THREE ESSAYS ON ECOMETRICS

by

Joonhwan Lee

SUBMITTED TO THE DEPARTMENT OF ECONOMICS IN PARTIAL
FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF
PHILOSOPHY IN ECONOMICS

ABSTRACT

This thesis consists of three chapters that cover separate topics in econometrics.

The first chapter demonstrates a negative result on the asymptotic sizes of subset Anderson-Rubin tests with weakly identified nuisance parameters and general covariance structure. The result of Guggenberger et al (2012) in case of homoskedasticity is shown to break down when general covariance structure is allowed. I provide a thorough simulation results to show that the break-down of the result can be observed in wide range of parameters that is plausible in empirical applications.

The second chapter propose an inference procedure on Quasi-Bayesian estimators accounting for Monte-Carlo numerical errors. Quasi-Bayesian method have been applied to numerous applications to tackle the non-convex shape arises in certain extremum estimations. The method involves drawing finite number of Monte Carlo Markov chains to make inference and thus some degree of numerical error is inevitable. This chapter quantifies the degree of numerical error arising from the finite draws and provides a method to incorporate such errors into the final inference. I show that a sufficient condition for establishing correct numerical standard errors is geometric ergodicity of the MCMC chain. It is also shown that geometric ergodicity is satisfied under Metropolis Hastings chains with quasi-posterior for the whole class of extremum estimators.

The third chapter considers fixed effects estimation and inference in nonlinear panel data models with random coefficients and endogenous regressors. The quantities of interest are estimated by cross sectional sample moments of generalized method of moments (GMM) estimators applied separately to the time series of each individual. To deal with the incidental parameter problem introduced by the noise of the within-individual estimators in short panels, we develop bias corrections. These corrections are based on higher-order asymptotic expansions of the GMM estimators and produce improved point and interval estimates in moderately long panels. Under asymptotic sequences where the cross sectional and time series dimensions of the panel pass to infinity at the same rate, the uncorrected estimators have asymptotic biases of the same order as their asymptotic standard deviations. The bias corrections remove the bias without increasing variance.

Thesis Supervisor : Victor Chernozhukov

Title : Professor of Economics
ACKNOWLEDGMENT

I would like to express the deepest appreciation to my advisers, Professor Victor Chernozhukov and Professor Anna Mikusheva. Without their guidance and support, I could not have gone through the PhD program. They have shown inexhaustible patience and generosity to my slow progress, and constantly helped me moving forward. Countless communications with them were definitely the most essential element in my thesis writing.

I would also like to show my gratitude to Professor Whitney Newey. Being a teaching assistant for him multiple times, I have witnessed the dedication to teaching and great amount of attention to every student. He has been truly an exemplary mentor. He has been always supportive and he was the very person that I sought help for when I had hard time. I also thank Ivan Fernandez-Val, who coauthored one of my thesis chapters. Co-work with him had been an enjoyable experience.

I would like to thank Gary King, the administrator of the economics department. As anybody who has been in this department would agree, his attention to detail and administrative support have been crucial in my years at MIT. My sincere thanks to my colleagues, especially Youngjun Jang who has been my closest friend at MIT.

I greatly appreciate the financial support from Korea Foundation for Advanced Studies. For five years, the foundation had provided generous financial support for my study at MIT.

Finally, I would like to thank my parents who have been always encouraging what I have been up to. Their spiritual support has been great part of my life. My special thanks to my lover and fiancee, Junhyeon Jeong, who has supported me from long distance for so long time.
Contents

ABSTRACT 2
ACKNOWLEDGMENT 3

Chapter 1. Asymptotic Sizes of Subset Anderson-Rubin Tests with Weakly Identified Nuisance Parameters and General Covariance Structure 6
1.1. Introduction 6
1.2. Linear Instrumental Regression Model and Weak IV 8
1.3. Simplification of the Model 11
1.4. Dominance of $AR(\beta_0)$ by $\chi^2(k-m_W)$ with Conditional Homoskedasticity 14
1.5. Simulation Results 17
1.6. Simulation of More General Setting 23
1.7. Conclusion 27

Bibliography 30

Chapter 2. Inference on Quasi-Bayesian Estimators Accounting for Monte-Carlo Markov Chain Numerical Errors 32
2.1. Introduction 32
2.2. Overview on Quasi-Bayesian Estimators 34
2.3. Geometric Ergodicity of MCMC Chain 38
2.4. Consistent Estimation of Variance-Covariance Matrix 42
2.5. Monte-Carlo Simulations 43
2.6. Conclusion 45

Bibliography 46
Bibliography

Chapter 3. PANEL DATA MODELS WITH NONADDITIVE UNOBSERVED HETEROGENEITY:

ESTIMATION AND INFERENCE

3.1. INTRODUCTION

3.2. MOTIVATING EXAMPLES

3.3. THE MODEL AND ESTIMATORS

3.4. ASYMPTOTIC THEORY FOR FE-GMM ESTIMATORS

3.5. BIAS CORRECTIONS

3.6. EMPIRICIAL EXAMPLE

3.7. CONCLUSION

Bibliography

APPENDIX A. NUMERICAL EXAMPLE

APPENDIX B. CONSISTENCY OF ONE-STEP AND TWO-STEP FE-GMM ESTIMATOR

APPENDIX C. ASYMPTOTIC DISTRIBUTION OF ONE-STEP AND TWO-STEP

FE-GMM ESTIMATOR

APPENDIX D. ASYMPTOTIC DISTRIBUTION OF BIAS-CORRECTED TWO-STEP

GMM

APPENDIX E. STOCHASTIC EXPANSION FOR $\tilde{\gamma} = \hat{\gamma}(\theta_0)$ AND $\tilde{\gamma}^0 = \hat{\gamma}(\theta_0)$

APPENDIX F. STOCHASTIC EXPANSION FOR $\hat{\delta}_t^W(\theta_0; \tilde{\gamma}^0)$ AND $\hat{\delta}_t(\theta_0; \tilde{\gamma}^0)$

APPENDIX G. SCORES AND DERIVATIVES

TABLES A1-A4
CHAPTER 1

ASYMPTOTIC SIZES OF SUBSET ANDERSON-RUBIN TESTS WITH
WEAKLY IDENTIFIED NUISANCE PARAMETERS AND GENERAL
COVARIANCE STRUCTURE

1.1. INTRODUCTION

Making inference on structural parameters in the linear instrumental variables (IVs) regression models has been one of the classic problems of econometrics. One of the biggest problems that many applications of the linear IV models encounter is that instruments are often weak, i.e. they are poorly correlated with the corresponding endogenous variables. Classical asymptotics has very bad finite sample behavior with weak instruments and thus the classical inference is practically unreliable. (See Stock, Wright and Yogo (2002)) Naturally, the problem of developing inference procedures that are robust to weak instruments has been one of the central questions of econometrics for the last 15 years.

There has been rich progress in constructing robust test statistics, most notably, the AR statistic by Anderson and Rubin (1949), the Lagrange multiplier (LM) statistic by Kleibergen (2002), and the conditional likelihood ratio (CLR) statistic by Moreira (2003). An important shortcoming of the above methods is that they are designed to test only the simple full vector hypothesis in the form of $H_0 : \beta = \beta_0$ where $\beta$ contains the coefficients for all the endogenous variables. Testing for a subset of parameters is not straightforward because the unrestricted structural parameters enter as additional nuisance parameters. Projection based tests are general solution for such problems but they are often very conservative especially when the number of dimensions projected out is large. When unrestricted structural parameters are strongly identified, the above test statistics can be adapted to have correct asymptotic size

The problem of testing without any assumption on the identification of unrestricted structural parameters was a long standing question. Guggenberger et al (2012) provided a partial answer to this question. They showed that with a Kronecker product structure on a certain covariance matrix, the subset AR statistic with LIML (limited information maximum likelihood) estimator plugged in has the correct asymptotic size and power improvement over projection based tests. The Kronecker product structure, however, essentially implies the conditional homoskedasticity among reduced form disturbances. Thus, the result of Guggenberger et al (2012) is not practically useful because most economic data involve high degree of heteroskedasticity or serial correlation. Also, an important question arises: Will the result of Guggenberger et al (2012) hold with general covariance structure?

This paper provides an answer to this question by documenting a counter-example. I consider a reduced model of the linear IV model with normal disturbances and perform thorough simulations. It is shown that the result of Guggenberger et al (2012) breaks down in wide range of covariance structure if the Kronecker product assumption is removed. Moreover, it is demonstrated via simulation that the projection based tests have sharp asymptotic size. The range of covariance structure where the break-down is observed in this paper, however, necessarily imply serial correlation among reduced form disturbances. Thus, the implication of this paper may shed light on the weak identification robust inference procedures with times series data.

The paper is organized as follows. Section 2 briefly discusses the model and the problem of interest. Section 3 considers a simplification of the model to make it tractable for analysis and simulations. Section 4 reiterates the result of Guggenberger et al (2012) in the simplified model and show that the Kronecker structure is actually isomorphic to the identity matrix in the context of the test statistic. Section 5 and 6 discuss the counter example to their result with general covariance structure and thorough simulation results.
1.2. **LINEAR INSTRUMENTAL REGRESSION MODEL AND WEAK IV**

Hausman (1983) wrote that an IV regression model can be represented as limited information simultaneous equations model, in which we only specify single structural equation of interest. The full structured model is

\[ y = Y\beta + W\gamma + \epsilon \]
\[ Y = Z\Pi_Y + V_Y \]
\[ W = Z\Pi_W + V_W \]

where \( y, Y, W \) are \( T \times 1, T \times m_Y, T \times m_W \) matrices that contain endogenous variables. \( W \) is consisted of solely endogenous variables, while \( Y \) may contain some exogenous variables that is of interest. \( Z \) is a \( T \times k \) matrix of instruments. We assume away any other included exogenous variable in the structural equation by regarding all the variables to be pre-multiplied by \( M_X = I_T - X(X'X)^{-1}X' \), where \( X \) is a \( T \times m_X \) matrix of exogenous variables that is not contained in \( Y \). As usual, we assume that \( Z \) is a full rank with \( k > m_Y + m_W \) to satisfy the rank condition. The hypothesis that we are interested in is

\[ H_0 : \beta = \beta_0 \text{ v.s. } H_1 : \beta \neq \beta_0. \]

With appropriate re-parametrization, we can also test a general linear restriction in this framework as well. If we have a test that have correct asymptotic size, we can construct a corresponding confidence interval of \( \beta \) by inverting the test. Under classic asymptotics when we have fixed full rank matrix of \( [\Pi_Y \Pi_W] \) and sample size \( T \) increases to infinity, we can easily establish asymptotic normality of the estimator for \( \beta \) and \( \gamma \). We can test these parameters or any function of them with conventional Wald (or t) statistics.

