evaluate I_{Nj2}, note that

$$I_{it}(\boldsymbol{\theta}_A) = 1\left[y_{it} - \tilde{\mathbf{w}}_{it}^{A}\boldsymbol{\theta}_A \leq 0\right] = 1\left[y_{it} - \tilde{\mathbf{w}}_{it}^{A}\boldsymbol{t}_m \leq \tilde{\mathbf{w}}_{it}^{A}(\boldsymbol{\theta}_A - \boldsymbol{t}_m)\right]$$

and that an indicator function is increasing. Define $I_{it}(\boldsymbol{\theta}, e) = 1\left(y_{it} - \tilde{\mathbf{w}}_{it}^{A}\boldsymbol{\theta} \le e\right)$ and $P_{it}(\boldsymbol{\theta}, e) = P\left(y_{it} - \tilde{\mathbf{w}}_{it}^{A}\boldsymbol{\theta} \le e|\tilde{\mathbf{w}}_{it}^{A}\right)$. Then, we have

$$\sup_{\left\|\tilde{\boldsymbol{\theta}}_A - \boldsymbol{t}_m\right\| \le C\sqrt{\frac{q_N}{N^5}}} \left|\sum_{i=1}^{N}\sum_{t=1}^{T} \tilde{w}_{itj}\left(I_{it}\left(\tilde{\boldsymbol{\theta}}_A\right) - I_{it}(\boldsymbol{t}_m) - P_{it}\left(\tilde{\boldsymbol{\theta}}_A\right) + P_{it}(\boldsymbol{t}_m)\right)\right|$$

$$\le \sum_{i=1}^{N}\sum_{t=1}^{T}\left|\tilde{w}_{itj}\right|\left[I_{it}\left(\boldsymbol{t}_m, \left\|\tilde{\mathbf{w}}_{it}^{A}\right\| C\sqrt{\frac{q_N}{N^5}}\right) - I_{it}(\boldsymbol{t}_m) - P_{it}\left(\boldsymbol{t}_m, -\left\|\tilde{\mathbf{w}}_{it}^{A}\right\| C\sqrt{\frac{q_N}{N^5}}\right) + P_{it}(\boldsymbol{t}_m)\right]$$

$$= \sum_{i=1}^{N}\sum_{t=1}^{T}\left|\tilde{w}_{itj}\right|\left[I_{it}\left(\boldsymbol{t}_m, \left\|\tilde{\mathbf{w}}_{it}^{A}\right\| C\sqrt{\frac{q_N}{N^5}}\right) - I_{it}(\boldsymbol{t}_m) - P_{it}\left(\boldsymbol{t}_m, \left\|\tilde{\mathbf{w}}_{it}^{A}\right\| C\sqrt{\frac{q_N}{N^5}}\right) + P_{it}(\boldsymbol{t}_m)\right]$$

$$+ \sum_{i=1}^{N}\sum_{t=1}^{T}\left|\tilde{w}_{itj}\right|\left[P_{it}\left(\boldsymbol{t}_m, \left\|\tilde{\mathbf{w}}_{it}^{A}\right\| C\sqrt{\frac{q_N}{N^5}}\right) - P_{it}\left(\boldsymbol{t}_m, -\left\|\tilde{\mathbf{w}}_{it}^{A}\right\| C\sqrt{\frac{q_N}{N^5}}\right)\right]$$

Note that

$$\sum_{i=1}^{N}\sum_{t=1}^{T}\left|\tilde{w}_{itj}\right|\left[P_{it}\left(\boldsymbol{t}_m, \left\|\tilde{\mathbf{w}}_{it}^{A}\right\| C\sqrt{\frac{q_N}{N^5}}\right) - P_{it}\left(\boldsymbol{t}_m, -\left\|\tilde{\mathbf{w}}_{it}^{A}\right\| C\sqrt{\frac{q_N}{N^5}}\right)\right]$$

$$= \sum_{i=1}^{N}\sum_{t=1}^{T}\left|\tilde{w}_{itj}\right|\left[F_{it}\left(\tilde{\mathbf{w}}_{it}^{A}(\boldsymbol{t}_m - \boldsymbol{\theta}_{oA}) + \left\|\tilde{\mathbf{w}}_{it}^{A}\right\| C\sqrt{\frac{q_N}{N^5}}\right)\right.$$

$$\left.-F_{it}\left(\tilde{\mathbf{w}}_{it}^{A}(\boldsymbol{t}_m - \boldsymbol{\theta}_{oA}) - \left\|\tilde{\mathbf{w}}_{it}^{A}\right\| C\sqrt{\frac{q_N}{N^5}}\right)\right]$$

$$\le C\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|\tilde{\mathbf{w}}_{it}^{A}\right\|\sqrt{\frac{q_N}{N^5}} \le Cq_N N^{-3/2} = o(N\lambda)$$

Hence, for all N sufficiently large, $I_{Nj2} \le \sum_{m=1}^{\mathcal{N}_1} P\left(\sum_{i=1}^{N}\alpha_{mi} \ge \frac{N\lambda}{4}\right)$ where

$$\alpha_{mi} = \sum_{t=1}^{T}\left|\tilde{w}_{itj}\right|\left[I_{it}\left(\boldsymbol{t}_m, \left\|\tilde{\mathbf{w}}_{it}^{A}\right\| C\sqrt{\frac{q_N}{N^5}}\right) - I_{it}(\boldsymbol{t}_m) - P_{it}\left(\boldsymbol{t}_m, \left\|\tilde{\mathbf{w}}_{it}^{A}\right\| C\sqrt{\frac{q_N}{N^5}}\right) + P_{it}(\boldsymbol{t}_m)\right]$$

Since α_{mi} are independent bounded random variables with mean zero, similarly as in the evaluation of I_{Nj1}, we can show that

$$V(\alpha_{mi}) \le C\sum_{t=1}^{T} E\left(\sqrt{q_N/N^5}\left\|\tilde{\mathbf{w}}_{it}^{A}\right\|\right) \le Cq_N N^{-5/2}$$

Applying Bernstein's inequality,

$$\begin{aligned}\sum_{m=1}^{\mathcal{N}_1} P\left(\sum_{i=1}^{N} \alpha_{mi} \geq \frac{N\lambda}{4}\right) &\leq \mathcal{N}_1 exp\left(-\frac{N^2\lambda^2/32}{Cq_N N^{-3/2} + CN\lambda}\right) \\ &\leq \mathcal{N}_1 exp\left(-CN\lambda\right) \\ &\leq Cexp\left(Cq_N \log N - CN\lambda\right)\end{aligned}$$

Therefore, we have

$$\sum_{j\in J_3} \left(I_{Nj1} + I_{Nj2}\right) \leq Cexp\left(\log p_N + Cq_N \log\left(N\right) - CN\lambda\right) = o\left(1\right)$$

since $N^{-\frac{1}{2}} q_N^{\frac{1}{2}} \log N = o\left(1\right)$ is implied by conditions given. Hence the result. ■

Lemma C.1.16 Assume Assumption 1–5, and 7–9. Suppose $\lambda = o\left(N^{-(1-C_4)/2}\right)$, $N^{-1/2} q_N^{1/2} = o\left(\lambda\right)$ and $\log\left(p_N\right) = o\left(N\lambda^2\right)$. Let $\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}\right)$ be the oracle estimator. Then, there exists a_{it}^* with $a_{it}^* = 0$ if $y_{it} - \mathbf{w}_{it}\boldsymbol{\beta} - \boldsymbol{\pi}\left(\mathbf{x}_i, \mathbf{z}_i\right)\boldsymbol{\gamma} \neq 0$ and $a_{it}^* \in [1-\tau, \tau]$ otherwise, such that, for $s_j\left(\boldsymbol{\beta}, \boldsymbol{\gamma}\right)$ with $a_{it} = a_{it}^*$, with probability approaching one

$$s_j\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}\right) = 0,\ j = 1, \cdots, K_4 + q_N \tag{A.15}$$

$$\left|\hat{\gamma}_j\right| \geq \left(a + \frac{1}{2}\right)\lambda,\ j = 1, \cdots, q_N \tag{A.16}$$

$$\left|s_j\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}\right)\right| \leq c\lambda,\ \forall c > 0,\ j = K_4 + q_N + 1, \cdots, K_4 + p_N \tag{A.17}$$

Proof. Define $\mathcal{D} = \left\{(i,t) : y_{it} - \mathbf{w}_{it}\hat{\boldsymbol{\beta}} - \boldsymbol{\pi}_A\left(\mathbf{x}_i, \mathbf{z}_i\right)\hat{\boldsymbol{\gamma}}_A = 0\right\}$. To show (A.15), note that, with probability tending to one, $(y_{it}, \tilde{\mathbf{w}}_{it})$ is in general position i.e. $|\mathcal{D}| = K_4 + q_N$ since y_{it} has a continuous density (2.2.2., Koenker, 2005). Thus, there exists $\left\{a_{it}^*\right\}$ with $(K_4 + q_N)$ nonzero elements such that (A.12) holds. (Alternatively, optimality of $\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}_A\right)$ implies $0_{K_4+p_N} \in \partial\left(\sum_i \sum_t \rho\left(y_{it} - \mathbf{w}_{it}\boldsymbol{\beta} - \boldsymbol{\pi}_A\left(\mathbf{x}_i, \mathbf{z}_i\right)\boldsymbol{\gamma}_A\right)\right)$ at $\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}\right)$ so that such $\left\{a_{it}^*\right\}$ exist.) To show (A.16), note that

$$\min_{1\leq j\leq q_N} \left|\hat{\gamma}_j\right| \geq \min_{1\leq j\leq q_N} \left|\gamma_{oj}\right| - \max_{1\leq j\leq q_N} \left|\hat{\gamma}_j - \gamma_{oj}\right|$$

By Assumption 9, $\min_{1\le j\le q_N} \left|\gamma_{oj}\right| \ge C_5 N^{-(1-C_4)/2}$. Then, by Lemma C.1.1 and Lemma C.1.9, it is implied that $\max_{1\le j\le q_N} \left|\hat{\gamma}_j - \gamma_{oj}\right| = O_p\left(\sqrt{\frac{q_N}{N}}\right) = o_p\left(N^{-(1-C_4)/2}\right)$. The result follows from $\lambda = o\left(N^{-(1-C_4)/2}\right)$. To show (A.17), define $J_3 \equiv \{j : K_4 + q_N + 1 \le j \le K_4 + p_N\}$. Then, for $j \in J_3$, by definition of $s_j\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}\right)$,

$$\begin{aligned} & s_j\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}\right) \\ & = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} \tilde{w}_{itj}\left(1\left[y_{it} - \mathbf{w}_{it}\hat{\boldsymbol{\beta}} - \boldsymbol{\pi}_A\left(\mathbf{x}_i, \mathbf{z}_i\right)\hat{\boldsymbol{\gamma}}_A \le 0\right] - \tau\right) - \frac{1}{N}\sum_{(i,t)\in\mathcal{D}} \tilde{w}_{itj}\left(a_{it}^* + (1-\tau)\right) \end{aligned}$$

where a_{it}^* satisfies the given condition. Thus, $\frac{1}{N}\sum_{(i,t)\in\mathcal{D}} \tilde{w}_{itj}\left(a_{it}^* + (1-\tau)\right) = O_p\left(\frac{q_N}{N}\right) = o_p(\lambda)$ since $|\mathcal{D}| = K_4 + q_N$ with probability tending to one. It will be shown that

$$\lim_{N\to\infty} P\left(\max_{j\in J_3}\left|\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{w}_{itj}\left(1\left[y_{it} - \mathbf{w}_{it}\hat{\boldsymbol{\beta}} - \boldsymbol{\pi}_A\left(\mathbf{x}_i, \mathbf{z}_i\right)\hat{\boldsymbol{\gamma}}_A \le 0\right] - \tau\right)\right| > c\lambda\right) = 0$$

Define

$$\begin{aligned} I_{it}(\boldsymbol{\theta}) &= 1\left[y_{it} - \tilde{\mathbf{w}}_{it}^A\boldsymbol{\theta} \le 0\right] \\ I_{it}^o &= 1\left[y_{it} - \mathbf{w}_{it}\boldsymbol{\beta}_o - g_o\left(\mathbf{x}_i, \mathbf{z}_i\right) \le 0\right] \\ P_{it}(\boldsymbol{\theta}) &= P\left(y_{it} - \tilde{\mathbf{w}}_{it}^A\boldsymbol{\theta} \le 0 | \tilde{\mathbf{w}}_{it}^A\right) \\ P_{it}^o &= P\left(y_{it} - \mathbf{w}_{it}\boldsymbol{\beta}_o - g_o\left(\mathbf{x}_i, \mathbf{z}_i\right) \le 0 | \tilde{\mathbf{w}}_{it}^A\right) \end{aligned}$$

Then, note

$$P\left(\max_{j\in J_3}\left|\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{w}_{itj}\left(1\left[y_{it} - \mathbf{w}_{it}\hat{\boldsymbol{\beta}} - \boldsymbol{\pi}_A\left(\mathbf{x}_i, \mathbf{z}_i\right)\hat{\boldsymbol{\gamma}}_A \le 0\right] - \tau\right)\right| > c\lambda\right)$$

$$
\begin{aligned}
&\le P\left(\max_{j\in J_3}\left|\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{w}_{itj}\left(I_{it}\left(\hat{\boldsymbol{\theta}}_A\right)-I_{it}^{o}\right)\right|>c\frac{\lambda}{2}\right)\\
&+P\left(\max_{j\in J_3}\left|\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{w}_{itj}\left(I_{it}^{o}-\tau\right)\right|>c\frac{\lambda}{2}\right)\\
&\le P\left(\max_{j\in J_3}\sup_{\|\boldsymbol{\theta}_A-\boldsymbol{\theta}_{oA}\|\le C\sqrt{\frac{q_N}{N}}}\left|\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{w}_{itj}\left(I_{it}\left(\boldsymbol{\theta}_A\right)-I_{it}^{o}\right)\right|>c\frac{\lambda}{2}\right)+o_p\left(1\right)
\end{aligned}
$$

$$
\begin{aligned}
&\le P\left(\max_{j\in J_3}\sup_{\|\boldsymbol{\theta}_A-\boldsymbol{\theta}_{oA}\|\le C\sqrt{\frac{q_N}{N}}}\left|\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{w}_{itj}(I_{it}\left(\boldsymbol{\theta}_A\right)-I_{it}^{o}-P_{it}\left(\boldsymbol{\theta}_A\right)+P_{it}^{o})\right|>c\frac{\lambda}{4}\right)\\
&+P\left(\max_{j\in J_3}\sup_{\|\boldsymbol{\theta}_A-\boldsymbol{\theta}_{oA}\|\le C\sqrt{\frac{q_N}{N}}}\left|\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{w}_{itj}(P_{it}\left(\boldsymbol{\theta}_A\right)-P_{it}^{o})\right|>c\frac{\lambda}{4}\right)+o_p\left(1\right)
\end{aligned}
$$

where the second inequality follows by Lemma C.1.17. Here, note that

$$
\begin{aligned}
&\max_{j\in J_3}\sup_{\|\boldsymbol{\theta}_A-\boldsymbol{\theta}_{oA}\|\le C\sqrt{\frac{q_N}{N}}}\left|\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{w}_{itj}(P_{it}\left(\boldsymbol{\theta}_A\right)-P_{it}^{o})\right|\\
&=\max_{j\in J_3}\sup_{\|\boldsymbol{\theta}_A-\boldsymbol{\theta}_{oA}\|\le C\sqrt{\frac{q_N}{N}}}\left|\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{w}_{itj}(F_{it}\left(\mathbf{w}_{it}\left(\boldsymbol{\beta}-\boldsymbol{\beta}_o\right)-r_i\right)-F_{it}\left(0\right)\right|\\
&\le C\sup_{\|\boldsymbol{\theta}_A-\boldsymbol{\theta}_{oA}\|\le C\sqrt{\frac{q_N}{N}}}\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}(|\mathbf{w}_{it}\left(\boldsymbol{\beta}-\boldsymbol{\beta}_o\right)|+|r_i|)\\
&\le C\sup_{\|\boldsymbol{\theta}_A-\boldsymbol{\theta}_{oA}\|\le C\sqrt{\frac{q_N}{N}}}[\sqrt{\lambda_{\max}\left(\frac{1}{N}\tilde{\mathbf{W}}_A\tilde{\mathbf{W}}_A'\right)}\left\|\boldsymbol{\theta}_A-\boldsymbol{\theta}_{oA}\right\|+\sup_i|r_i|]\\
&\le C\left(\sqrt{\frac{q_N}{N}}+\sqrt{\frac{q_N}{N}}\right)=o\left(\lambda\right)
\end{aligned}
$$

Thus, now it suffices to show

$$
P\left(\max_{j\in J_3}\sup_{\|\boldsymbol{\theta}_A-\boldsymbol{\theta}_{oA}\|\le C\sqrt{\frac{q_N}{N}}}\left|\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{w}_{itj}(I_{it}\left(\boldsymbol{\theta}_A\right)-I_{it}^{o}-P_{it}\left(\boldsymbol{\theta}_A\right)+P_{it}^{o})\right|>c\lambda/4\right)
$$

tends to 0 as N goes to infinity, which is implied by Lemma C.1.18. ■

Lemma C.1.17 Assume Assumption 1–5, and 7–9. Suppose $\log(p_N) = o\left(N\lambda^2\right)$ and $N\lambda^2 \to \infty$. Then

$$P\left(\max_{j\in J_3}\left|\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{w}_{itj}\left(I_{it}^{o}-\tau\right)\right| > c\frac{\lambda}{2}\right) \to 0.$$

Proof. The argument is similar to Wang, Wu, Li (2012). Note that, for some constant C , the random variable $\frac{1}{C}\sum_{t=1}^{T}\tilde{w}_{itj}I_{it}^{o}$ is independent across i, bounded by the interval $[0, 1]$. Then, by Hoeffding's inequality, it is implied

$$P\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{w}_{itj}\left(I_{it}^{o}-\tau\right) > c\frac{\lambda}{2}\right) \leq exp\left(-CN\lambda^2\right)$$

so that

$$\begin{aligned}
&P\left(\max_{j\in J_3}\left|\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{w}_{itj}\left(I_{it}^{o}-\tau\right)\right| > c\frac{\lambda}{2}\right)\\
&\leq 2p_N exp\left(-CN\lambda^2\right)\\
&= 2exp\left(\log p_N - CN\lambda^2\right) \to 0
\end{aligned}$$

■

Lemma C.1.18 Assume Assumption 1–5, and 7–9. Suppose $\lambda = o\left(N^{-(1-C_4)/2}\right)$, $N^{-1/2}q_N^{1/2} = o(\lambda)$, and $\log(p_N) = o\left(N\lambda^2\right)$. Then, for any given positive constant C,

$$\lim_{N\to\infty} P\left(\max_{j\in J_3}\sup_{\|\boldsymbol{\theta}_A-\boldsymbol{\theta}_{oA}\|\leq C\sqrt{\frac{q_N}{N}}}\left|\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{w}_{itj}\left(I_{it}\left(\boldsymbol{\theta}_A\right)-I_{it}^{o}-P_{it}\left(\boldsymbol{\theta}_A\right)+P_{it}^{o}\right)\right| > N\lambda\right) = 0$$

Proof. Let $\mathcal{B} = \left\{\boldsymbol{\theta}_A : \|\boldsymbol{\theta}_A-\boldsymbol{\theta}_{oA}\| \leq C\sqrt{\frac{q_N}{N}}\right\}$. $\mathcal{B}$ can be covered with a net of balls with radius $C\sqrt{\frac{q_N}{N^5}}$ with cardinality $\mathcal{N}_1 \equiv |\mathcal{B}| \leq CN^{2q_N}$. Denote the $\mathcal{N}_1$ balls centered at $\boldsymbol{t}_m = (\boldsymbol{t}_{m1}, \boldsymbol{t}_{m2})$

by $b\left(\boldsymbol{t}_m\right)$ for $m = 1, \cdots, \mathcal{N}_1$.

$$
\begin{aligned}
& P\left(\sup_{\left\|\boldsymbol{\theta}_A-\boldsymbol{\theta}_{oA}\right\| \leq C\sqrt{\frac{q_N}{N}}}\left|\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{w}_{itj}\left(I_{it}\left(\boldsymbol{\theta}_A\right)-I_{it}^{o}-P_{it}\left(\boldsymbol{\theta}_A\right)+P_{it}^{o}\right)\right|>N\lambda\right) \\
& \leq \sum_{m=1}^{\mathcal{N}_1} P\left(\left|\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{w}_{itj}\left(I_{it}\left(\boldsymbol{t}_m\right)-I_{it}^{o}-P_{it}\left(\boldsymbol{t}_m\right)+P_{it}^{o}\right)\right|>N\lambda/2\right) \\
& +\sum_{m=1}^{\mathcal{N}_1} P\left(\sup_{\left\|\tilde{\boldsymbol{\theta}}_A-\boldsymbol{t}_m\right\| \leq C\sqrt{\frac{q_N}{N^5}}} \begin{array}{c}\left|\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{w}_{itj}\left(I_{it}\left(\tilde{\boldsymbol{\theta}}_A\right)-I_{it}\left(\boldsymbol{t}_m\right)\right.\right. \\ \left.\left.-P_{it}\left(\tilde{\boldsymbol{\theta}}_A\right)+P_{it}\left(\boldsymbol{t}_m\right)\right)\right|\end{array} > N\lambda/2\right) \\
& \equiv I_{Nj1}+I_{Nj2}.
\end{aligned}
$$