Testing the parameters under potential weak identification, that is when \( [\Pi_Y \Pi_W] \) is close to degenerate along some direction, is problematic because usual asymptotic approximation does not work well even with very large \( T \). Staiger and Stock (1997), among others, examine this problem by considering an alternative asymptotics where \( [\Pi_Y \Pi_W] \) are changing with
sample size $T$ with order of $\frac{1}{\sqrt{T}}$. More recent works have focused on finding a set of robust
tests that have asymptotically correct size under arbitrarily weak identification. These robust
tests are tests based on Anderson-Rubin statistic (Anderson and Rubin, 1949), conditional
likelihood ratio statistic (Moreira, 2003) and a Lagrange multiplier statistic (Kleibergen,
2002). The aforementioned statistics are known to have limiting distributions that do not
depend on nuisance parameters when testing a hypothesis that contains the whole set of
endogenous variables, in our case that would be $H_0 : \beta = \beta_0, \gamma = \gamma_0$.

Contrary to classic asymptotics, it is not straightforward to perform a test on a subset
of parameters based on weak-instrument robust statistics. This is due to the fact that
unrestricted structural parameters constitute additional nuisance parameters in the testing
problem. In our model, the hypothesis of interest is

$$H_0 : \beta = \beta_0,$$

while allowing $\gamma$ to be unrestricted. If $\gamma$ is strongly identified, the robust tests above can
be adapted by replacing $\gamma$ with $\hat{\gamma}$, which is a consistent estimator of $\gamma$. Stock and Wright
(2000) show that such modification of AR statistic in GMM setting provides a valid test.
Kleibergen (2004) extends the result to CLR and LM statistic for a linear regression model.
Guggenberger and Smith (2005) and Otsu (2006) address the similar issue in a more general
GEL (Generalized Empirical Likelihood) framework.

Without the assumption of strong identification of $\gamma$, a natural approach is to apply
projection type tests. See Dufour (1997) and Dufour and Taamouti (2005), among others.
Projection test based on AR statistic can be described as follows. Consider AR statistic for
both $\beta$ and $\gamma$, $AR(\beta, \gamma)$. For testing the hypothesis $H_0 : \beta = \beta_0$, the projection test rejects
the null when $AR(\beta_0, \gamma) > \chi^2(k)_{1-\alpha}$ for all values of $\gamma$. Thus, the corresponding test statistic
is

$$AR(\beta_0) = \min_{\gamma \in \mathbb{R}^m} AR(\beta_0, \gamma)$$

The problem of the projection approach is that it does not provide efficient test if $\gamma$ happens
to be strongly identified, i.e. it has lower power than potentially optimal tests in some sense.
One can note that level-\(\alpha\) projected AR test uses \(\chi^2(k)_{1-\alpha}\) as a critical value while the subset AR test with the assumption that \(\gamma_0\) is strongly identified uses \(\chi^2(k - m_W)_{1-\alpha}\) while having the same test statistic.

Guggenberger et al (2012) show that we can actually improve upon projection tests even with weakly identified \(\gamma\). They show that under a Kronecker product covariance of \((\epsilon, V_W)\), that is

\[
E [\text{vec}(Z_i U_i') (\text{vec}(Z_i U_i'))'] = E[U_i U_i'] \otimes E[Z_i Z_i']
\]

where \(U_i = (\epsilon_i, V_W')\), the same subset AR test statistic \(AR(\beta_0) = \min_{\gamma \in \mathbb{R}^{m_W}} AR(\beta_0, \gamma)\) has the limit distribution that is stochastically dominated by \(\chi^2(k - m_W)\). Along with the fact that the limit distribution is exactly \(\chi^2(k - m_W)\) with strong identification of \(\gamma\), one can conclude that the test based on the subset AR statistic with critical value of \(\chi^2(k - m_W)_{1-\alpha}\) has correct asymptotic size of \(\alpha\), and provides power improvements over the projected AR test.

The crucial Kronecker product assumption essentially corresponds to conditional homoskedasticity of \(U_i = (\epsilon_i, V_W')\). This can be very restrictive in many empirical applications, especially where the weak instrument robust procedures are widely used. For example, Kleibergen and Mavroeidis (2009) applied the subset AR test based on \(\chi^2(k - m_W)\) critical value for times series data to make inference on New Keynesian Phillips curve. Presumably, the data has significant auto-correlation and conditional heteroskedasticity which are common in any time series data. They applied subset tests with AR statistic based on a conjecture that the power improvement over projection tests would hold in the case of general covariance structure. However, as pointed out later by Guggenberger et al (2012), subset LM test and CLR test does not give correct asymptotic size even under the conditional homoskedasticity and the positive result holds only for subset AR test with the Kronecker product covariance assumption. A question whether the result of Guggenberger et al (2012) holds in case of general covariance structure which allows heteroskedasticity and auto-correlation remains open. If this question has a negative answer, in other words, if the stochastic domination by \(\chi^2(k - m_W)\) does not hold with general covariance structure, then we are back to the lower
power of projection tests unless we are willing to accept the very restrictive assumption of conditional homoskedasticitity.

This paper tries to address the question by providing an counter-example and show that the stochastic domination by \( \chi^2(k - m_W) \) breaks down in wide range of non-Kronecker covariance structure. Also, a thorough Monte-Carlo simulation experiment is done to examine the region of parameters that causes break-down of Guggenberger et al (2012)'s result.

1.3. SIMPLIFICATION OF THE MODEL

Here, I analyze the reduced form in case of fixed instruments, normal errors and a known covariance matrix. The model is canonical in the literature for several reasons. First, it provides a benchmark that allows simple exposition and finite sample analysis of the statistic of interest. Second, it is a ground for asymptotics of more general models. See Moreira(2003, 2009), Andrews, Moreira and Stock(2006) and Guggenberger et al(2012) among others. Since the purpose of the paper is to provide an counter-example along with a thorough Monte-Carlo study, the notations and definitions of the following benchmark model will be used in the remaining parts.

We can rewrite the model in reduced form as

\[
\begin{pmatrix}
    y \\
    Y \\
    W
\end{pmatrix}
= \begin{pmatrix}
    Z \Pi_1 + U \\
    Z \Pi_Y + V_Y \\
    Z \Pi_W + V_W
\end{pmatrix},
\]

where \( U = \epsilon + V_Y \beta_0 + V_W \gamma_0 \) and \( \Pi_1 = \Pi_Y \beta_0 + \Pi_W \gamma_0 \). Since error terms are normal and instruments are fixed, the model can be reduced to

\[
\begin{pmatrix}
    \tilde{\Pi}_1 \\
    vec(\tilde{\Pi}_Y) \\
    vec(\tilde{\Pi}_W)
\end{pmatrix}
\sim N \left( \begin{pmatrix} \Pi_Y \beta_0 + \Pi_W \gamma_0 \\ vec(\Pi_Y) \\ vec(\Pi_W) \end{pmatrix}, \Sigma \right),
\]

where

\[
\Sigma = \left( I_{(1+m_Y+m_W)} \otimes (Z'Z)^{-1} \right) \text{Var} \left( vec(Z'(U V_Y V_W)) \right) \left( I_{(1+m_Y+m_W)} \otimes (Z'Z)^{-1} \right)
\]
and \((\hat{\Pi}_1, \hat{\Pi}_Y, \hat{\Pi}_W)\) are OLS estimator of \((\Pi_1, \Pi_Y, \Pi_W)\). Under the null hypothesis of \(H_0: \beta = \beta_0\), we can further concentrate the model by incorporating the information of true \(\beta\). We have

\[
\begin{pmatrix}
\hat{\Pi}_1 - \hat{\Pi}_Y \beta_0 \\
\text{vec}(\hat{\Pi}_W)
\end{pmatrix}
\sim N
\begin{pmatrix}
\Pi_W \gamma_0 \\
\Pi_W
\end{pmatrix},
\tilde{\Sigma}.
\]

Note that

\[
\tilde{\Sigma} = (I_{(1+m_w)} \otimes (Z'Z)^{-1}) \text{Var} \left( \text{vec} \left( Z'(\bar{U} V_W) \right) \right) (I_{(1+m_w)} \otimes (Z'Z)^{-1}),
\]

where \(\bar{U} = \epsilon + V_W \gamma\). Thus, there is no need to specify the covariance structure between \(V_Y\) and other stochastic terms to analyze this model under the null. That is exactly why Guggenberger et al (2012) did not need Kronecker product assumption for terms involving \(V_Y\). Note also that the reduced form covariance matrix \(\tilde{\Sigma}\) can be consistently estimable and thus we can treat \(\tilde{\Sigma}\) as known in asymptotic analysis of the model. Here, unknown parameters are \(\Pi_W\) and \(\gamma_0\).

The statistic of interest is the subset AR test statistic defined as

\[
AR(\beta_0) = \min_{\gamma} \min_{\Pi_W} \left( \begin{pmatrix}
\xi - \Pi_W \gamma \\
\text{vec}(\hat{\Pi}_W) - \text{vec}(\Pi_W)
\end{pmatrix}, \tilde{\Sigma}^{-1} \begin{pmatrix}
\xi - \Pi_W \gamma \\
\text{vec}(\hat{\Pi}_W) - \text{vec}(\Pi_W)
\end{pmatrix} \right),
\]

where \(\xi = \hat{\Pi}_1 - \hat{\Pi}_Y \beta_0\).