To evaluate I_{Nj1}, define $v_{ij} = \sum_{t=1}^{T}\tilde{w}_{itj}\left(I_{it}\left(\boldsymbol{t}_m\right)-I_{it}^{o}-P_{it}\left(\boldsymbol{t}_m\right)+P_{it}^{o}\right)$, which are bounded, independent mean zero random variables. First, note that $I_{it}\left(\boldsymbol{t}_m\right)-I_{it}^{o}$ is nonzero only if ($I_{it}\left(\boldsymbol{t}_m\right) = 1$ and $I_{it}^{o} = 0$) or ($I_{it}\left(\boldsymbol{t}_m\right) = 0$ and $I_{it}^{o} = 1$), which implies $\left|\varepsilon_{it}\right| < \left|\mathbf{w}_{it}\left(\boldsymbol{t}_{m1}-\boldsymbol{\beta}_o\right)-r_i\right|$ a.s. Then,

$$
\begin{aligned}
V\left(v_{ij}\right) & \leq C\sum_{t=1}^{T} V\left(\tilde{w}_{itj}\left(I_{it}\left(\boldsymbol{t}_m\right)-I_{it}^{o}-P_{it}\left(\boldsymbol{t}_m\right)+P_{it}^{o}\right)\right) \\
& \leq C\sum_{t=1}^{T} E\left(\tilde{w}_{itj}^{2}\left[I_{it}\left(\boldsymbol{t}_m\right)-I_{it}^{o}\right]^{2}\right) \\
& \leq C\sum_{t=1}^{T} P\left(\left[I_{it}\left(\boldsymbol{t}_m\right)-I_{it}^{o}\right]^{2}=1\right) \\
& \leq C\sum_{t=1}^{T} P\left(\left|\varepsilon_{it}\right| < \left|\mathbf{w}_{it}\left(\boldsymbol{t}_{m1}-\boldsymbol{\beta}_o\right)-r_i\right|\right) \\
& \leq C\sum_{t=1}^{T} E\left(\left[F_{it}\left(\left|\mathbf{w}_{it}\left(\boldsymbol{t}_{m1}-\boldsymbol{\beta}_o\right)-r_i\right|\right)-F_{it}\left(-\left|\mathbf{w}_{it}\left(\boldsymbol{t}_{m1}-\boldsymbol{\beta}_o\right)-r_i\right|\right)\right]\right) \\
& \leq CE\left(\sum_{t=1}^{T}\left|\mathbf{w}_{it}\left(\boldsymbol{t}_{m1}-\boldsymbol{\beta}_o\right)-r_i\right|\right)
\end{aligned}
$$

Therefore,

$$\sum_{i=1}^{N} V\left(v_{ij}\right) \leq CE\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\left|\mathbf{w}_{it}\left(\boldsymbol{t}_{m1}-\boldsymbol{\beta}_{o}\right)-r_i\right|\right) \leq CNE\left(\sqrt{\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\mathbf{w}_{it}\left(\boldsymbol{t}_{m1}-\boldsymbol{\beta}_{o}\right)-r_i\right)^2}\right)$$
$$\leq CN\left[E\left(\sqrt{\lambda_{\max}\left(\frac{1}{N}\tilde{W}_A\tilde{W}_A'\right)}\right)\left\|\boldsymbol{t}_m-\boldsymbol{\theta}_{oA}\right\| + \sup\left|r_i\right|\right] \leq C\sqrt{Nq_N}$$

Applying Bernstein's inequality,

$$I_{Nj1} \leq \mathcal{N}_1 exp\left(-\frac{N^2\lambda^2/8}{C\sqrt{Nq_N}+CN\lambda}\right) \leq \mathcal{N}_1 exp\left(-CN\lambda\right) \leq exp\left(Cq_N\log\left(N\right)-CN\lambda\right)$$

To evaluate I_{Nj2}, note that

$$I_{it}\left(\boldsymbol{\theta}_A\right) = 1\left[y_{it}-\tilde{\mathbf{w}}_{it}^{A}\boldsymbol{\theta}_A \leq 0\right] = 1\left[y_{it}-\tilde{\mathbf{w}}_{it}^{A}\boldsymbol{t}_m \leq \tilde{\mathbf{w}}_{it}^{A}\left(\boldsymbol{\theta}_A-\boldsymbol{t}_m\right)\right]$$

and that an indicator function is increasing. Define $I_{it}\left(\boldsymbol{\theta},e\right) = 1\left(y_{it}-\tilde{\mathbf{w}}_{it}^{A}\boldsymbol{\theta} \leq e\right)$ and $P_{it}\left(\boldsymbol{\theta},e\right) = P\left(y_{it}-\tilde{\mathbf{w}}_{it}^{A}\boldsymbol{\theta} \leq e|\tilde{\mathbf{w}}_{it}^{A}\right)$. Then, we have

$$\sup_{\left\|\tilde{\boldsymbol{\theta}}_A-\boldsymbol{t}_m\right\|\leq C\sqrt{\frac{q_N}{N^5}}}\left|\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{w}_{itj}\left(I_{it}\left(\tilde{\boldsymbol{\theta}}_A\right)-I_{it}\left(\boldsymbol{t}_m\right)-P_{it}\left(\tilde{\boldsymbol{\theta}}_A\right)+P_{it}\left(\boldsymbol{t}_m\right)\right)\right|$$
$$\leq \sum_{i=1}^{N}\sum_{t=1}^{T}\left|\tilde{w}_{itj}\right|\left[I_{it}\left(\boldsymbol{t}_m,\left\|\tilde{\mathbf{w}}_{it}^{A}\right\|C\sqrt{\frac{q_N}{N^5}}\right)-I_{it}\left(\boldsymbol{t}_m\right)-P_{it}\left(\boldsymbol{t}_m,-\left\|\tilde{\mathbf{w}}_{it}^{A}\right\|C\sqrt{\frac{q_N}{N^5}}\right)+P_{it}\left(\boldsymbol{t}_m\right)\right]$$
$$= \sum_{i=1}^{N}\sum_{t=1}^{T}\left|\tilde{w}_{itj}\right|\left[I_{it}\left(\boldsymbol{t}_m,\left\|\tilde{\mathbf{w}}_{it}^{A}\right\|C\sqrt{\frac{q_N}{N^5}}\right)-I_{it}\left(\boldsymbol{t}_m\right)-P_{it}\left(\boldsymbol{t}_m,\left\|\tilde{\mathbf{w}}_{it}^{A}\right\|C\sqrt{\frac{q_N}{N^5}}\right)+P_{it}\left(\boldsymbol{t}_m\right)\right]$$
$$+ \sum_{i=1}^{N}\sum_{t=1}^{T}\left|\tilde{w}_{itj}\right|\left[P_{it}\left(\boldsymbol{t}_m,\left\|\tilde{\mathbf{w}}_{it}^{A}\right\|C\sqrt{\frac{q_N}{N^5}}\right)-P_{it}\left(\boldsymbol{t}_m,-\left\|\tilde{\mathbf{w}}_{it}^{A}\right\|C\sqrt{\frac{q_N}{N^5}}\right)\right]$$

Note that

$$\begin{aligned}
&\sum_{i=1}^{N}\sum_{t=1}^{T}\left|\tilde{w}_{itj}\right|\left[P_{it}\left(\boldsymbol{t}_m,\left\|\tilde{\mathbf{w}}_{it}^{A}\right\|C\sqrt{\frac{q_N}{N^5}}\right)-P_{it}\left(\boldsymbol{t}_m,-\left\|\tilde{\mathbf{w}}_{it}^{A}\right\|C\sqrt{\frac{q_N}{N^5}}\right)\right]\\
&=\sum_{i=1}^{N}\sum_{t=1}^{T}\left|\tilde{w}_{itj}\right|\Bigg[F_{it}\left(\mathbf{w}_{it}\left(\boldsymbol{t}_{m1}\boldsymbol{-}\boldsymbol{\beta}_o\right)-r_i+\left\|\tilde{\mathbf{w}}_{it}^{A}\right\|C\sqrt{\frac{q_N}{N^5}}\right)\\
&-F_{it}\left(\mathbf{w}_{it}\left(\boldsymbol{t}_{m1}\boldsymbol{-}\boldsymbol{\beta}_o\right)-r_i-\left\|\tilde{\mathbf{w}}_{it}^{A}\right\|C\sqrt{\frac{q_N}{N^5}}\right)\Bigg]\\
&\le C\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|\tilde{\mathbf{w}}_{it}^{A}\right\|\sqrt{\frac{q_N}{N^5}}\le Cq_N N^{-3/2}=o\left(N\lambda\right)
\end{aligned}$$

Hence, for all N sufficiently large, $I_{Nj2}\le\sum_{m=1}^{\mathcal{N}_1}P\left(\sum_{i=1}^{N}\alpha_{mi}\ge\frac{N\lambda}{4}\right)$ where

$$\alpha_{mi}=\sum_{t=1}^{T}\left|\tilde{w}_{itj}\right|\left[I_{it}\left(\boldsymbol{t}_m,\left\|\tilde{\mathbf{w}}_{it}^{A}\right\|C\sqrt{\frac{q_N}{N^5}}\right)-I_{it}\left(\boldsymbol{t}_m\right)-P_{it}\left(\boldsymbol{t}_m,\left\|\tilde{\mathbf{w}}_{it}^{A}\right\|C\sqrt{\frac{q_N}{N^5}}\right)+P_{it}\left(\boldsymbol{t}_m\right)\right]$$

Since α_{mi} are independent bounded random variables with mean zero, similarly as in the evaluation of I_{Nj1}, we can show that

$$V\left(\alpha_{mi}\right)\le C\sum_{t=1}^{T}E\left(\sqrt{q_N/N^5}\left\|\tilde{\mathbf{w}}_{it}^{A}\right\|\right)\le Cq_N N^{-5/2}$$

Applying Bernstein's inequality,

$$\begin{aligned}
\sum_{m=1}^{\mathcal{N}_1}P\left(\sum_{i=1}^{N}\alpha_{mi}\ge\frac{N\lambda}{4}\right)&\le\mathcal{N}_1 exp\left(-\frac{N^2\lambda^2/32}{Cq_N N^{-3/2}+CN\lambda}\right)\\
&\le\mathcal{N}_1 exp\left(-CN\lambda\right)\\
&\le Cexp\left(Cq_N\log N-CN\lambda\right)
\end{aligned}$$

Therefore, we have

$$\sum_{j\in J_3}\left(I_{Nj1}+I_{Nj2}\right)\le Cexp\left(\log p_N+Cq_N\log\left(N\right)-CN\lambda\right)=o\left(1\right)$$

since $N^{-\frac{1}{2}}q_N^{\frac{1}{2}}\log N=o\left(1\right)$ is implied by conditions given. Hence the result. ■