We can decompose \(\tilde{\Sigma}^{-1} = Q'Q\) where \(Q\) can be represented as

\[
Q = \begin{pmatrix}
Q_{11} & Q_{12} \\
0 & Q_{22}
\end{pmatrix}.
\]

Such decomposition includes Cholesky decomposition, which makes \(Q\) an upper-diagonal matrix with positive diagonal entries. Then we can rewrite the statistic as

\[
AR(\beta_0) = \min_{\gamma} \min_{\Phi} \left( \begin{pmatrix}
\hat{\Phi} - H(\gamma) \text{vec}(\Phi) \\
\text{vec}(\hat{\Phi}) - \text{vec}(\Phi)
\end{pmatrix}, \begin{pmatrix}
\hat{\Phi} - H(\gamma) \text{vec}(\Phi) \\
\text{vec}(\hat{\Phi}) - \text{vec}(\Phi)
\end{pmatrix} \right),
\]
where

\[ \tilde{\xi} = Q_{11}\xi + Q_{12}\text{vec}(\hat{\Pi}_w); \]

\[ \tilde{\phi} = Q_{22}\text{vec}(\hat{\Pi}_w); \]

\[ \Phi = Q_{22}\text{vec}(\Pi_w); \]

\[ H(\gamma) = (\gamma' \otimes Q_{11})Q_{22}^{-1} + Q_{12}Q_{22}^{-1}, \]

and

\[ \left( \begin{array}{c} \tilde{\xi} \\ \text{vec}(\tilde{\phi}) \end{array} \right) \sim N \left( \left( \begin{array}{c} H(\gamma_0)\text{vec}(\Phi) \\ \text{vec}(\Phi) \end{array} \right), I_{k(1+m_w)} \right). \]

The parameter \( \Phi \) can be straightforwardly concentrated out because the statistic is a quadratic function of \( \text{vec}(\Phi) \) given value of \( \gamma \). The first order condition with respect to \( \text{vec}(\Phi) \) is,

\[ 2H(\gamma)'H(\gamma)\text{vec}(\Phi) - 2H(\gamma)'\tilde{\xi} - 2(\text{vec}(\tilde{\phi}) - \text{vec}(\Phi)) = 0, \]

which gives,

\[ \text{vec}(\Phi^*) = (H(\gamma)'H(\gamma) + I_{kmw})^{-1} \left( H(\gamma)'\tilde{\xi} + \text{vec}(\tilde{\phi}) \right). \]

Plugging in the optimal value of \( \Phi \), we have

\[ AR(\beta_0) = \min_{\gamma} - \left( H(\gamma)'\tilde{\xi} + \text{vec}(\tilde{\phi}) \right)' \left( H(\gamma)'H(\gamma) + I_{kmw} \right)^{-1} \left( H(\gamma)'\tilde{\xi} + \text{vec}(\tilde{\phi}) \right) + \| \text{vec}(\tilde{\phi}) \|^2 + \| \tilde{\xi} \|^2 \]

\[ = \min_{\gamma} \left( \tilde{\xi} - H(\gamma)\text{vec}(\tilde{\phi}) \right)' \left( H(\gamma)H(\gamma)' + I_{k} \right)^{-1} \left( \tilde{\xi} - H(\gamma)\text{vec}(\tilde{\phi}) \right), \]

where the second equality uses the following decomposition of \( I_{k(mw+1)} \),

\[ I_{k(mw+1)} = \begin{pmatrix} H(\gamma)' & I_{kmw} \\ I_k & -H(\gamma) \end{pmatrix}' \begin{pmatrix} (H(\gamma)'H(\gamma) + I_{kmw})^{-1} & 0_{kmw} \\ 0_k & (H(\gamma)H(\gamma)' + I_k)^{-1} \end{pmatrix} \begin{pmatrix} H(\gamma)' & I_{kmw} \\ I_k & -H(\gamma) \end{pmatrix}. \]

\[ ^1 \text{Note that for any } m \times n \text{ matrix } H, \text{ the following holds.} \]

\[ (I_n + H'H)^{-1} = I_n - H'(I_m + HH')^{-1}H. \]
1.4. Dominance of $AR(\beta_0)$ by $\chi^2(k - m_W)$ with Conditional Homoskedasticity

Here I briefly discuss the result and proof of Guggenberger et al (2012) using the concentration and simplification I developed above for the linear IV model with conditional homoskedasticity. Given the Kronecker product assumption (i.e. conditional homoskedasticity), we have

$$\tilde{\Sigma} = \Omega \otimes (Z'Z)^{-1},$$

where $\Omega = Var((U_i, V_{W,i})'$. Let $P'P = \Omega^{-1}$ be the Cholesky decomposition of $\Omega^{-1}$. Consider the following decomposition

$$\tilde{\Sigma}^{-1} = (P \otimes (Z'Z)^{1/2})(P \otimes (Z'Z)^{1/2}).$$

Also, let

$$P = \begin{pmatrix} P_{11} & P_{12} \\ 0 & P_{22} \end{pmatrix},$$

where $P_{11}$ is a scalar, $P_{22}$ is a $m_W \times m_W$ matrix. One can note that in this case,

$$H(\gamma) = (\gamma'P_{11}^{-1}P_{22}^{-1} + P_{12}P_{22}^{-1}) \otimes I_k \equiv \tilde{\gamma} \otimes I_k,$$

$$H(\gamma)vec(\Phi) = \Phi \tilde{\gamma},$$

with a re-parametrization. The $AR$ statistic can be written as

$$AR(\beta_0) = \min_{\tilde{\gamma}} \frac{\|\tilde{\xi} - \Phi \tilde{\gamma}\|^2}{1 + \tilde{\gamma}'\tilde{\gamma}}.$$

This shows the Kronecker product assumption basically makes the model equivalent to a model from $\tilde{\Sigma} = I_{k(1+m_W)}$ with different set of parameters. That is, the class of model $\{(\gamma_0, \Pi, \tilde{\Sigma})|\tilde{\Sigma} = \Omega \otimes (Z'Z)^{-1}\}$ is equivalent to the class of model $\{\tilde{\gamma}_0, \tilde{\Phi}, I_{k(m_W+1)}\}$ and we can can achieve significant reduction in dimensionality by assuming the Kronecker product
structure. Now let us define

$$\eta_1 = \tilde{\xi} - \Phi\tilde{\gamma}_0,$$

$$\eta_2 = \tilde{\Phi} - \Phi.$$

The proof of the statement that the statistic $AR(\beta_0)$ is dominated by $\chi^2(k - m_W)$ in Guggenberger et al (2012), hinges on the following equivalence,

$$\min_{\gamma} \frac{||\tilde{\xi} - \tilde{\Phi}\gamma||^2}{1 + \gamma'\gamma} = \min_{d_1, d_2} \frac{||\epsilon_1d_1 + \tilde{\Phi}d_2||^2}{d_1^2 + d_2^2} \text{ s.t. } d_1^2 + d_2^2 = C,$$

where

$$\epsilon_1 = \frac{1}{\sqrt{1 + \gamma_0'\gamma_0}}(\eta_1 - \eta_2\gamma_0),$$

$$\epsilon_2 = \frac{1}{\sqrt{1 + \gamma_0'\gamma_0}}(\eta_1\gamma_0' + \eta_2),$$

$$\tilde{\Phi} = \sqrt{1 + \gamma_0'\gamma_0}\Phi + \epsilon_2,$$

and $C$ is any positive number. By plugging in

$$d_1^* = 1, \quad d_2^* = -(\tilde{\Phi}'\tilde{\Phi})^{-1}\tilde{\Phi}'\epsilon_1,$$

we have

$$\min_{\gamma} \frac{||\tilde{\xi} - \tilde{\Phi}\gamma||^2}{1 + \gamma'\gamma} \leq \frac{\epsilon_1' \left(I_k - \tilde{\Phi}(\tilde{\Phi}'\tilde{\Phi})^{-1}\tilde{\Phi}'\right)}{1 + d_2^*d_2^*} \leq \epsilon_1' \left(I_k - \tilde{\Phi}(\tilde{\Phi}'\tilde{\Phi})^{-1}\tilde{\Phi}'\right) \sim \chi^2(k - m_W),$$

since $\epsilon_1$ and $\epsilon_2$ are independent. Obviously, all the elements above is not feasible but that is not of concern here because we basically use them to show that there exists $\gamma^*$ for every realization of $\eta_1$ and $\eta_2$ such that the criterion function evaluated at $\gamma^*$ be dominated by $\chi^2(k - m_W)$. That is sufficient to show that $AR$ statistic is indeed dominated by $\chi^2(k - m_W)$.

**Negative Result in General Covariance Case.** The above result, however, does not hold generally with potential heteroskedasticity and auto-correlation. Here, I document
a case where the above result breaks down in a case of general covariance structure. Since one counter-example is enough to prove the claim, I demonstrate such an example with simplest possible setting, i.e. with $m_W = 1$ and $k = 2$. I set parameters to be

$$\gamma_0 = 1 \quad \Phi = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad Q_1 \equiv Q_{11} Q_{22}^{-1} = I_2 \quad Q_2 \equiv Q_{12} Q_{22}^{-1} = 4 \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$ 

Note that the parameter setup seems to be not much a departure from the homoskedastic case except that $Q_{12} Q_{22}^{-1}$ is not a scalar multiple of $I_2$. Also, the value of $\Phi$ indicates that the strength of identification is very weak. As shown later in more thorough simulation experiments, the upward departure from $\chi^2(k - m_W)$ is more pronounced as $\Phi$ and $\gamma_0$ get closer to zero. I ran a Monte-Carlo simulation based on the concentrated and reduced model

$$\begin{pmatrix} \bar{\xi} \\ \bar{\Phi} \end{pmatrix} \sim N \left( \begin{pmatrix} H(\gamma_0) \Phi \\ \Phi \end{pmatrix}, I_{2k} \right),$$

and examined the distribution of $AR$ statistic

$$AR(\beta_0) = \min_\gamma \left( \bar{\xi} - H(\gamma) \hat{\Phi} \right)' \left( H(\gamma) H(\gamma)' + I_k \right)^{-1} \left( \bar{\xi} - H(\gamma) \hat{\Phi} \right).$$

Projection method guarantee that $AR(\beta_0)$ is dominated by $\chi^2(2)$, and the question we wish to address is whether $AR(\beta_0)$ is dominated by $\chi^2(1)$. Although the minimization involved is over just a single dimensional space, the criterion function has potentially many humps and thus guaranteeing global minimization for every simulation draw is not an easy task. I employed a Newton type algorithm with multiple starting values to do the task. Figure 1 shows the quantile function of $AR(\beta_0)$ along with that of $\chi^2(1)$ and $\chi^2(2)$. As we can see from Figure 1, there is a clear stochastic ordering in this case. The statistic $AR(\beta_0)$ is stochastically larger than $\chi^2(1)$ and thus using $\chi^2(1)$ critical values would clearly produce over rejection, i.e. size distortion.
100,000 simulation draws. Parameters are set to $\gamma_0 = 1$, $\Phi = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, $Q_1 = I_2$, $Q_2 = 4 \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$.

This case is with $k = 2$ and $m_W = 1$.

### 1.5. Simulation Results

In this section, I report some Monte-Carlo simulation results with varying model parameters $\gamma_0$, $\Phi$, $Q_1$ and $Q_2$ as well as the number of instruments $k$. For most part, I look into the case of $k = 2$ for simplicity of specification. As in the previous example, I set $Q_1 = I_2$ and consider the combinations of the following set of parameters.

- $\gamma_0 \in \{0, 0.1, 0.5, 1, 5, 20\}$,
- $\Phi \in \{\lambda \mu | \mu = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \lambda = 0, 0.1, 0.5, 1, 5, 20\}$,
- $Q_2 \in \{\eta R | R = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \eta = 0, 1, 4, 10\}$. 

![Quantile Plot of AR, $\chi^2(1)$ and $\chi^2(2)$](image-url)
Due to the nature of the $AR(\beta_0)$, the distribution of it is expected to be continuous in the change of the above parameters. The criterion function is a continuous function in those parameters. Since minimum is a continuous functional, one can argue that the distribution of $AR(\beta_0)$ and its functionals are also continuous in all the parameters. ²

Before examining the simulation result, it may be helpful to see what the above parameter represent in terms of the original model of interest. The reduced form is written

$$
\begin{pmatrix}
y \\
Y \\
W
\end{pmatrix} = 
\begin{pmatrix}
Z\Pi_1 + U \\
Z\Pi_Y + V_Y \\
Z\Pi_W + V_W
\end{pmatrix},
$$

and in case of the model we consider for the simulation, where $k = 2$ and $m_W = 1$, under the null hypothesis of $H_0 : \beta = \beta_0$, the model is reduced to

$$
\begin{pmatrix}
\hat{\Pi}_1 - \hat{\Pi}_Y \beta_0 \\
\hat{\Pi}_W
\end{pmatrix} \sim N \left( \begin{pmatrix} \Pi_W \gamma_0 \\ \Pi_W \end{pmatrix}, \bar{\Sigma} \right).
$$

If we assume that $Z$ is normalized so that $Z'Z = I_k$, we have

$$
\bar{\Sigma} = \text{Var} \left( \text{vec} \left( Z' (\bar{U} V_W) \right) \right),
$$

where $\bar{U} = U - \beta_0 V_Y$.

We can see how $Q_1$ and $Q_2$ are related to $\bar{\Sigma}$, the reduced form covariance as follows. Since the model is a reduced version, there are infinitely many $\bar{\Sigma}$ that corresponds to the same value of $Q_1$ and $Q_2$. If we normalize so that $\bar{\Sigma}_{22} = I_2$, then we have

$$
\bar{\Sigma} = \begin{pmatrix}
(Q_1' (I + Q_2 Q_2')^{-1} Q_1)^{-1} & -(Q_1' Q_1)^{-1} Q_1' Q_2 \\
-Q_2' Q_1 (Q_1' Q_1)^{-1} & I_2
\end{pmatrix} = \begin{pmatrix}
(1 + \eta^2) I_2 & -\eta R \\
-\eta R' & I_2
\end{pmatrix}.
$$

Note that when $Q_2$ is zero matrix, that is when $\eta = 0$, then we have conditional homoskedasticity. Larger values of $\eta$ means that the reduced form error of $Z'\bar{U}$ has much larger variance.

²Note that the minima $\gamma^*$ is not necessarily continuous in parameters.
(at the order of $\eta$) than that of $Z'V_W$. Also, larger values of $\eta$ translate into higher correlation, either negative or positive, between $Z_1'U$ and $Z_2'V_W$, and $Z_2'U$ and $Z_1'V_W$ because the correlation coefficients are

$$\pm \frac{\eta}{\sqrt{1 + \eta^2}}.$$ 

Given the value of $R$ in the simulation, we have opposite signs in those two correlations. The matrix $R$ is chosen in this way because it produced the most amount of upward deviation from $\chi^2(k - m_W)$. However, a slight change in the value of $R$ also generated such deviation and therefore the deviation is not a singularity, but occurs over a wide range of set of parameters.

One can ask whether the structure of $\Sigma$ that generates the deviation is plausible or realistic in empirical application. This question will be addressed more thoroughly in the later sections.

Vector parameter $\Phi$ indicates the strength of identification because under conditional homoskedasticity with $\tilde{\Sigma} = \begin{pmatrix} \sigma_u^2 & \sigma_{uw} \\ \sigma_{uw} & \sigma_v^2 \end{pmatrix} \otimes (Z'Z)^{-1}$, we have

$$\Phi = Q_{22} \Pi_W = \frac{1}{\sigma_v} (Z'Z)^{\frac{1}{2}} \Pi_W,$$

$$||\Phi||^2 = \Pi_W Z'Z \Pi_W / \sigma_v^2,$$

i.e., the norm of $\Phi$ is the familiar concentration parameter for $W$ under the assumption that $H_0 : \beta = \beta_0$ holds. See Stock, Wright and Yogo (2002) for the definition. In general case, we can note that

$$||\Phi||^2 = \Pi_W \tilde{\Sigma}_{22}^{-1} \Pi_W,$$

where $\tilde{\Sigma}_{22}$ is the variance-covariance matrix of $vec(\tilde{\Phi})$. In case of $\Phi = 20\mu$ or $\Phi = 5\mu$ in this experiment setting, they correspond to the concentration parameter of 400 and 25, which are regarded as strong identification in most empirical literature in case of $k = 2$. The value of $\gamma_0$ also presumably affects the distribution of $AR(\beta_0)$. For conditional homoskedastic case, I showed that a model with $(\gamma, \Pi, \tilde{\Sigma})$ is equivalent to $(\tilde{\gamma}, \Phi, I)$ for some values of $\tilde{\gamma}$ and $\Phi$ in terms of the $AR(\beta_0)$ distribution. As shown above, the value of $\tilde{\gamma}$ is affected by the
elements of $\Sigma$. The simulation results clearly indicates that the value of $\gamma_0$ indeed affects the distribution of $AR(\beta_0)$ in a subtle manner.

The number of simulation draw is 50,000 for each combination of parameters and I tabulate the 95% and 90% quantile values of empirical distribution of the simulated $AR(\beta_0)$, denoted by $AR_{95}$ and $AR_{90}$ respectively, along with the standard errors. Also, I report $P(AR(\beta_0) > \chi^2(1)_{1-\alpha})$ for $\alpha = 0.05$ and 0.1, which is the true size of the test based on $\chi^2(1)_{1-\alpha}$ critical values. Those values will show the size distortion when we apply the subset test in Guggenberger et al (2012) to a described model without the assumption of Kronecker product. Through the simulation experiment, I found that the behavior of $AR_{95}$ or $AR_{90}$ is a good representation of the behavior of the whole distribution of $AR(\beta_0)$. That is, at least in this experiment, when $AR_{95}$ is above $\chi^2(1)_{0.95}$, then the distribution of $AR(\beta_0)$ stochastically dominates $\chi^2(1)$. Also, when $AR_{95}$ converges to $\chi^2(1)_{0.95}$ in some change of parameter, then the distribution of $AR(\beta_0)$ is also found to converge to $\chi^2(1)$. Therefore, the simulation results can be interpreted accordingly.

The simulation results in Table 1 and Table 2 show some notable tendencies. First, as $\Phi$ is further apart from zero, the distribution of $AR(\beta_0)$ converge to $\chi^2(1)$. This has a natural explanation that $||\Phi||^2$ is the concentration parameter and larger value of it indicates strong identification. This is consistent with the result of Stock and Wright (2000) that we have the subset test statistic following $\chi^2(k - m\nu)$ with strongly identified $\gamma_0$. However, the speed of convergence varies for different values of $\gamma$ and $Q_2$. Most notably, higher values of $\gamma_0$ seems to be highly correlated with the speed of convergence. Figure 2 shows how $AR_{95}$ changes in the identification strength of $\gamma_0$. In homoskedastic case where $Q_2 = 0$, we can observe monotonic increase of $AR_{95}$ until it converges to $\chi^2(k - 1)_{0.95}$. In general covariance case where $Q_2 = 4R$, we can observe $AR_{95}$ is well above $\chi^2(1)_{0.95}$ with weaker identification and then decreases way below $\chi^2(1)_{0.95}$, and finally increase monotonically while converging. In both cases, the speed of convergence is faster when the value of $\gamma_0$ is larger. We can also note that the speed of convergence is much slower in case of $Q_2 = 4R$. 
### Table 1. Simulation Results of AR95 for $k = 2, m_W = 1$

95% quantile of $\chi^2(1)$ is 3.84 and that of $\chi^2(2)$ is 5.99. The result is from 50,000 simulation draws for each configuration.

Break-down of the dominance by $\chi^2(k - m_W)$ is observed in the weak identification region of $\Phi$ and it is more pronounced when $\eta$ is larger. Figure 3 shows that $AR(\beta_0)$ can indeed get very close to $\chi^2(2)$ when $\eta$ is sufficiently large and there is no identification. This example demonstrate that at least with $k = 2$ and $m_W = 1$, the asymptotic size of the subset $AR$ test based on critical value of the projection $AR$ test sharply equals $\alpha$ if we consider every possible $\tilde{\Sigma}$, not just a class of $\tilde{\Sigma}$’s that have Kronecker product structure. In fact, in the next section, it is demonstrated that this is generally true for $k > 2$ and $m_W > 1$. These results show that applying the critical value of $\chi^2(k - m_W)_{1-\alpha}$ in empirical applications.
may generate significant size distortion when we have weakly identified η and high degree of heteroskedasticity and auto-correlation. Even in the most severe case of break-down in this setup, the upward deviation from $\chi^2(1)$ seems to disappear when $\lambda \geq 2$ which can be translated to concentration parameter of 4. This seems pretty small number but obviously one cannot use such threshold to decide whether we are safe to use the $\chi^2(k - m_W)$ critical values in empirical applications.