C.2 Appendix: Supplementary Tables

Figure A.1 Pooled Birth Weights

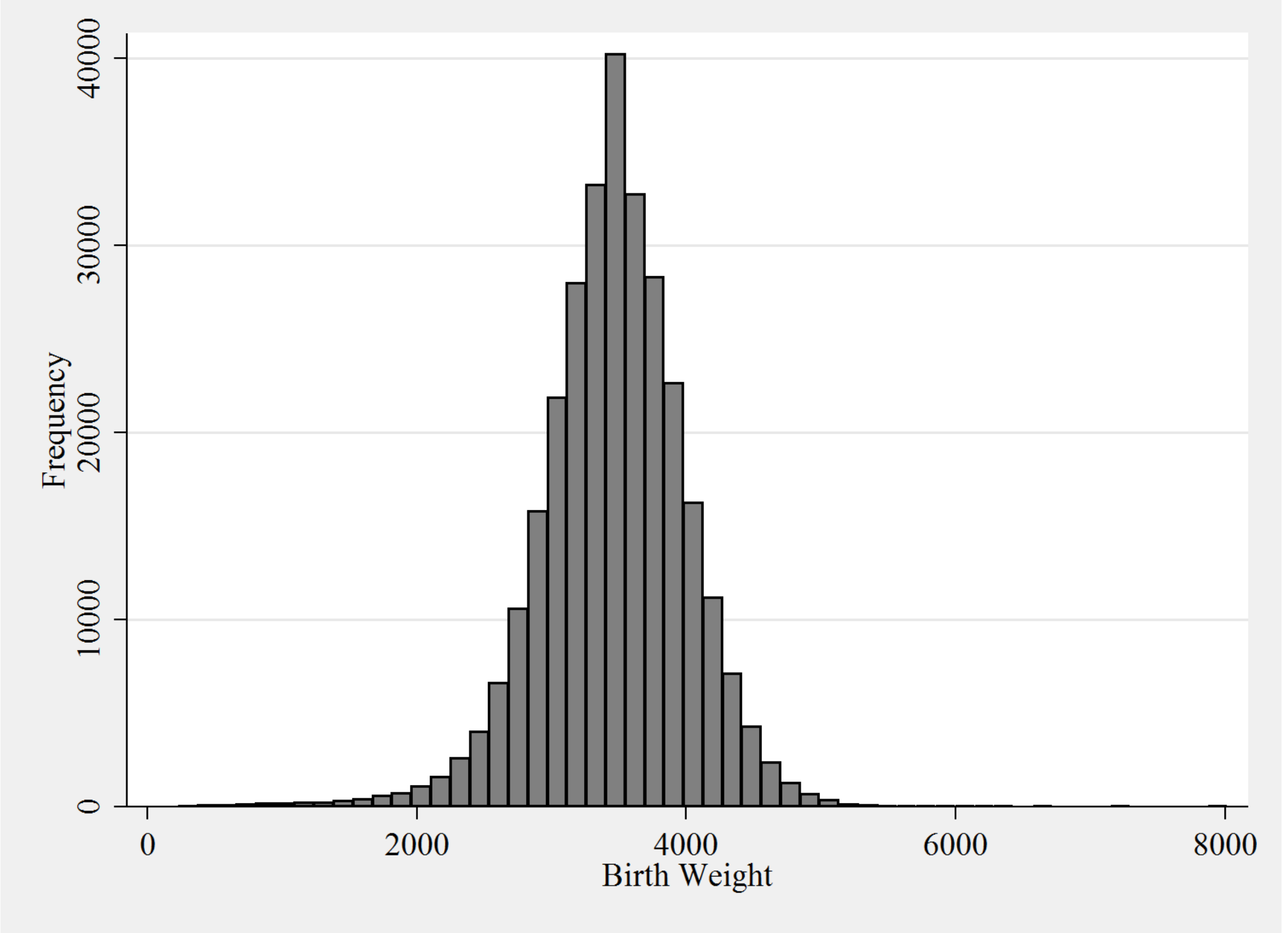

Table A.1 Estimator performance, DGP 1, $\beta 2$

			$\tau = 0.1$			$\tau = 0.5$			$\tau = 0.9$		
Method	N	IC	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
gMund	300	QBIC	-0.0005	0.0217	0.0216	0.0010	0.0356	0.0356	-0.0003	0.0215	0.0215
gMund	300	QBICL	-0.0006	0.0216	0.0216	0.0011	0.0357	0.0357	-0.0003	0.0215	0.0215
gMund	300	BIC	-0.0006	0.0216	0.0216	0.0011	0.0357	0.0357	-0.0003	0.0215	0.0215
gMund	300	BICL	-0.0006	0.0216	0.0216	0.0011	0.0357	0.0357	-0.0003	0.0215	0.0215
gMund	300	AIC1	-0.0000	0.0217	0.0217	0.0019	0.0355	0.0356	-0.0000	0.0213	0.0212
gMund	300	AIC2	-0.0006	0.0216	0.0216	0.0011	0.0358	0.0358	-0.0003	0.0215	0.0215
gMund	1000	QBIC	0.0001	0.0111	0.0111	0.0012	0.0185	0.0186	-0.0000	0.0111	0.0111
gMund	1000	QBICL	0.0001	0.0112	0.0112	0.0011	0.0186	0.0186	-0.0001	0.0111	0.0111
gMund	1000	BIC	0.0001	0.0112	0.0112	0.0011	0.0186	0.0186	-0.0001	0.0111	0.0111
gMund	1000	BICL	0.0001	0.0112	0.0112	0.0011	0.0186	0.0186	-0.0001	0.0111	0.0111
gMund	1000	AIC1	0.0001	0.0112	0.0112	0.0011	0.0183	0.0183	-0.0001	0.0110	0.0110
gMund	1000	AIC2	0.0001	0.0112	0.0112	0.0012	0.0186	0.0186	-0.0001	0.0111	0.0111
gCham	300	QBIC	0.0009	0.0201	0.0201	-0.0005	0.0347	0.0347	0.0010	0.0195	0.0195
gCham	300	QBICL	0.0008	0.0200	0.0200	-0.0008	0.0346	0.0346	0.0009	0.0196	0.0196
gCham	300	BIC	0.0008	0.0200	0.0200	-0.0008	0.0346	0.0346	0.0009	0.0196	0.0196
gCham	300	BICL	0.0008	0.0200	0.0200	-0.0008	0.0346	0.0346	0.0009	0.0196	0.0196
gCham	300	AIC1	0.0010	0.0205	0.0205	-0.0002	0.0346	0.0346	0.0008	0.0191	0.0191
gCham	300	AIC2	0.0008	0.0200	0.0200	-0.0007	0.0346	0.0346	0.0009	0.0196	0.0196
gCham	1000	QBIC	0.0002	0.0110	0.0110	0.0008	0.0196	0.0196	0.0004	0.0115	0.0115
gCham	1000	QBICL	0.0002	0.0110	0.0110	0.0007	0.0196	0.0196	0.0005	0.0116	0.0116
gCham	1000	BIC	0.0002	0.0110	0.0110	0.0007	0.0196	0.0196	0.0005	0.0116	0.0116
gCham	1000	BICL	0.0002	0.0110	0.0110	0.0007	0.0196	0.0196	0.0005	0.0116	0.0116
gCham	1000	AIC1	0.0001	0.0109	0.0109	0.0010	0.0196	0.0196	0.0007	0.0117	0.0117
gCham	1000	AIC2	0.0002	0.0110	0.0110	0.0007	0.0196	0.0196	0.0005	0.0116	0.0116

Table A.2 Estimator performance, DGP 2, $\beta 2$

			$\tau = 0.1$			$\tau = 0.5$			$\tau = 0.9$		
Method	N	IC	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
gMund	300	QBIC	0.0105	0.1643	0.1645	-0.0078	0.2587	0.2587	-0.0078	0.1596	0.1597
gMund	300	QBICL	0.0085	0.1642	0.1643	-0.0021	0.2635	0.2634	-0.0063	0.1621	0.1622
gMund	300	BIC	0.0082	0.1638	0.1639	-0.0075	0.2587	0.2587	-0.0065	0.1622	0.1622
gMund	300	BICL	0.0085	0.1642	0.1643	-0.0021	0.2635	0.2634	-0.0063	0.1621	0.1622
gMund	300	AIC1	0.0186	0.1665	0.1674	-0.0041	0.2580	0.2579	-0.0119	0.1598	0.1601
gMund	300	AIC2	0.0119	0.1648	0.1652	-0.0039	0.2579	0.2578	-0.0080	0.1610	0.1612
gMund	1000	QBIC	0.0057	0.0887	0.0889	0.0093	0.1473	0.1476	-0.0054	0.0876	0.0878
gMund	1000	QBICL	0.0045	0.0882	0.0883	0.0101	0.1479	0.1481	-0.0049	0.0857	0.0858
gMund	1000	BIC	0.0045	0.0883	0.0884	0.0091	0.1474	0.1476	-0.0048	0.0856	0.0857
gMund	1000	BICL	0.0045	0.0882	0.0883	0.0101	0.1479	0.1481	-0.0049	0.0857	0.0858
gMund	1000	AIC1	0.0081	0.0878	0.0882	0.0080	0.1478	0.1480	-0.0067	0.0878	0.0880
gMund	1000	AIC2	0.0068	0.0895	0.0897	0.0078	0.1477	0.1479	-0.0045	0.0876	0.0876
gCham	300	QBIC	0.0238	0.1623	0.1640	0.0129	0.2653	0.2655	-0.0121	0.1608	0.1612
gCham	300	QBICL	0.0269	0.1639	0.1660	0.0243	0.2830	0.2839	-0.0089	0.1644	0.1646
gCham	300	BIC	0.0266	0.1640	0.1660	0.0135	0.2650	0.2652	-0.0109	0.1622	0.1625
gCham	300	BICL	1.0523	0.4443	1.1421	0.0239	0.2825	0.2834	-0.0036	0.1691	0.1691
gCham	300	AIC1	0.0240	0.1593	0.1610	0.0055	0.2659	0.2658	-0.0201	0.1646	0.1657
gCham	300	AIC2	0.0237	0.1613	0.1629	0.0059	0.2659	0.2658	-0.0143	0.1623	0.1628
gCham	1000	QBIC	0.0004	0.0891	0.0890	-0.0078	0.1433	0.1434	0.0047	0.0865	0.0866
gCham	1000	QBICL	0.0011	0.0886	0.0885	-0.0061	0.1445	0.1446	0.0049	0.0859	0.0860
gCham	1000	BIC	0.0006	0.0886	0.0886	-0.0078	0.1433	0.1434	0.0050	0.0859	0.0860
gCham	1000	BICL	0.0014	0.0886	0.0886	-0.0061	0.1445	0.1446	0.0050	0.0862	0.0863
gCham	1000	AIC1	0.0004	0.0878	0.0878	-0.0097	0.1432	0.1435	0.0035	0.0867	0.0867
gCham	1000	AIC2	0.0015	0.0881	0.0880	-0.0098	0.1434	0.1437	0.0040	0.0863	0.0863

Table A.3 Selection Performance, DGP 3

					$\tau = 0.1$			$\tau = 0.5$			$\tau = 0.9$		
Method	p_N	q_o	N	IC	TV	FV	True	TV	FV	True	TV	FV	True
gMund	136	6	300	QBIC	5.14	2.61	0.07	4.98	2.19	0.10	5.72	0.80	0.59
gMund	136	6	300	QBICL	4.86	1.52	0.07	4.74	1.58	0.11	5.73	0.36	0.71
gMund	136	6	300	BIC	4.64	1.36	0.07	4.71	1.49	0.11	5.70	0.31	0.71
gMund	136	6	300	BICL	4.42	1.12	0.07	4.64	1.36	0.11	5.37	0.44	0.67
gMund	136	6	300	AIC1	5.15	4.33	0.06	5.30	4.83	0.07	5.43	3.62	0.16
gMund	136	6	300	AIC2	4.76	1.42	0.07	4.81	1.82	0.11	5.72	0.33	0.71
gMund	136	6	1000	QBIC	5.52	1.81	0.19	5.89	0.93	0.42	5.92	0.28	0.83
gMund	136	6	1000	QBICL	5.53	1.54	0.20	5.51	0.65	0.46	5.96	0.12	0.92
gMund	136	6	1000	BIC	5.15	1.05	0.20	5.41	0.63	0.46	5.92	0.09	0.92
gMund	136	6	1000	BICL	4.85	1.15	0.20	5.35	0.65	0.46	5.91	0.09	0.92
gMund	136	6	1000	AIC1	5.62	3.06	0.13	5.89	4.38	0.23	5.79	3.35	0.23
gMund	136	6	1000	AIC2	5.52	1.53	0.20	5.83	0.74	0.45	5.96	0.12	0.92
gCham	102	18	300	QBIC	14.89	5.64	0.00	13.46	8.86	0.00	14.71	6.45	0.00
gCham	102	18	300	QBICL	14.31	4.51	0.00	12.65	5.84	0.00	14.22	5.04	0.00
gCham	102	18	300	BIC	13.49	4.54	0.00	12.52	5.61	0.00	13.32	4.87	0.00
gCham	102	18	300	BICL	2.11	0.40	0.00	8.89	3.89	0.00	2.76	5.65	0.00
gCham	102	18	300	AIC1	15.02	9.44	0.00	14.11	11.92	0.00	14.88	10.49	0.00
gCham	102	18	300	AIC2	14.11	4.44	0.00	12.96	6.95	0.00	14.08	4.96	0.00
gCham	102	18	1000	QBIC	16.61	3.98	0.05	15.59	5.12	0.00	16.89	3.31	0.10
gCham	102	18	1000	QBICL	16.39	2.45	0.05	14.34	4.67	0.00	16.62	2.24	0.10
gCham	102	18	1000	BIC	15.66	2.43	0.05	14.11	4.71	0.00	15.91	2.20	0.10
gCham	102	18	1000	BICL	15.29	2.50	0.05	13.43	4.70	0.00	15.40	2.44	0.10
gCham	102	18	1000	AIC1	16.50	7.18	0.02	16.63	7.98	0.00	16.65	7.30	0.02
gCham	102	18	1000	AIC2	16.35	2.36	0.05	14.92	4.67	0.00	16.57	2.15	0.10