Table 2. Simulation Results of AR90 for $k = 2, m_W = 1$

<table>
<thead>
<tr>
<th>η = 0</th>
<th>η = 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma_0 = 0$</td>
<td>$\gamma_1 = 0.1$</td>
</tr>
<tr>
<td>$\lambda = 0$</td>
<td>1.17</td>
</tr>
<tr>
<td>(0.009)</td>
<td>(0.009)</td>
</tr>
<tr>
<td>$\lambda = 0.1$</td>
<td>1.19</td>
</tr>
<tr>
<td>(0.010)</td>
<td>(0.009)</td>
</tr>
<tr>
<td>$\lambda = 0.5$</td>
<td>1.24</td>
</tr>
<tr>
<td>(0.010)</td>
<td>(0.010)</td>
</tr>
<tr>
<td>$\lambda = 1$</td>
<td>1.43</td>
</tr>
<tr>
<td>(0.011)</td>
<td>(0.011)</td>
</tr>
<tr>
<td>$\lambda = 5$</td>
<td>2.59</td>
</tr>
<tr>
<td>(0.020)</td>
<td>(0.020)</td>
</tr>
<tr>
<td>$\lambda = 20$</td>
<td>2.71</td>
</tr>
<tr>
<td>(0.021)</td>
<td>(0.021)</td>
</tr>
</tbody>
</table>

TABLE 2. Simulation Results of AR90 for $k = 2, m_W = 1$

90% quantile of $\chi^2(1)$ is 2.71 and that of $\chi^2(2)$ is 4.61. The result is from 50,000 simulation draws for each configuration. Standard error from simulation in parenthesis.
FIGURE 1.5.1. Change of AR95 in Identification Strength
For $\eta = 0$ (Conditional homoskedastic case) and $\eta = 4$ (Non-Kronecker case). 50,000 simulation draws for each configuration.

1.6. SIMULATION OF MORE GENERAL SETTING

The previous section discussed the simulation results for $k = 2$ and $Q_2$ of scalar multiple of $R = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$. One may ask if we can observe the break-down of dominance by $\chi^2(k-m_W)$ in wider range of parameters. The answer is positive both for $Q \neq \eta R$ and $k > 2$. This section is devoted to exploration of the region of parameters that generates $AR(\beta_0)$ distribution.
Figure 1.5.2. Case of AR being close to $\chi^2(k)$

Note that the value of $\gamma_0$ does not matter here because we set $\lambda = 0$ so that $\Phi = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$.

dominating $\chi^2(k - m_W)$ and discussion of whether the region of parameter is plausible in empirical applications.

First, I consider $Q_2$ that is not a scalar multiple of $R$ in case of $k = 2$. Instead, I consider a class of $Q_2$,

$$Q_2 \subset \{ \eta R(\theta) | \eta \in \mathbb{R}, \theta \in [-\frac{\pi}{2}, \frac{\pi}{2}] \},$$

that is, $Q_2$ is a scalar multiple of a rotation matrix in $\mathbb{R}^2$. The class contains the values of $Q_2$ used in the previous section as a special case of $\theta = \frac{\pi}{2}$. Figure 4 shows how $AR95$ changes over different values of $\theta$, where other parameters are set to $\{Q_1 = I_2, \Phi = 0.1\mu, \gamma_0 = 1, \eta \in \{2, 8, 20\}\}$. Interestingly, $AR95$ is increasing to the rotation angle and it takes maximum at $\theta = \pm \frac{\pi}{2}$. We can note that when $\eta$ is sufficiently large, we can observe $AR95 \geq \chi^2(1)_{0.95}$ for wide range of $\theta$. For the case of $\eta = 2$, $|\theta| > \frac{\pi}{3}$. It should be noted that there exists $Q_2$ that produces $AR$ statistic dominating $\chi^2(1)$ outside the class considered here. The class of scalar multiple of
rotation matrices is considered just for convenience of characterization. In fact, any $Q_2$ with complex eigenvalues whose imaginary parts are sufficiently large could generate AR statistic that dominates $\chi^2(1)$. The exact mechanic of this, however, could not be clearly described analytically.

![Figure 1.6.1. Change of AR95 on $\theta$](image)

50,000 simulation draws with increment of $\pi/180$ of $\theta$.

Second, I consider the case of $k > 2$ and $k = 2l$, i.e. $k$ is even with $l$ being a positive integer. Then an obvious extension of $Q_2$ with $\eta R(\theta)$ is

$$Q_2 = I_l \otimes \eta R(\theta),$$

where $\eta = 20$, $\theta = \frac{\pi}{2}$ and and $0_k$ is a $k \times 1$ vector of zeros. The other parameters are set as

$$Q_1 = I_k, \quad \gamma_0 = 0, \quad \Phi = 0_k.$$

The Table 3 shows the simulation results for $k \in \{4, 6, 10, 30\}$. We can see that both AR95 and AR90 exceeds the value of $\chi^2(k - 1)_{0.95}$ and $\chi^2(k - 1)_{0.9}$ by a significant margin in all
<table>
<thead>
<tr>
<th>$k$</th>
<th>$\chi^2(k-1)_{0.90}$</th>
<th>$AR90$</th>
<th>$\chi^2(k)_{0.90}$</th>
<th>$\chi^2(k-1)_{0.95}$</th>
<th>$AR95$</th>
<th>$\chi^2(k)_{0.95}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>6.25 (0.023)</td>
<td>7.56</td>
<td>7.78 (0.033)</td>
<td>7.81 (0.033)</td>
<td>9.28</td>
<td>9.49 (0.067)</td>
</tr>
<tr>
<td>6</td>
<td>9.24 (0.027)</td>
<td>10.42</td>
<td>10.64 (0.037)</td>
<td>11.07 (0.037)</td>
<td>12.36</td>
<td>12.59 (0.044)</td>
</tr>
<tr>
<td>10</td>
<td>14.68 (0.033)</td>
<td>15.75</td>
<td>15.99 (0.044)</td>
<td>16.92 (0.044)</td>
<td>18.05</td>
<td>18.31 (0.067)</td>
</tr>
<tr>
<td>30</td>
<td>39.09 (0.051)</td>
<td>39.93</td>
<td>40.26 (0.067)</td>
<td>42.56 (0.067)</td>
<td>43.45</td>
<td>43.77 (0.067)</td>
</tr>
</tbody>
</table>

Table 3. $AR95$ and $AR90$ in case of $k = 2l$

Note that $Q_2$'s in these cases have quite a sparse structure when $k$ is larger, and they still generate the $AR(\beta_0)$ statistic that stochastically dominates $\chi^2(k-1)$.

For $k = 2l + 1$, it is also possible to observe a breakdown of $\chi^2(k-1)$ dominance. By setting the parameters as

$$Q_1 = \begin{pmatrix} \epsilon & 0'_{k-1} \\ 0_{k-1} & I_{k-1} \end{pmatrix}, \quad Q_2 = \begin{pmatrix} 1 & 0'_{k-1} \\ 0_{k-1} & I_{k-1} \otimes \eta R(\theta) \end{pmatrix}, \quad \gamma_0 = 0, \quad \Phi = 0_k,$$

where $\eta = 20$ and $\theta = \pi$. The table 4 shows the simulation results for $k \in \{3, 7, 11, 31\}$ with $\epsilon = 0.0001$. I could not observe a case with $Q_1$ not being close to singular matrix for $k = 2l + 1$. When $k$ is a odd number, I was not able to find any $Q_2$ that generates $AR > \chi^2(k-1)$ with $Q_1 = I_k$ among 5000 randomly generated $Q_2$'s. Such cases was only observed with $Q_1$ being near singular. $Q_1$ being above with very small $\epsilon$ implies that one of the instruments has nearly no information.

The whole region of parameters where I found the break-down of $\chi^2(k-m_w)$ dominance, however, implies some degree of auto-correlation in errors. If we only have conditional heteroskedasticity without any auto-correlation, we have

$$\tilde{\Sigma} = (I_{(1+m_w)} \otimes (Z'Z)^{-1}) Var (vec(Z\omega)) (I_{(1+m_w)} \otimes (Z'Z)^{-1})$$

$$= (I_{(1+m_w)} \otimes (Z'Z)^{-1}) E[E[\omega_i|Z_i] \otimes (Z_iZ_i')] (I_{(1+m_w)} \otimes (Z'Z)^{-1}),$$
which implies all $k \times k$ blocks that consist $\tilde{\Sigma}$ should be symmetric. It can be shown that
the parameters values where I documented the stochastic dominance of AR statistic over $\chi^2(k - m_W)$ are not feasible under the block-wise symmetry of $\tilde{\Sigma}$. This does not necessarily imply that AR statistic is dominated by $\chi^2(k - m_W)$ when we have only conditional heteroskedasticity. As Guggenberger et al(2012) noted, proving the result analytically is not an easy feat. Thus, the findings in this paper have something to say for applying subset AR test for data prone to auto-correlation, e.g. time series data. A notable application is inference on New-Keynesian Phillips curve as in Kleibergen and Mavroeidis (2009).

1.7. CONCLUSION

Reducing the degree of freedom for testing a subset of parameters with weak identification robust test statistic and weakly identified nuisance parameters has been a challenging problem in the literature. Such dimension reduction is important in practical perspective because the efficiency of a test can be substantially improved. Recent work of Guggenberger et al (2012) showed that subset Anderson-Rubin statistic can be applied with reduced degree of freedom with conditional homoskedasticity, or Kronecker product structure. This paper is first to document that the result of Guggenberger et al (2012) does not hold in general covariance structure. With a thorough simulation study, I show that the projection based tests have sharp asymptotic size and this cannot be improved without further assumptions on the

<table>
<thead>
<tr>
<th>$k$</th>
<th>$\chi^2(k-1)_{0.90}$</th>
<th>$AR90$</th>
<th>$\chi^2(k)_{0.90}$</th>
<th>$\chi^2(k-1)_{0.95}$</th>
<th>$AR95$</th>
<th>$\chi^2(k)_{0.95}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>4.61</td>
<td>5.83</td>
<td>6.25</td>
<td>5.99</td>
<td>7.37</td>
<td>7.81</td>
</tr>
<tr>
<td></td>
<td>(0.022)</td>
<td></td>
<td>(0.030)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>10.64</td>
<td>11.65</td>
<td>12.02</td>
<td>12.59</td>
<td>13.69</td>
<td>14.07</td>
</tr>
<tr>
<td></td>
<td>(0.029)</td>
<td>(0.039)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>15.99</td>
<td>16.85</td>
<td>17.28</td>
<td>18.31</td>
<td>19.21</td>
<td>19.68</td>
</tr>
<tr>
<td></td>
<td>(0.034)</td>
<td>(0.045)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>40.26</td>
<td>41.02</td>
<td>41.42</td>
<td>43.77</td>
<td>44.53</td>
<td>44.99</td>
</tr>
<tr>
<td></td>
<td>(0.052)</td>
<td>(0.067)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4. $AR95$ and $AR90$ in case of $k = 2l + 1$
covariance structure. Also, it is shown that the break down of the result is most pronounced when the identification of nuisance parameters is weak.