Table A.4 Estimator performance, DGP 3, $\beta 1$

			$\tau = 0.1$			$\tau = 0.5$			$\tau = 0.9$		
Method	N	IC	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
gMund	300	QBIC	0.0015	0.0210	0.0210	0.0016	0.0364	0.0365	-0.0001	0.0200	0.0200
gMund	300	QBICL	0.0005	0.0212	0.0212	0.0012	0.0366	0.0366	-0.0001	0.0203	0.0203
gMund	300	BIC	0.0004	0.0214	0.0214	0.0009	0.0363	0.0363	-0.0002	0.0204	0.0204
gMund	300	BICL	0.0074	0.0404	0.0411	0.0002	0.0359	0.0359	0.0100	0.0419	0.0431
gMund	300	AIC1	0.0016	0.0207	0.0208	0.0021	0.0364	0.0365	0.0001	0.0202	0.0202
gMund	300	AIC2	0.0005	0.0214	0.0214	0.0013	0.0362	0.0362	-0.0001	0.0204	0.0203
gMund	1000	QBIC	0.0007	0.0116	0.0116	0.0016	0.0191	0.0191	-0.0000	0.0115	0.0115
gMund	1000	QBICL	0.0007	0.0116	0.0116	0.0015	0.0192	0.0193	0.0000	0.0116	0.0116
gMund	1000	BIC	0.0008	0.0120	0.0120	0.0013	0.0192	0.0193	-0.0001	0.0115	0.0115
gMund	1000	BICL	0.0008	0.0123	0.0123	0.0013	0.0192	0.0193	-0.0001	0.0115	0.0115
gMund	1000	AIC1	0.0009	0.0117	0.0117	0.0015	0.0188	0.0189	-0.0001	0.0115	0.0115
gMund	1000	AIC2	0.0007	0.0116	0.0116	0.0016	0.0191	0.0192	0.0000	0.0116	0.0116
gCham	300	QBIC	0.0012	0.0211	0.0212	0.0033	0.0332	0.0333	-0.0004	0.0226	0.0226
gCham	300	QBICL	0.0010	0.0215	0.0216	0.0004	0.0335	0.0335	-0.0007	0.0229	0.0229
gCham	300	BIC	0.0011	0.0235	0.0235	-0.0003	0.0331	0.0331	-0.0002	0.0247	0.0247
gCham	300	BICL	0.8448	0.5553	1.0109	0.0629	0.0530	0.0823	0.1379	0.0987	0.1695
gCham	300	AIC1	0.0006	0.0209	0.0209	0.0037	0.0332	0.0334	-0.0005	0.0219	0.0219
gCham	300	AIC2	0.0010	0.0220	0.0220	0.0017	0.0336	0.0337	-0.0008	0.0234	0.0234
gCham	1000	QBIC	0.0001	0.0116	0.0116	0.0004	0.0190	0.0190	0.0001	0.0110	0.0110
gCham	1000	QBICL	0.0001	0.0118	0.0118	0.0001	0.0189	0.0189	0.0002	0.0110	0.0110
gCham	1000	BIC	0.0002	0.0119	0.0119	0.0002	0.0189	0.0189	0.0002	0.0114	0.0114
gCham	1000	BICL	0.0010	0.0141	0.0142	-0.0004	0.0186	0.0186	0.0039	0.0249	0.0252
gCham	1000	AIC1	-0.0000	0.0116	0.0116	0.0004	0.0191	0.0191	0.0002	0.0110	0.0110
gCham	1000	AIC2	0.0001	0.0117	0.0117	0.0003	0.0189	0.0189	0.0002	0.0111	0.0111

Table A.5 Estimator performance, DGP 3, $\beta 2$

			$\tau = 0.1$			$\tau = 0.5$			$\tau = 0.9$		
Method	N	IC	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
gMund	300	QBIC	0.0015	0.0215	0.0215	0.0015	0.0355	0.0355	0.0010	0.0212	0.0212
gMund	300	QBICL	0.0011	0.0219	0.0219	0.0008	0.0356	0.0356	0.0010	0.0211	0.0211
gMund	300	BIC	0.0004	0.0222	0.0222	0.0006	0.0355	0.0355	0.0011	0.0214	0.0214
gMund	300	BICL	0.0060	0.0404	0.0409	0.0004	0.0352	0.0352	0.0121	0.0443	0.0459
gMund	300	AIC1	0.0012	0.0211	0.0211	0.0015	0.0359	0.0359	0.0011	0.0209	0.0209
gMund	300	AIC2	0.0008	0.0222	0.0222	0.0011	0.0355	0.0355	0.0010	0.0212	0.0212
gMund	1000	QBIC	0.0005	0.0117	0.0117	0.0001	0.0195	0.0195	0.0011	0.0114	0.0115
gMund	1000	QBICL	0.0005	0.0117	0.0117	-0.0002	0.0195	0.0195	0.0011	0.0115	0.0115
gMund	1000	BIC	0.0004	0.0117	0.0117	-0.0001	0.0195	0.0195	0.0010	0.0114	0.0114
gMund	1000	BICL	0.0004	0.0118	0.0118	-0.0001	0.0195	0.0195	0.0010	0.0114	0.0114
gMund	1000	AIC1	0.0005	0.0117	0.0117	0.0002	0.0196	0.0196	0.0011	0.0115	0.0116
gMund	1000	AIC2	0.0005	0.0117	0.0117	0.0001	0.0196	0.0196	0.0011	0.0115	0.0115
gCham	300	QBIC	-0.0002	0.0216	0.0216	0.0037	0.0345	0.0347	-0.0001	0.0220	0.0220
gCham	300	QBICL	0.0005	0.0223	0.0223	0.0011	0.0329	0.0329	0.0003	0.0227	0.0227
gCham	300	BIC	0.0009	0.0236	0.0237	0.0006	0.0329	0.0329	0.0004	0.0240	0.0240
gCham	300	BICL	0.8354	0.5497	0.9999	0.0606	0.0533	0.0806	0.1367	0.0998	0.1692
gCham	300	AIC1	-0.0005	0.0213	0.0213	0.0044	0.0348	0.0350	-0.0005	0.0217	0.0217
gCham	300	AIC2	0.0008	0.0227	0.0227	0.0022	0.0333	0.0334	0.0004	0.0227	0.0227
gCham	1000	QBIC	-0.0006	0.0118	0.0119	0.0008	0.0191	0.0191	0.0002	0.0113	0.0113
gCham	1000	QBICL	-0.0004	0.0119	0.0119	0.0007	0.0190	0.0190	0.0000	0.0113	0.0113
gCham	1000	BIC	-0.0005	0.0120	0.0120	0.0008	0.0189	0.0189	0.0002	0.0115	0.0115
gCham	1000	BICL	0.0004	0.0148	0.0148	0.0005	0.0191	0.0191	0.0040	0.0264	0.0267
gCham	1000	AIC1	-0.0007	0.0120	0.0120	0.0008	0.0191	0.0191	-0.0000	0.0115	0.0115
gCham	1000	AIC2	-0.0006	0.0119	0.0119	0.0007	0.0192	0.0192	0.0000	0.0113	0.0113

Table A.6 Selection Performance, DGP 4

					$\tau = 0.1$			$\tau = 0.5$			$\tau = 0.9$		
Method	p_N	q_o	N	IC	TV	FV	True	TV	FV	True	TV	FV	True
gMund	136	6	300	QBIC	3.41	2.23	0.00	2.67	2.40	0.00	3.71	3.36	0.00
gMund	136	6	300	QBICL	1.65	1.01	0.00	1.17	1.05	0.00	2.31	2.06	0.00
gMund	136	6	300	BIC	3.08	1.63	0.00	3.21	4.42	0.00	3.60	2.78	0.00
gMund	136	6	300	BICL	1.38	0.94	0.00	1.66	1.30	0.00	1.58	1.93	0.00
gMund	136	6	300	AIC1	4.20	7.48	0.00	3.62	7.22	0.00	3.99	10.00	0.00
gMund	136	6	300	AIC2	4.04	5.49	0.00	4.05	11.82	0.00	3.90	7.45	0.00
gMund	136	6	1000	QBIC	4.11	2.40	0.00	3.65	2.52	0.00	4.38	3.13	0.00
gMund	136	6	1000	QBICL	3.17	1.34	0.00	1.97	1.43	0.00	3.90	2.04	0.00
gMund	136	6	1000	BIC	4.00	2.06	0.00	4.02	4.30	0.00	4.30	2.60	0.00
gMund	136	6	1000	BICL	2.54	1.11	0.00	3.23	1.95	0.00	3.66	2.02	0.00
gMund	136	6	1000	AIC1	4.70	8.67	0.00	4.52	8.66	0.00	4.66	10.86	0.00
gMund	136	6	1000	AIC2	4.57	6.39	0.00	4.94	16.30	0.00	4.56	7.98	0.00
gCham	102	18	300	QBIC	6.56	4.06	0.00	4.83	3.82	0.00	7.83	7.27	0.00
gCham	102	18	300	QBICL	1.39	1.13	0.00	0.93	0.88	0.00	3.18	4.04	0.00
gCham	102	18	300	BIC	5.75	3.46	0.00	6.11	5.06	0.00	6.84	6.49	0.00
gCham	102	18	300	BICL	0.46	0.43	0.00	3.37	2.77	0.00	2.86	3.79	0.00
gCham	102	18	300	AIC1	9.18	8.27	0.00	7.31	6.96	0.00	9.79	11.46	0.00
gCham	102	18	300	AIC2	8.53	6.57	0.00	8.69	10.22	0.00	9.44	10.01	0.00
gCham	102	18	1000	QBIC	9.79	5.52	0.00	7.84	5.37	0.00	11.51	7.60	0.00
gCham	102	18	1000	QBICL	4.52	2.81	0.00	3.39	3.15	0.00	7.19	6.56	0.00
gCham	102	18	1000	BIC	8.85	4.85	0.00	9.34	6.88	0.00	11.15	7.29	0.00
gCham	102	18	1000	BICL	3.99	2.66	0.00	4.61	3.66	0.00	5.12	6.07	0.00
gCham	102	18	1000	AIC1	11.88	10.11	0.00	10.41	9.05	0.00	12.38	12.00	0.00
gCham	102	18	1000	AIC2	11.55	8.30	0.00	11.45	12.97	0.00	12.17	10.11	0.00

Table A.7 Estimator performance, DGP 4, $\beta 1$

			$\tau = 0.1$			$\tau = 0.5$			$\tau = 0.9$		
Method	N	IC	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
gMund	300	QBIC	0.0259	0.3429	0.3437	0.0434	0.5665	0.5678	-0.0069	0.3193	0.3192
gMund	300	QBICL	0.0796	0.3854	0.3933	0.1254	0.6198	0.6321	0.0966	0.3656	0.3780
gMund	300	BIC	0.0322	0.3489	0.3503	0.0324	0.5584	0.5591	0.0008	0.3270	0.3268
gMund	300	BICL	0.1409	0.4682	0.4887	0.0866	0.5692	0.5755	0.1341	0.3740	0.3971
gMund	300	AIC1	0.0177	0.3421	0.3424	0.0293	0.5506	0.5511	0.0005	0.3233	0.3232
gMund	300	AIC2	0.0177	0.3431	0.3434	0.0170	0.5498	0.5498	-0.0009	0.3214	0.3213
gMund	1000	QBIC	0.0060	0.1788	0.1788	0.0186	0.3064	0.3068	-0.0023	0.1883	0.1882
gMund	1000	QBICL	0.0443	0.1950	0.1999	0.0763	0.3240	0.3327	0.0088	0.1979	0.1980
gMund	1000	BIC	0.0086	0.1805	0.1806	0.0100	0.3030	0.3030	-0.0027	0.1877	0.1876
gMund	1000	BICL	0.0625	0.1943	0.2040	0.0313	0.3152	0.3166	0.0336	0.2165	0.2190
gMund	1000	AIC1	0.0053	0.1760	0.1760	0.0063	0.3036	0.3035	-0.0030	0.1929	0.1928
gMund	1000	AIC2	0.0069	0.1783	0.1784	0.0086	0.3025	0.3025	-0.0031	0.1940	0.1940
gCham	300	QBIC	0.1006	0.3474	0.3615	0.1191	0.5735	0.5855	0.0432	0.3505	0.3530
gCham	300	QBICLP	0.9157	0.6470	1.1210	1.3293	0.8436	1.5742	0.1974	0.4299	0.4728
gCham	300	BIC	0.1263	0.3550	0.3766	0.0694	0.5610	0.5650	0.0725	0.3544	0.3615
gCham	300	BICLP	1.2698	0.5288	1.3754	0.3097	0.7033	0.7681	0.2272	0.4578	0.5109
gCham	300	AIC1	0.0437	0.3407	0.3434	0.0429	0.5478	0.5492	0.0062	0.3382	0.3381
gCham	300	AIC2	0.0517	0.3419	0.3456	0.0225	0.5443	0.5445	0.0107	0.3414	0.3414
gCham	1000	QBIC	0.0305	0.1835	0.1859	0.0531	0.3109	0.3153	0.0011	0.1881	0.1880
gCham	1000	QBICLP	0.0775	0.1901	0.2052	0.1170	0.3379	0.3575	0.1160	0.2168	0.2458
gCham	1000	BIC	0.0426	0.1863	0.1910	0.0348	0.3064	0.3082	0.0036	0.1900	0.1899
gCham	1000	BICLP	0.0831	0.1974	0.2140	0.0951	0.3092	0.3234	0.1638	0.2254	0.2785
gCham	1000	AIC1	0.0098	0.1815	0.1817	0.0276	0.3046	0.3057	-0.0052	0.1890	0.1890
gCham	1000	AIC2	0.0128	0.1810	0.1814	0.0230	0.3024	0.3032	-0.0028	0.1902	0.1902

Table A.8 Estimator performance, DGP 4, $\beta 2$

			$\tau = 0.1$			$\tau = 0.5$			$\tau = 0.9$		
Method	N	IC	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
gMund	300	QBIC	0.0423	0.3284	0.3309	0.0432	0.5459	0.5474	0.0055	0.3433	0.3432
gMund	300	QBICL	0.0882	0.3400	0.3511	0.1255	0.5864	0.5993	0.1222	0.3856	0.4043
gMund	300	BIC	0.0542	0.3286	0.3329	0.0313	0.5319	0.5325	0.0072	0.3433	0.3432
gMund	300	BICL	0.1200	0.3883	0.4062	0.0873	0.5363	0.5431	0.1595	0.3912	0.4223
gMund	300	AIC1	0.0342	0.3250	0.3267	0.0288	0.5270	0.5275	-0.0005	0.3371	0.3369
gMund	300	AIC2	0.0332	0.3238	0.3253	0.0177	0.5267	0.5267	-0.0004	0.3392	0.3390
gMund	1000	QBIC	0.0148	0.1835	0.1840	0.0238	0.3133	0.3141	0.0051	0.1823	0.1823
gMund	1000	QBICL	0.0477	0.1956	0.2012	0.0883	0.3228	0.3345	0.0160	0.1915	0.1921
gMund	1000	BIC	0.0164	0.1861	0.1867	0.0175	0.3102	0.3105	0.0047	0.1799	0.1798
gMund	1000	BICL	0.0637	0.1998	0.2096	0.0387	0.3205	0.3227	0.0433	0.2108	0.2151
gMund	1000	AIC1	0.0159	0.1819	0.1825	0.0181	0.3070	0.3074	0.0051	0.1783	0.1783
gMund	1000	AIC2	0.0170	0.1829	0.1836	0.0162	0.3046	0.3049	0.0042	0.1800	0.1799
gCham	300	QBIC	0.0723	0.3467	0.3540	0.0962	0.5598	0.5677	0.0496	0.3642	0.3674
gCham	300	QBICLP	0.8832	0.6648	1.1052	1.2828	0.8614	1.5449	0.1973	0.4430	0.4847
gCham	300	BIC	0.0982	0.3581	0.3712	0.0447	0.5466	0.5481	0.0728	0.3712	0.3781
gCham	300	BICLP	1.2474	0.5287	1.3547	0.2693	0.6714	0.7230	0.2520	0.4945	0.5548
gCham	300	AIC1	0.0127	0.3366	0.3367	0.0280	0.5434	0.5439	0.0062	0.3535	0.3534
gCham	300	AIC2	0.0284	0.3383	0.3393	0.0091	0.5350	0.5348	0.0071	0.3532	0.3531
gCham	1000	QBIC	0.0249	0.1887	0.1902	0.0317	0.3141	0.3156	-0.0009	0.1844	0.1843
gCham	1000	QBICLP	0.0757	0.1926	0.2068	0.0948	0.3322	0.3453	0.1151	0.2145	0.2433
gCham	1000	BIC	0.0387	0.1891	0.1930	0.0171	0.3109	0.3113	0.0053	0.1856	0.1855
gCham	1000	BICLP	0.0784	0.1966	0.2116	0.0794	0.3121	0.3219	0.1628	0.2199	0.2735
gCham	1000	AIC1	0.0060	0.1902	0.1902	0.0111	0.3101	0.3102	-0.0059	0.1852	0.1852
gCham	1000	AIC2	0.0056	0.1915	0.1915	0.0068	0.3080	0.3079	-0.0054	0.1850	0.1850

Table A.9 Selection Performance, DGP 5

					$\tau = 0.1$			$\tau = 0.5$			$\tau = 0.9$		
Method	p_N	q_o	N	IC	TV	FV	True	TV	FV	True	TV	FV	True
gMund	136	18	300	QBIC	7.54	3.67	0.00	6.01	3.88	0.00	8.77	6.34	0.00
gMund	136	18	300	QBICL	3.20	1.44	0.00	2.35	1.75	0.00	4.20	4.30	0.00
gMund	136	18	300	BIC	6.87	3.08	0.00	7.26	5.30	0.00	8.00	5.70	0.00
gMund	136	18	300	BICL	2.56	1.22	0.00	3.65	2.39	0.00	2.92	3.82	0.00
gMund	136	18	300	AIC1	9.21	7.09	0.00	8.13	6.76	0.00	10.57	10.81	0.00
gMund	136	18	300	AIC2	8.72	5.70	0.00	9.23	9.66	0.00	10.11	9.03	0.00
gMund	136	18	1000	QBIC	9.65	4.82	0.00	8.44	4.93	0.00	12.12	7.14	0.00
gMund	136	18	1000	QBICL	7.43	2.17	0.00	4.56	2.52	0.00	7.87	5.17	0.00
gMund	136	18	1000	BIC	9.23	4.20	0.00	9.45	6.24	0.00	11.72	6.67	0.00
gMund	136	18	1000	BICL	5.93	1.49	0.00	7.17	3.81	0.00	6.99	5.08	0.00
gMund	136	18	1000	AIC1	11.72	8.97	0.00	10.65	8.74	0.00	13.08	11.36	0.00
gMund	136	18	1000	AIC2	11.22	7.50	0.00	11.85	12.94	0.00	12.87	9.64	0.00
gCham	102	12	300	QBIC	5.96	3.42	0.00	4.85	3.34	0.00	6.65	5.16	0.00
gCham	102	12	300	QBICL	3.09	1.08	0.00	2.08	1.20	0.00	3.81	3.47	0.00
gCham	102	12	300	BIC	5.54	2.72	0.00	5.67	4.96	0.00	6.29	4.63	0.00
gCham	102	12	300	BICL	2.28	0.83	0.00	3.59	2.12	0.00	3.18	3.34	0.00
gCham	102	12	300	AIC1	7.23	8.79	0.00	6.39	7.43	0.00	7.47	9.64	0.00
gCham	102	12	300	AIC2	6.92	6.75	0.00	7.27	11.71	0.00	7.27	7.80	0.00
gCham	102	12	1000	QBIC	7.51	4.17	0.00	6.53	4.10	0.00	8.41	5.03	0.00
gCham	102	12	1000	QBICL	5.51	1.96	0.00	4.61	2.51	0.00	6.97	3.99	0.00
gCham	102	12	1000	BIC	7.23	3.71	0.00	7.36	5.65	0.00	8.24	4.64	0.00
gCham	102	12	1000	BICL	5.24	1.55	0.00	5.40	3.02	0.00	6.12	3.84	0.00
gCham	102	12	1000	AIC1	8.54	10.60	0.00	8.20	8.74	0.00	8.95	10.53	0.00
gCham	102	12	1000	AIC2	8.31	8.01	0.00	9.03	15.04	0.00	8.81	8.05	0.00