The region of parameters that I found the break down, however, necessarily imply there is some degree of auto-correlation in errors. This leaves an important question: can we have the dimension reduction with only conditional heteroskedasticity? The simulation results suggest that the answer might be positive. Theoretical proof may be a daunting task, but it will definitely allow subset AR test to be applied to much wider range of problems.
Bibliography


CHAPTER 2

INFERENCES ON QUASI-BAYESIAN ESTIMATORS ACCOUNTING FOR MONTE-CARLO MARKOV CHAIN NUMERICAL ERRORS

2.1. INTRODUCTION

Quasi-Bayesian estimators (QBEs) or Laplace type estimators (LTEs) are defined similarly to Bayesian estimators with replacing general statistical criterion functions for parametric likelihood function. Chernozhukov and Hong (2003) provides a detailed discussion on properties of QBEs. They show that under some primitive conditions that are necessary for asymptotic normality of general class of extremum estimators, QBEs are asymptotically equivalent to the corresponding extremum estimators. QBEs are particularly useful when the statistical criterion function of interest exhibits irregular and non-smooth behavior in finite sample. Such functional behavior makes it hard to calculate extremum point estimates since the process involves global optimization on highly irregular non-smooth function. Several prominent econometric models such as censored quantile regression model and instrumental variable quantile regression model are well known to have the problem. These models have well defined global minima in finite sample whose behavior follows standard asymptotics. Thus, one can say that the challenge of dealing with such statistical models is purely computational.

QBEs essentially transform the extremum estimation problem into a Bayesian estimation problem with a properly set Quasi-posterior. Thus, the point estimation comes down to taking mean or median of MCMC draws of the Quasi-posterior, and constructing confidence interval can be as simple as taking appropriate quantiles of the Quasi-posterior.

Theoretically, we can make the number of draws for MCMC algorithm large enough to assure certain degree of precision for the computation of Bayesian or Quasi-Bayesian estimators of interest. In practice, however, there are many computational restrictions that
makes large number of draws very costly or nearly impossible. In QBE context, the problem at hand is to obtain the value of $E[g(\theta)]$ as precise as possible, where $\theta$ is parameter of interest and the expectation is taken on Quasi-posterior. As noted in Gelman and Shirley (2011), inference on $E[g(\theta)]$ or a functional of the posterior requires considerably more MCMC draws than the problem of making Bayesian inference on $g(\theta)$. Therefore, with practical number of draws, it is plausible that there is non-trivial amount of error in the computed QBE purely due to finiteness of MCMC draws. Rarely this error is addressed in application of not only QBEs but also general Bayesian estimation using MCMC draws. For Bayesian inference, one can argue that the Monte-Carlo error does not affect much on the construction of credible interval, which often is the main object of interest. For QBEs or Bayesian point estimates, one should take close look at the Monte-Carlo error and incorporate it into the statistical inference procedure.

Similar problems have been treated in context of simulated method of moments. McFadden (1989) discusses the numerical error from using simulated moments in discrete response models. Pakes and Pollard (1989) provides asymptotic theory for general class of simulation based estimators. Simulated method of moments or its variant, however, generally assume the possibility of independent draws. For QBE problems, independent draws are generally impossible and thus we rely on Markov Chain Monte-Carlo, which makes the asymptotics substantially more complicated than independent simulation draws.

This paper provides a framework of incorporating Monte-Carlo error in Quasi-Bayesian estimation problem. First, I briefly review the statistical inference based on QBEs following Chernozhukov and Hong (2003), and quantify the Monte-Carlo error adjusted standard errors for QBEs. Then, I establish conditions for a central limit theorem that is needed to ensure asymptotic normality of QBE when the number of draws $B$ in MCMC is large. Next, I establish conditions for consistent estimation of variance-covariance matrix of Monte-Carlo errors. Finally, I provide a simple Monte-Carlo study using censored quantile regression model similar to Chernozhukov and Hong (2003) and show that incorporating Monte-Carlo error improves the coverage probability.
2.2. OVERVIEW ON QUASI-BAYESIAN ESTIMATORS

In Chernozhukov and Hong (2003), QBEs are proposed as an alternative estimation method for general extremum estimation problems. Thus, we first define the extremum estimator of interest and basic assumptions for the estimator to have consistency and asymptotic normality. Given a probability space \((\Omega, \mathcal{F}, P)\), define

\[
\hat{\theta}_n^E = \arg \max_\theta L_n(\theta)
\]

where \(L_n(\theta)\) is a measurable function of \(\theta\) on \(\Theta\), and a random variable on \(\Omega\). This framework of extremum estimation encompasses various estimation methods such as maximum likelihood estimation and generalized method of moments. The following assumptions are generally used in the literature to ensure consistency and asymptotic normality of an extremum estimator. They are primitive assumptions for identification and validity of asymptotic expansion. See Newey and McFadden (1994) or Gallant and White (1988) for further discussion.

(1) The true parameter \(\theta_0\) belongs to the interior of a compact convex subset \(\Theta \subset \mathbb{R}^d\) and \(\theta_0\) is identifiable in the sense that there exists \(\epsilon > 0\) such that

\[
\liminf_{n \to \infty} P^* \left\{ \sup_{|\theta - \theta_0| \geq \delta, n} \frac{1}{n} (L_n(\theta) - L_n(\theta_0)) \leq -\epsilon \right\} = 1
\]

for all \(\delta > 0\).

(2) For \(\theta\) in an open neighborhood of \(\theta_0\), we have

\[
L_n(\theta) - L_n(\theta_0) = (\theta - \theta_0)'\Delta_n(\theta_0) - \frac{1}{2}(\theta - \theta_0)'[nJ_n(\theta_0)](\theta - \theta_0) + R_n(\theta)
\]

where \(\Omega_n^{-1/2}(\theta_0)\Delta_n(\theta_0)/\sqrt{n} \overset{d}{\to} N(0, I_d)\) and \(J_n(\theta_0) = O(1)\) and \(\Omega_n(\theta_0) = \text{Var}(\Delta_n(\theta_0)/\sqrt{n}) = O(1)\) are uniformly in \(n\) positive definite constant matrices. For residual term \(R_n(\theta)\),
for any $\epsilon > 0$ there exist $\delta > 0$ and $M > 0$ such that

$$\limsup_{n \to \infty} P^* \left\{ \sup_{M/\sqrt{n} \leq |\theta - \theta_0| \leq n|\theta - \theta_0|^2} \frac{|R_n(\theta)|}{n|\theta - \theta_0|^2} > \epsilon \right\} < \epsilon$$

and

$$\limsup_{n \to \infty} P^* \left\{ \sup_{|\theta - \theta_0| \leq M/\sqrt{n}} |R_n(\theta)| > \epsilon \right\} = 0$$

Under the Assumption 1 and 2, we have consistency of $\hat{\theta}_n^E$ and

$$\sqrt{n} \Omega_n^{-1/2} \left( \theta_0 \right) R_n(\theta_0) \left( \hat{\theta}_n^E - \theta_0 \right) \overset{d}{\to} N \left( 0, I_d \right).$$

In theory, the extremum estimation provides a simple unified framework for many standard statistical models. However, actually computing the extremum estimates in practice is not a trivial problem. As noted by Andrews (1997), the problem can be cumbersome for some of important econometric models such as censored quantile regression models and quantile instrumental variables models.

Quasi-Bayesian Estimator (QBE) or Laplace-Type Estimator (LTE) provide an asymptotic equivalent to the extremum estimator. QBE is defined as

$$\hat{\theta}_n^Q = \arg \inf_{\zeta \in \Theta} \int_{\Theta} \rho_n(\theta - \zeta) \left\{ \frac{\exp(L_n(\theta)\pi(\theta))}{\int_{\Theta} \exp(L_n(\theta)\pi(\theta))d\theta} \right\} d\theta,$$

where $\rho_n(u)$ is a penalty function and $\pi(\theta)$ is a prior density. By construction, $\hat{\theta}_n^Q$ is a Bayesian decision based on a penalty function $\rho_n$ and quasi-posterior

$$p_n(\theta) = \frac{\exp(L_n(\theta)\pi(\theta))}{\int_{\Theta} \exp(L_n(\theta)\pi(\theta))d\theta}.$$ 

Chernozhukov and Hong (2003) considers fairly general class of $\rho_n(u)$ and states their main results accordingly. However, I mainly consider $\rho_n(u) = n|u|^2$, which makes the value of $\hat{\theta}_n^Q$ be the mean of the quasi-posterior $p_n(\theta)$. The key results of Chernozhukov and Hong (2003) (Theorem 1 and 2) state that the quasi-posterior, with appropriate normalization, converges.
to Gaussian distribution in total variation of moments norm and thus we have
\[ \sqrt{n}(\hat{\theta}_n^Q - \theta_0) = \sqrt{n}(\hat{\theta}_n^E - \theta_0) + o_p(1) \]
\[ \sqrt{n}\Omega_n^{-1/2}(\theta_0)J_n(\theta_0)(\hat{\theta}_n^Q - \theta_0) \overset{d}{\rightarrow} N(0, I_d). \]

This essentially states that we can substitute QBE for an extremum estimator. Furthermore, we can make statistical inference on \( \theta_0 \) or any smooth function of \( \theta_0 \) with consistent estimates of \( \Omega_n(\theta_0) \) and \( J_n(\theta_0) \). \( \Omega_n(\theta_0) \) can be estimated straightforwardly by standard methods, and \( J_n(\theta_0) \) can be estimated by the variance of the quasi-posterior.