Table A.10 Estimator performance, DGP 5, $\beta 1$

			$\tau = 0.1$			$\tau = 0.5$			$\tau = 0.9$		
Method	N	IC	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
gMund	300	QBIC	0.0459	0.3642	0.3669	-0.0024	0.5246	0.5244	0.0090	0.3442	0.3442
gMund	300	QBICL	0.1934	0.4944	0.5306	0.1810	0.6825	0.7057	0.1306	0.4318	0.4510
gMund	300	BIC	0.0575	0.3677	0.3720	-0.0092	0.5188	0.5186	0.0202	0.3543	0.3547
gMund	300	BICL	0.3393	0.6274	0.7130	0.0659	0.5413	0.5451	0.1669	0.4681	0.4968
gMund	300	AIC1	0.0346	0.3594	0.3609	-0.0166	0.5249	0.5249	0.0017	0.3366	0.3365
gMund	300	AIC2	0.0371	0.3593	0.3610	-0.0193	0.5259	0.5260	0.0042	0.3374	0.3372
gMund	1000	QBIC	0.0152	0.1861	0.1866	0.0038	0.3030	0.3028	-0.0013	0.1798	0.1797
gMund	1000	QBICL	0.0339	0.1926	0.1955	0.0632	0.3224	0.3284	0.0338	0.2153	0.2179
gMund	1000	BIC	0.0155	0.1856	0.1862	0.0041	0.3028	0.3027	-0.0019	0.1792	0.1791
gMund	1000	BICL	0.0513	0.2024	0.2087	0.0134	0.3038	0.3040	0.0634	0.2334	0.2418
gMund	1000	AIC1	0.0153	0.1836	0.1841	0.0045	0.3034	0.3033	-0.0054	0.1789	0.1789
gMund	1000	AIC2	0.0151	0.1832	0.1838	0.0044	0.3056	0.3055	-0.0040	0.1786	0.1786
gCham	300	QBIC	0.0202	0.3090	0.3095	0.0732	0.5035	0.5086	0.0128	0.3246	0.3247
gCham	300	QBICL	0.2321	0.4315	0.4898	0.4136	0.6022	0.7303	0.0861	0.3633	0.3732
gCham	300	BIC	0.0277	0.3097	0.3108	0.0538	0.4978	0.5005	0.0221	0.3240	0.3245
gCham	300	BICL	0.3751	0.4462	0.5828	0.1377	0.5489	0.5657	0.1168	0.3952	0.4119
gCham	300	AIC1	0.0009	0.3242	0.3240	0.0529	0.5055	0.5080	-0.0024	0.3244	0.3242
gCham	300	AIC2	0.0091	0.3209	0.3208	0.0462	0.5086	0.5104	0.0023	0.3201	0.3199
gCham	1000	QBIC	0.0141	0.1704	0.1709	0.0226	0.2826	0.2834	-0.0037	0.1704	0.1703
gCham	1000	QBICL	0.0255	0.1691	0.1710	0.0413	0.2899	0.2927	0.0202	0.1730	0.1741
gCham	1000	BIC	0.0161	0.1713	0.1719	0.0177	0.2830	0.2834	-0.0031	0.1706	0.1705
gCham	1000	BICL	0.0256	0.1704	0.1722	0.0317	0.2777	0.2794	0.0359	0.1766	0.1801
gCham	1000	AIC1	0.0137	0.1785	0.1790	0.0210	0.2877	0.2883	-0.0052	0.1772	0.1772
gCham	1000	AIC2	0.0100	0.1755	0.1757	0.0232	0.2894	0.2902	-0.0051	0.1763	0.1763

Table A.11 Estimator performance, DGP 5, $\beta 2$

			$\tau = 0.1$			$\tau = 0.5$			$\tau = 0.9$		
Method	N	IC	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
gMund	300	QBIC	0.0468	0.3428	0.3458	0.0297	0.5341	0.5347	-0.0059	0.3508	0.3506
gMund	300	QBICL	0.1815	0.4963	0.5282	0.1784	0.6529	0.6765	0.1127	0.4217	0.4363
gMund	300	BIC	0.0535	0.3491	0.3530	0.0139	0.5313	0.5312	0.0040	0.3557	0.3555
gMund	300	BICL	0.3173	0.6270	0.7024	0.0868	0.5283	0.5351	0.1521	0.4538	0.4784
gMund	300	AIC1	0.0305	0.3319	0.3331	0.0090	0.5312	0.5310	-0.0101	0.3416	0.3416
gMund	300	AIC2	0.0372	0.3360	0.3379	0.0029	0.5309	0.5306	-0.0105	0.3444	0.3444
gMund	1000	QBIC	0.0150	0.1813	0.1818	0.0081	0.2922	0.2922	-0.0015	0.1866	0.1866
gMund	1000	QBICL	0.0308	0.1889	0.1913	0.0646	0.3073	0.3138	0.0344	0.2085	0.2112
gMund	1000	BIC	0.0140	0.1840	0.1844	0.0065	0.2914	0.2913	-0.0019	0.1854	0.1853
gMund	1000	BICL	0.0473	0.2013	0.2066	0.0132	0.3000	0.3001	0.0679	0.2284	0.2382
gMund	1000	AIC1	0.0144	0.1819	0.1824	0.0069	0.2923	0.2923	-0.0041	0.1848	0.1848
gMund	1000	AIC2	0.0164	0.1817	0.1823	0.0072	0.2953	0.2952	-0.0046	0.1870	0.1869
gCham	300	QBIC	0.0467	0.3172	0.3204	0.0185	0.5172	0.5173	0.0130	0.3223	0.3224
gCham	300	QBICL	0.2552	0.4229	0.4937	0.3702	0.6165	0.7188	0.0782	0.3827	0.3904
gCham	300	BIC	0.0488	0.3241	0.3276	0.0067	0.5211	0.5209	0.0153	0.3261	0.3263
gCham	300	BICL	0.3876	0.4361	0.5833	0.0823	0.5595	0.5652	0.1005	0.4033	0.4154
gCham	300	AIC1	0.0268	0.3250	0.3259	-0.0023	0.5236	0.5233	0.0022	0.3292	0.3291
gCham	300	AIC2	0.0289	0.3246	0.3257	-0.0070	0.5261	0.5259	0.0048	0.3278	0.3277
gCham	1000	QBIC	0.0133	0.1727	0.1732	0.0266	0.2898	0.2908	0.0003	0.1708	0.1708
gCham	1000	QBICL	0.0269	0.1707	0.1727	0.0450	0.2962	0.2995	0.0240	0.1743	0.1759
gCham	1000	BIC	0.0134	0.1722	0.1726	0.0230	0.2964	0.2972	-0.0007	0.1683	0.1682
gCham	1000	BICL	0.0294	0.1744	0.1768	0.0342	0.2886	0.2905	0.0420	0.1802	0.1849
gCham	1000	AIC1	0.0102	0.1821	0.1823	0.0232	0.3005	0.3012	0.0010	0.1768	0.1767
gCham	1000	AIC2	0.0084	0.1818	0.1819	0.0226	0.3061	0.3068	0.0006	0.1748	0.1747

Table A.12 Birthweight, pooled quantile regression, all moms, (unit: grams)

	Quantile				
	0.1	0.25	0.5	0.75	0.9
Smoke	-258.82	-250.75	-238.49	-234.27	-227.57
	(6.57)	(4.38)	(3.79)	(4.21)	(5.09)
Male	93.59	114.26	131.27	145.08	157.47
	(3.52)	(2.38)	(2.10)	(2.35)	(3.05)
Age	18.11	7.28	2.59	-0.10	-2.81
	(3.73)	(2.41)	(2.10)	(2.38)	(2.89)
Age^2	-0.34	-0.13	-0.04	0.02	0.09
	(0.06)	(0.04)	(0.04)	(0.04)	(0.05)
Kessner index = 2	-157.10	-108.39	-81.71	-66.64	-63.02
	(8.50)	(5.39)	(4.56)	(4.84)	(6.22)
Kessner index = 3	-297.62	-212.68	-149.85	-120.34	-91.31
	(24.05)	(15.39)	(12.48)	(11.17)	(17.79)
No prenatal visit	-118.36	0.77	7.87	30.44	25.41
	(40.77)	(24.07)	(21.29)	(17.82)	(29.10)
First prenatal visit in 2nd trimester	139.51	93.27	72.21	60.64	57.45
	(10.11)	(6.32)	(5.40)	(5.89)	(7.52)
First prenatal visit in 3rd trimester	282.43	194.33	119.48	89.75	56.94
	(27.28)	(17.32)	(14.34)	(14.15)	(20.60)

Table A.13 Birthweight, quantile regression with Classical CRE, all moms, (unit: grams)

	Quantile				
	0.1	0.25	0.5	0.75	0.9
Smoke	-144.27	-145.56	-147.59	-147.18	-147.37
	(10.43)	(7.51)	(6.62)	(7.56)	(9.87)
Male	98.49	122.15	139.71	153.22	162.48
	(4.59)	(3.30)	(2.91)	(3.32)	(4.34)
Age	-15.43	-20.22	-22.62	-17.66	-13.34
	(8.69)	(6.26)	(5.51)	(6.30)	(8.22)
Age^2	0.30	0.35	0.41	0.32	0.34
	(0.12)	(0.08)	(0.07)	(0.08)	(0.11)
Kessner index = 2	-139.90	-88.61	-60.36	-57.38	-50.37
	(10.29)	(7.06)	(6.22)	(7.11)	(9.28)
Kessner index = 3	-257.35	-173.21	-126.93	-93.61	-65.79
	(22.65)	(16.31)	(14.36)	(16.42)	(21.42)
No prenatal visit	-133.51	-31.59	3.45	14.55	57.79
	(36.44)	(26.23)	(23.10)	(26.41)	(34.46)
First prenatal visit in 2nd trimester	109.02	66.49	43.00	40.34	39.56
	(11.60)	(8.35)	(7.36)	(8.41)	(10.97)
First prenatal visit in 3rd trimester	209.63	137.03	84.09	60.87	55.49
	(27.53)	(19.82)	(17.46)	(19.96)	(26.04)

BIBLIOGRAPHY

BIBLIOGRAPHY

Abrevaya, Jason. 2006. Estimating the effect of smoking on birth outcomes using a matched panel data approach. *Journal of Applied Econometrics* 21(4). 489–519. doi:10.1002/jae.851.

Abrevaya, Jason & Christian M Dahl. 2008. The Effects of Birth Inputs on Birthweight. *Journal of Business & Economic Statistics* 26(4). 379–397. doi:10.1198/073500107000000269. http://dx.doi.org/10.1198/073500107000000269.

Amemiya, T. 1985. *Advanced econometrics*. Havard University Press.

Amemiya, Takeshi. 1978. The Estimation of a Simultaneous Equation Generalized Probit Model. *Econometrica* 46(5). 1193–1205. doi:10.2307/1911443.

Amemiya, Takeshi. 1979. The estimation of a simultaneous-eqution Tobit model. *International Economic Review* 20(1). 169–181.

Belloni, Alexandre, D Chen, Victor Chernozhukov & Christian Hansen. 2012a. Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain. *Econometrica* 80(6). 2369–2429. doi:10.3982/ECTA9626.

Belloni, Alexandre, V Chernozhukov & C Hansen. 2012b. Inference for high-dimensional sparse econometric models. *arXiv:1201.0220v1* (June 2010). 1–41. doi:10.1017/CBO9781139060035.008. http://arxiv.org/abs/1201.0220$\delimiter"026E30F$npapers3://publication/uuid/9974B31C-5AB7-4852-AA3C-B6A2D25D3901.

Belloni, Alexandre & Victor Chernozhukov. 2011a. Ch.3 High Dimensional Sparse Econometric Models:. In Pierre Alquier, Eric Gautier & Gilles Stoltz (eds.), *An introduction, inverse problems and high-dimensional estimation (lecture notes in statistics volume 203)*, 121–156. Springer-Verlag. doi:10.1007/978-3-642-19989-9.

Belloni, Alexandre & Victor Chernozhukov. 2011b. l1-Penalized Quantile Regression in High-Dimensional Sparse Models. *The Annals of Statistics* 39(1). 82–130. doi:10.1214/10-AOS827. http://projecteuclid.org/euclid.aos/1291388370.

Belloni, Alexandre, Victor Chernozhukov & Christian Hansen. 2014a. Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies* 81(2). 608–650. doi:10.1093/restud/rdt044.

Belloni, Alexandre, Victor Chernozhukov, Christian Hansen & Damian Kozbur. 2014b. Inference in High Dimensional Panel Models with an Application to Gun Control. *arXiv:1411.6507v1* (April 2013). 1–80.

Breusch, Trevor, Hailong Qian, Peter Schmidt & Donald Wyhowski. 1999. Redundancy of moment conditions. *Journal of Econometrics* 91(1). 89–111. doi:10.1016/S0304-4076(98)00050-5.

Bunea, Florentina, Alexandre Tsybakov & Marten Wegkamp. 2007. Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics* 1. 169–194. doi:10.1214/07-EJS008. http://eprints.pascal-network.org/archive/00003861/.

Canay, Ivan a. 2011. A simple approach to quantile regression for panel data. *Econometrics Journal* 14(3). 368–386. doi:10.1111/j.1368-423X.2011.00349.x.

Chamberlain, Gary. 1980. Analysis of Covariance with Qualitative Data. *Review of Economic Studies* 47(146). 225. doi:10.3386/w0325.

Chamberlain, Gary. 1984. Panel Data. In Zvi Griliches & Michael D. Intriligator (eds.), *Handbook of econometrics*, chap. 22, 1247–1318. Amsterdam: North Holland.

Chernozhukov, Victor, Ivan Fernandez-Val, Jinyong Hahn & Whitney Newey. 2009. Average and Quantile Effects in Nonseparable Panel Models. *Econometrica* 81(2). 71. doi:10.3982/ECTA8405. `http://arxiv.org/abs/0904.1990`.

Crepon, Bruno, Francis Kramarz & Alain Trognon. 1997. Parameters of Interest, Nuisance Parameters and Orthogonality Conditions: An Application to Autoregressive Error Component Models. *Journal of Econometrics* 82(1). 135–156. doi:10.1016/S0304-4076(97)00054-7.

Fan, Jianqing & Runze Li. 2001. Variable Selection via Nonconcave Penalized. *Journal of American Statistical Association* 96(456). 1348–1360.

Fan, Jianqing & Jinchi Lv. 2010. A Selective Overview of Variable Selection in High Dimensional Feature Space. *Statistica Sinica* 18(11). 1492–1501. doi:10.1016/j.str.2010.08.012.Structure.

Fan, Jianqing & Jinchi Lv. 2011. Nonconcave penalized likelihood with NP-dimensionality. *IEEE Transactions on Information Theory* 57(8). 5467–5484. doi:10.1109/TIT.2011.2158486.

Fan, Yingying & Cheng Yong Tang. 2013. Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B* 75(3). 531–552.

Friedberg, Stephen H., Arnold J. Insel & Lawrence E. Spence. 2003. *Linear algebra*. Prentice Hall 4th edn.

Gouriéroux, Christian, Alain Monfort & Alain Trognon. 1984. Pseudo maximum likelihood methods: theory. *Econometrica* 52(3).

Graham, Bryan, Jinyong Hahn & James Powell. 2009a. A Quantile Correlated Random Coefficient Panel Data Model. *manuscript* (March 2008).

Graham, Bryan S., Jinyong Hahn & James L. Powell. 2009b. The incidental parameter problem in a non-differentiable panel data model. *Economics Letters* 105(2). 181–182. doi:10.1016/j.econlet.2009.07.015. `http://dx.doi.org/10.1016/j.econlet.2009.07.015`.

Harding, Matthew & Carlos Lamarche. 2016. Penalized Quantile Regression with Semiparametric Correlated Effects: An Application with Heterogeneous Preferences. *Journal of Applied Econometrics* doi:10.1002/jae.2520.

Hastie, Trevor, Robert Tibshirani & Martin Wainwright. 2015. *Statistical Learning with Sparsity*. CRC Press.

He, Xuming & Peide Shi. 1994. Convergence rate of b-spline estimators of nonparametric conditional quantile functions. *Journal of Nonparametric Statistics* 3(3-4). 299–308. doi:10.1080/10485259408832589. http://www.tandfonline.com/doi/abs/10.1080/10485259408832589.

He, Xuming & Peide Shi. 1996. Bivariate Tensor-Product B-Splines in a Partly Linear Model. *Journal of Multivariate Analysis* 58(2). 162–181. doi:10.1006/jmva.1996.0045. http://www.sciencedirect.com/science/article/pii/S0047259X96900457.

He, Xuming, Zhong Yi Zhu & Wing Kam Fung. 2002. Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika* 89(3). 579–590. doi:10.1093/biomet/89.3.579.

Horowitz, Joel L & Sokbae Lee. 2005. Nonparametric Estimation of an Additive Quantile Regression Model. *Journal of the American Statistical Association* 100(472). 1238–1249. doi:10.1198/016214505000000583. http://dx.doi.org/10.1198/016214505000000583.

Hurwicz, Leonid. 1950. *Generalization of the Concept of Identification, "Statistical Inference in Dynamic Economic Models"*. New York: John Wiley.

Jun, Sung Jae, Yoonseok Lee & Youngki Shin. 2016. Treatment Effects With Unobserved Heterogeneity: A Set Identification Approach. *Journal of Business & Economic Statistics* 34(2). 302–311. doi:10.1080/07350015.2015.1044008. http://dx.doi.org/10.1080/07350015.2015.1044008.

Kato, Kengo, Antonio F. Galvao Jr. & Gabriel V. Montes-Rojas. 2012. Asymptotics for panel quantile regression models with individual effects. *Journal of Econometrics* 170(1). 76–91. doi:10.1016/j.jeconom.2012.02.007. http://dx.doi.org/10.1016/j.jeconom.2012.02.007.

Kim, Tae-Hwan & Halbert White. 2003. Estimation, Inference, and Specification Testing for Possibly Misspecified Quantile Regression. In *Maximum likelihood estimation of misspecified models: Twenty years later, advances in econometrics, volume 17*, 107–132. Emerald Group Publishing Limited.

Kim, Yongdai & Sunghoon Kwon. 2012. Global optimality of nonconvex penalized estimators. *Biometrika* 99(2). 315–325. doi:10.1093/biomet/asr084.

Koenker, Roger. 2004. Quantile regression for longitudinal data. *Journal of Multivariate Analysis* 91(1). 74–89. doi:10.1016/j.jmva.2004.05.006.

Koenker, Roger. 2005. *Quantile regression*. Cambridge University Press. https://books.google.com/books?hl=en{&}lr={&}id=hdkt7V4NXsgC{&}oi=fnd{&}pg=PP1{&}dq=koenker+quantile+regression+{&}ots=FtvlcdE2xv{&}sig=yIkQo3GawnC7IRgvmAlbtPE3NtI.

Koenker, Roger & Gilbert Bassett Jr. 1982. Robust Tests for Heteroscedasticity Based on Regression Quantiles. *Econometrica* 50(1). 43–61. doi:10.2307/1912528. http://ideas.repec.org/a/ecm/emetrp/v50y1982i1p43-61.html.

Lamarche, Carlos. 2010. Robust penalized quantile regression estimation for panel data. *Journal of Econometrics* 157(2). 396–408. doi:10.1016/j.jeconom.2010.03.042. http://dx.doi.org/10.1016/j.jeconom.2010.03.042.

Lee, Sokbae. 2007. Endogeneity in quantile regression models: A control function approach. *Journal of Econometrics* 141(2). 1131–1158. doi:10.1016/j.jeconom.2007.01.014.

Li, Qi & Jefferey S. Racine. 2007. *Nonparametric Econometrics: theory and practice*. Princeton University Press.

Lv, Jinchi & Yingying Fan. 2009. A unified approach to model selection and sparse recovery using regularized least squares. *Annals of Statistics* 37(6 A). 3498–3528. doi:10.1214/09-AOS683.

Matzkin, Rosa. 2007. Nonparametric identification. *Handbook of Econometrics* http://www.sciencedirect.com/science/article/pii/S1573441207060734.

Mundlak, Yair. 1978. On the pooling of time series and cross section data. *Econometrica* 46(1). 69–85. doi:10.1017/CBO9781107415324.004.

Newey, Whitney K. 1987. Efficient Estimation of Limited Dependent Variable Models with Endogenous Explanatory Variables. *Journal of Econometrics* 36(3). 231–250. doi:10.1016/0304-4076(87)90001-7.

Newey, Whitney K. & Daniel McFadden. 1994. Chapter 36 Large sample estimation and hypothesis testing. doi:10.1016/S1573-4412(05)80005-4.

Noh, Hoh Suk & Byeong U Park. 2010. Sparse Varying Coefficient Models for Longitudinal Data. *Statistica Sinica* 20. 1183–1202.

Parente, Paulo & J.M.C Santos Silva. 2016. Quantile regression with clustered data. *Journal of Econometric Methods (forthcoming)* 5(1). 1–15. doi:10.1515/jem-2014-0011.

Pollard, David. 1985. New Ways to Prove Central Limit Theorems. *Econometric Theory* 1(03). 295–313. doi:10.1017/S0266466600011233.

Rivers, Douglas & Quang H. Vuong. 1988. Limited information estimators and exogeneity tests for simultaneous probit models. *Journal of Econometrics* 39(3). 347–366. doi:10.1016/0304-4076(88)90063-2.

Rosen, Adam M. 2012. Set identification via quantile restrictions in short panels. *Journal of Econometrics* 166(1). 127–137. doi:10.1016/j.jeconom.2011.06.011. http://dx.doi.org/10.1016/j.jeconom.2011.06.011.

Schumaker, Larry L. 2007. *Spline functions: basic theory*. Cambridge University Press.

Sherwood, Ben & Lan Wang. 2016. Partially linear additive quantile regression in ultra-high dimension. *Annals of Statistics* 44(1). 288–317. doi:10.1214/12-AAP876.

Tao, Pd & Lth An. 1997. Convex analysis approach to dc programming: Theory, algorithms and applications. *Acta Mathematica Vietnamica* 22(1). 289–355. http://www.math.ac.vn/publications/acta/pdf/9701289.pdf.

Tibshirani, Ryan J. 2013. The lasso problem and uniqueness. *Electronic Journal of Statistics* 7(1). 1456–1490. doi:10.1214/13-EJS815.

Turkington, Darrell A. 2013. *Generalized vectorization, cross-products, and matrix calculus*. Cambridge.

Vershynin, R. 2010. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027* http://arxiv.org/abs/1011.3027.

Wang, Lan, Yichao Wu & Runze Li. 2012. Quantile Regression for Analyzing Heterogeneity in Ultra-high Dimension. *Journal of the American Statistical Association* 107(497). 214–222. doi:10.1080/01621459.2012.656014. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3471246{&}tool=pmcentrez{&}rendertype=abstract.

Welsh, A.H. 1989. On M-Processes and M-Estimation. *The Annals of Statistics* 17(1). 337–361.

White, Halbert. 1994. *Estimation, Inference and Specification Analysis*. Cambridge University Press.

Wooldridge, Jeffrey M. 2009. Correlated random effects models with unbalanced panels. *Manuscript (version July 2009) Michigan State* http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.472.4787{&}rep=rep1{&}type=pdf.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. MIT Press 2nd edn. doi:10.1515/humr.2003.021.

Wooldridge, Jeffrey M. 2014. Quasi-maximum Likelihood Estimation and Testing for Nonlinear Models with Endogenous Explanatory Variables. *Journal of Econometrics* 182(1). 226–234. doi:10.1016/j.jeconom.2014.04.020. http://dx.doi.org/10.1016/j.jeconom.2014.04.020.

Wooldridge, Jeffrey M. & Ying Zhu. 2016. L1-Regularized Quasi-Maximum likelihood Estimation and Inference in High-Dimensional Correlated Random Effects Probit. *manuscript* .

Zhang, Cun-Hui. 2010. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics* 38(2). 894–942. doi:10.1214/09-AOS729.

Zhang, F. 2005. *The Schur complement and its applications*, vol. 4. Springer Science & Business Media. doi:10.1007/b105056.

Zou, Hui. 2006. The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association* 101(476). 1418–1429. doi:10.1198/016214506000000735.