Let us consider a statistical inference problem in the context of QBE. Let \( g(\theta) \) be a continuously differentiable function and \( g(\theta_0) \) be the object of interest. Asymptotic 100(1 - \( \alpha \))% confidence interval for \( \theta_0 \) can be constructed as
\[
\left[ g(\hat{\theta}_n^Q) + \Phi^{-1}(\frac{\alpha}{2}) \cdot s.e(g(\hat{\theta}_n^Q)), g(\hat{\theta}_n^Q) + \Phi^{-1}(1 - \frac{\alpha}{2}) \cdot s.e(g(\hat{\theta}_n^Q)) \right]
\]
where
\[
s.e(g(\hat{\theta}_n^Q)) = \sqrt{\frac{\nabla \theta g(\hat{\theta}_n^Q)' \hat{J}_n(\theta_0)^{-1} \hat{\Omega}(\theta_0)^{-1} \hat{\Omega}(\theta_0)^{-1} \nabla \theta g(\hat{\theta}_n^Q)}}{n},
\]
and \( \hat{J}(\theta_0)^{-1} = \int \nabla \theta n(\theta - \hat{\theta}_n^Q)(\theta - \hat{\theta}_n^Q)' p_n(\theta) d\theta \), \( \hat{\Omega}(\theta_0) \) is a consistent estimate for \( \Omega_n(\theta_0) \), and \( \Phi \) is the standard normal distribution function. In practice, the value of \( g(\hat{\theta}_n^Q) \) and \( \hat{J}(\theta_0)^{-1} \) is obtained by calculating the sample mean and the variance for MCMC sequence \( S = \{\theta^{(1)}, \theta^{(2)}, \theta^{(3)}, \ldots, \theta^{(B)}\} \). For simplicity of the discussion, I assume that the MCMC sequence is properly burned in. The confidence interval constructed above, therefore, implicitly assumes that we can have arbitrarily high precision for calculating the mean and the variance of quasi-posterior. In fact, the feasible point estimate and confidence interval are calculated based on
\[
\hat{\theta}_n^Q = \frac{1}{B} \sum_{s=1}^{B} \theta^{(s)}, \theta^{(s)} \in S
\]
instead of \( \hat{\theta}_n^Q \), the true mean of quasi-posterior. Hence, the standard error of the feasible estimate \( \hat{\theta}_n^Q \) should reflect the additional error from the finite number of MCMC draws. Since the MCMC draws are statistically independent from \( \hat{\theta}_n^Q \) by construction, the Monte-Carlo
error corrected standard error of \( g(\tilde{\theta}_n^Q) \) can be written as

\[
\sqrt{n} \frac{\nabla_{\theta} g(\tilde{\theta}_n^Q)' \tilde{J}_n(\theta_0)^{-1} \hat{\Omega}(\theta_0) \tilde{J}(\theta_0)^{-1} \nabla_{\theta} g(\tilde{\theta}_n^Q)}{\nabla_{\theta} g(\tilde{\theta}_n^Q)' V \text{ar}(\frac{1}{B} \sum_{s=1}^{B} \theta^{(s)}) \nabla_{\theta} g(\tilde{\theta}_n^Q)} + V \text{ar}(\frac{1}{B} \sum_{s=1}^{B} \theta^{(s)}) \nabla_{\theta} g(\tilde{\theta}_n^Q)
\]

where

\[
\tilde{J}(\theta_0)^{-1} = \frac{1}{B} \sum_{s=1}^{B} (\theta^{(s)} - \tilde{\theta}_n^Q)(\theta^{(s)} - \tilde{\theta}_n^Q)'
\]

Whether the later part is of significance or not mainly depends on how large \( B \) is. That is, we can arbitrarily reduce the later part of the error by simply increasing the number of MCMC draws. However, many practical computational restrictions hinder the possibility of having very many MCMC draws, i.e. setting \( B \) very large. For example, some problems have highly complicated criterion functions \( L_n(\theta) \) which makes the repeated evaluation of \( L_n(\theta) \) required in MCMC quite costly. When calculating \( L_n(\theta) \) involves nested iterative computation, increasing \( B \) may be very costly because the marginal computational cost involves additional set of nested iterations.

Even if we consider fixed \( B \), there is another factor that affects the size of the Monte-Carlo error. This factor is the persistence of the Markov chain. For example, for a random walk Metropolis-Hastings algorithm for a high dimensional problem, the optimal acceptance rate is known to be about 0.234. (See Gelman, Roberts and Gilks (1996) or Roberts and Rosenthal (2007)) This means that the Markov chain \( S \) is quite persistent which makes the variance of \( \sum \theta^{(s)} \), if it exists, large in general compared to independent Monte-Carlo draws.

There are two questions to address before we attempt to quantify the key component of the Monte-Carlo error,

\[
\sqrt{B} \text{V} \text{ar}(\frac{1}{B} \sum_{s=0}^{B} \theta^{(s)}).
\]

First, we need to verify whether the quantity exists at all in the limit. That is, we need to ensure that a central limit theorem goes through for the sum of MCMC draws. More precisely, we need to show that

\[
\sqrt{B}(\frac{1}{B} \sum_{s=1}^{B} \theta^{(s)} - \tilde{\theta}_n^Q) \overset{d}{\rightarrow} N(0, \Omega_n^M).
\]
Second, we need to have a consistent estimator for $\Omega_n^M$, which can be used as a proxy for $\sqrt{B} \cdot \text{Var}(\frac{1}{B} \sum_{s=1}^{B} \theta^{(s)})$. With these two points addressed, we can construct the Monte-Carlo error corrected standard error for the feasible QBE as follows.

$$s.e.(\tilde{\theta}_n^Q) = \sqrt{\frac{\nabla_{\theta} g(\tilde{\theta}_n^Q) \cdot \tilde{J}_n(\theta_0)^{-1} \tilde{\Omega}(\theta_0) \tilde{J}(\theta_0)^{-1} \nabla_{\theta} g(\tilde{\theta}_n^Q) + \nabla_{\theta} g(\tilde{\theta}_n^Q) \tilde{\Omega}_n^M \nabla_{\theta} g(\tilde{\theta}_n^Q)}{n B}}$$

where $\tilde{\Omega}_n^M$ is a consistent estimator for $\Omega_n^M$.

### 2.3. Geometric Ergodicity of MCMC Chain

In this section, I show that the MCMC draws for Quasi-posterior follow a central limit theorem. I mainly consider the Metropolis-Hastings algorithm with random walk proposal density, which is mostly used in practice for QBEs and other non-hierarchical Bayesian models. For the detailed discussion, I review some of the results on central limit theorems for Markov chains. The approach in this section follows that of Meyn and Tweedie (2009), Jones (2004) and Roberts and Tweedie (1996). We start with some basic definitions of Markov chain theory.

**Definitions.**

1. A Markov chain transition kernel $P$ is $\pi$-irreducible if for any $x \in \mathcal{X}$ and for any set $A$ with $\pi(A) > 0$, there exists and $n$ such that $P^n(x, A) > 0$.
2. A $\pi$-irreducible $P$ is periodic if there exists an integer $d \geq 2$ and a collection of disjoint sets $A_1, \ldots, A_d \in B$ such that for each $x \in A_j$, $P(x, A_{j+1}) = 1$ for $j = 1, \ldots, d - 1$ and for $x \in A_d$, $P(x, A_1) = 1$. Otherwise, $P$ is said to be aperiodic.
3. A $\pi$-irreducible Markov chain $\{X_n\}$ with stationary distribution $\pi$ is Harris recurrent if for every set $A$ with $\pi(A) > 0$ $\Pr(X_n \in A \text{ i.o.} | X_1 = x) = 1$ for all $x$.
4. If $\pi$ is a probability distribution, the Markov chain $X$ is said to be Harris ergodic if it is $\pi$-irreducible, aperiodic and Harris recurrent.

One obvious example that satisfies all the conditions above but stationarity is random walk in single dimension, because random walk can potentially visit any set with positive probability measure in $\mathcal{X}$ infinitely often. Straightforwardly, Metropolis-Hastings algorithm with
proposal density of random walk satisfies irreducibility, aperiodicity and Harris recurrency as long as its acceptance probability is bounded away from zero. In fact, Mengersen and Tweedie (1996) showed that this is true for any Metropolis-Hastings chain in general. The following proposition from Athreya et al (1996) provides a useful implication of this.

**PROPOSITION.** If $P$ is Harris ergodic, then we have

$$||P^n(x, \cdot) - \pi(\cdot)|| \to 0 \quad \text{as} \quad n \to \infty$$

where the norm is total variation norm,

$$||P^n(x, \cdot) - \pi(\cdot)|| \equiv \sup_{A \in \mathcal{B}} |P^n(x, A) - \pi(A)|$$

This shows that the Markov chain eventually converges to the stationary distribution in terms of total variation norm. With Harris ergodicity, we can treat a Markov chain as if it is a stationary process in terms of our problem. This follows from the following proposition of Meyn and Tweedie (2009).

**PROPOSITION.** If a CLT holds for any one initial distribution for a Harris ergodic Markov chain, then it holds for every initial distribution.

The convergence in total variation norm provides direct link to Markov chain central limit theorems. In order to establish proper rate of convergence and obtain a limit distribution, we need stronger notion of ergodicity, which is geometric ergodicity. A Harris ergodic Markov chain $P$ is said to be geometrically ergodic if there exist $t \in (0, 1)$ and a non-negative function $M(x)$ such that

$$||P^n(x, \cdot) - \pi(\cdot)|| < M(x)t^n$$

We have established that Metropolis-Hastings random walk chain is Harris ergodic by construction. Roberts and Tweedie (1996) give a set of sufficient conditions for a multivariate Metropolis-Hastings chain to be geometrically ergodic. They show that the geometric ergodicity of a Metropolis-Hastings random walk chain is purely determined by the shape.
and tail behavior of the target density, the quasi-posterior in our case. The following theorem from Roberts and Tweedie (1996) sets sufficient conditions for the shape.

**THEOREM.** Let \( p \) and \( h \) be polynomials on \( \mathbb{R}^d \), and

\[
P = \{ \pi | \pi(x) = h(x) \exp\{p(x)\} \} \quad \text{s.t.} \quad \lim_{|x| \to \infty} p_m(x) \to -\infty \forall m \geq 2,
\]

where \( p_m \) denotes the polynomial consisting only of \( p \)'s \( m \)th order terms. For any target density \( \pi \in P \), symmetric random walk Metropolis-Hastings chain is geometrically ergodic.

The large sample limiting distribution of quasi-posterior under the assumptions in Chernozhukov and Hong (2003) is Gaussian distribution which is contained in \( P \). Moreover, we assumed compact parameter space \( \Theta \) for the extremum estimation. Thus, there is no tail issue for the quasi-posterior and all of the primitive conditions for geometric ergodicity in Theorem 3.2 of Roberts and Tweedie (1996) are easily satisfied.

**THEOREM.** A random walk Metropolis-Hastings chain for a quasi-posterior that has acceptance probability bounded away from zero is geometrically ergodic under the Assumption 1 and 2.

Geometric ergodicity turns out to be a sufficient condition for establishing a CLT. This can be proved by showing that geometric ergodic Markov chain is strongly mixing (or \( \alpha \)-mixing) with exponential rate.

**LEMMA.** Geometric ergodic Markov chain is strongly mixing with exponential mixing rate if

\[
E_\pi[M(X)] < \infty,
\]

where \( M(x) \) is defined as above.

**PROOF.** A sequence \( \{X_n\} \) is said to be strongly mixing if the mixing coefficient \( \alpha(n) \to 0 \) as \( n \to \infty \) where

\[
\alpha(n) = \sup_{k \geq 1} \sup_{B \in \mathcal{F}_k} \sup_{A \in \mathcal{F}_{k+n}} |P(A \cap B) - P(A)P(B)|,
\]
where $\mathcal{F}_{t_1}^{t_2}$ is a sigma field generated by \{\(X_t : t_1 \leq t \leq t_2\). Let \(P^n(x, \cdot)\) be the transition kernel of a geometrically ergodic Markov chain with stationary distribution \(\pi(x)\). Note that
\[
\int_B |P^n(x, A) - \pi(A)| \pi(dx) \geq \int_B (P^n(x, A) - \pi(A)) \pi(dx) \geq \left| \Pr(X_n \in A, X_0 \in B) - \pi(A) \pi(B) \right|
\]
and thus we have
\[
\alpha(n) \leq \int \sup_{A \in B} |P^n(x, A) - \pi(A)| \pi(dx) \leq E_{\pi}[M(X)]^n,
\]
where \(B\) is the Borel sigma field on the support of \(\pi, X\).

Note that the assumption on the existence of \(E_{\pi}[M(X)]\) holds in our case because the parameter set \(\Theta\) is assumed to be compact. Using this result along with the following classic central limit theorem for strongly mixing sequences, we can establish the desired result.

**THEOREM.** Let \(X_n\) be a centered strictly stationary strongly mixing sequence with mixing coefficient \(\alpha(n)\) such that \(E|X_n|^{2+\delta} < \infty\) for some \(\delta > 0\) and
\[
\sum_n \alpha(n)^{\delta/(2+\delta)} < \infty.
\]
Then we have,
\[
\sigma^2 = E[X_0^2] + 2 \sum_{j=1}^{\infty} E[X_0 X_j] < \infty,
\]
and if \(\sigma^2 > 0\),
\[
\frac{1}{\sqrt{n}} \sum_{k=1}^{n} X_k \overset{d}{\rightarrow} N(0, \sigma^2).
\]

**COROLLARY.** Let \(\{\theta(s), s = 1, \ldots\}\) be a random walk Metropolis-Hastings Markov chain for a quasi-posterior \(p_n(\theta)\). Suppose \(E_{p_n}|\theta|^{2+\delta} < \infty\) for some \(\delta > 0\). Let \(\hat{\theta}_n^Q = E_{p_n}[\theta]\) and
\[
\Omega_M = E_{p_n}[(\theta^{(1)} - \hat{\theta}_n^Q)(\theta^{(1)} - \hat{\theta}_n^Q)'] + \sum_{i=2}^{\infty} E_{p_n}[(\theta^{(i)} - \hat{\theta}_n^Q)(\theta^{(i)} - \hat{\theta}_n^Q)'] + \sum_{i=2}^{\infty} E_{p_n}[(\theta^{(i)} - \hat{\theta}_n^Q)(\theta^{(1)} - \hat{\theta}_n^Q)']
\]
Then we have,
\[
\sqrt{B} \left( \frac{1}{B} \sum_{s=1}^{B} \theta(s) - \hat{\theta}_n^Q \right) \overset{d}{\rightarrow} N(0, \Omega_M).
\]
2.4. CONSISTENT ESTIMATION OF VARIANCE-COVARIANCE MATRIX

We have shown that the central limit theorem holds for random walk Metropolis-Hastings chains for Quasi-posteriors. For actual calculation of the Monte-Carlo error, we need to obtain a consistent estimator for the variance-covariance matrix $\Omega_n^M$. In this section, I show that geometric ergodicity is indeed a sufficient condition for the consistency of classic non-parametric HAC estimators. Andrews (1991) provides a set of primitive conditions for kernel based non-parametric estimator to be consistent. Let $V_t = \theta(t) - \hat{\theta}_n^Q \in \mathbb{R}^p$ and $\Gamma(j)$ be the covariance of $V_t$ and $V_{t+j}$. Let $\kappa_{abcd}(t, t + j, t + m, t + n)$ be the element-by-element fourth order cumulant of $(V_t, V_{t+j}, V_{t+m}, V_{t+n})$ where $a, b, c, d$ indicate indices of elements. Suppose that the following conditions hold:

1. $\{V_t\}$ is a mean zero, fourth order stationary sequence of rv’s with $\sum_{j=-\infty}^{\infty} ||\Gamma(j)|| < \infty$ and
   $$\sum_{j=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \kappa_{abcd}(0, j, m, n) < \infty \quad \forall a, b, c, d \leq p$$

2. $\sqrt{T} \frac{1}{h} \sum_t V_t = O_p(1)$, $\sup_{t \geq 1} E||V_t||^2 < \infty$ and $\int_{-\infty}^{\infty} |k(x)| dx < \infty$ where $k(x)$ is a kernel.

Under 1 and 2, a consistent estimator for $\Omega_n^M = \sum_{j=-\infty}^{\infty} \Gamma(j)$ exists and it can be written as,

$$\hat{\Omega}_n^M = \sum_{j=-T+1}^{T-1} k(j) \hat{\Gamma}(j)$$

where

$$\hat{\Gamma}(j) = \begin{cases} \frac{1}{T} \sum_{t=j+1}^{T} \hat{V}_t \hat{V}'_{t-j} & \text{for } j \geq 0 \\ \frac{1}{T} \sum_{t=-j+1}^{T} \hat{V}_{t+j} \hat{V}'_{t+j} & \text{for } j < 0 \end{cases}$$

Here, $S_T$ is the truncation parameter that is increasing in $T$ and $\hat{V}_t = \theta(t) - \hat{\theta}_n^Q$. See Andrews (1991) for detailed discussion.

The following theorem connects geometric ergodicity and consistency of $\hat{\Omega}_n^M$.

**THEOREM.** Let $\{\theta(s), s = 1, \ldots\}$ be a random walk Metropolis-Hastings Markov chain for a quasi-posterior $p_n(\theta)$ and $\hat{\Omega}_n^M$ be defined as above. Then $\hat{\Omega}_n^M$ is a consistent estimator of $\Omega_n^M$ if $\frac{S_T^2}{T} \to 0$ and $S_T \to \infty$. 

42
PROOF. Lemma 1 of Andrews (1991) states that if \( \{V_t\} \) is a mean zero \( \alpha \)-mixing sequence of rv's with

\[
\sup_{t \geq 1} E[|V_t|^4] < \infty \\
\sum_{j=1}^{\infty} j^2 \alpha(j)^{(v-1)/v} < \infty
\]

for some \( v > 1 \), then we have

\[
\sum_{j=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \kappa_{abcd}(0, j, m, n) < \infty \quad \forall a, b, c, d \leq p.
\]

The moment condition holds for any \( v \) because the quasi-posterior has compact support, and the mixing coefficient condition is easily satisfied by the geometric ergodicity. Thus, the first condition for the consistency holds. Also, the second condition is satisfied by the result of the previous section. Therefore, by the Theorem 1 of Andrews (1991), we have

\[
\hat{\Omega}_n^M \overset{p}{\to} \Omega_n^M.
\]

\[\square\]

2.5. MONTE-CARLO SIMULATIONS

There has been a fairly large literature concerning the censored quantile regression model since Powell (1988). One of the main issues has been the difficulty of calculating the extremum estimate. Buchinsky and Hahn (1998) proposed an iterated linear programming (ILP) algorithm to tackle the problem. As shown in Chernozhukov and Hong (2003), the ILP algorithm seems to be unreliable in cases with high degree of censoring compared to quasi-Bayesian estimation. Thus, censored quantile regression is one of good examples that show merits of QBE.

I consider a simple version of censored quantile regression model for expository purpose. The model is the same as that of Chernozhukov and Hong (2003) for comparison. The actual computation cost for calculating QBE in this model is relatively low since it is a simple
toy model that does not involve any nested iterations. However, the simulation result from this simple model can shed light on the importance of accounting for Monte-Carlo errors. Consider the following model:

$$Y^* = \beta_0 + X'\beta + u$$

$$X \sim N(0, I_3)$$

$$u = X_1^2 N(0, 1)$$

$$Y = \max\{0, Y^*\}$$

where the true parameter $$(\beta_0, \beta_1, \beta_2, \beta_3) = (-1, 3, 3, 3)$$. We observe data set of copies of $$(X, Y)$$ with sample size $$n$$. The setting produces approximately 55% censoring. The criterion function for extremum estimation is

$$L_n(\theta) = -\sum_{i=1}^{n} |Y_i - \max(0, \beta_0 + X_i'\beta)|.$$ 

For quasi-Bayesian estimation, I set the initial values equal to OLS estimates and burned first 500 draws. Gaussian random walk Metropolis-Hastings algorithm is employed and the step size is adjusted to have acceptance probability of 20% ~ 25%, which is considered to be optimal in the literature. The Table 1 shows the percentage of Monte-Carlo error contained in mean-squared error. The ratio is calculated as

$$R = \frac{\hat{\Omega}_{n,11}^M / B}{MSE(\hat{\beta}_n)}$$

where $$\hat{\Omega}_{n,11}^M$$ is the first diagonal element of $$\hat{\Omega}_n^M$$ defined as above. It also shows the coverage probability of 90% confidence intervals using regular standard error calculated from QBE and Monte-Carlo error corrected standard error. Finally, the table provides accuracy of the Monte-Carlo variance term, $$\hat{\Omega}_{n,11}^M$$ by calculating normalized RMSE of it. Following the usual convention, normalizing constant is the difference between maximum and minimum among the set of $$\hat{\Omega}_{n,11}^M$$'s. For the last statistic, I fixed the values of $$(Y, X)$$ and generated many different Markov chains for calculating QBEs.
Table 1. Monte-Carlo Error in Censored Quantile Regression (based on 1,000 repetitions)

<table>
<thead>
<tr>
<th>$\beta_0$</th>
<th>$n = 200$</th>
<th>$n = 800$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$B = 1000$</td>
<td>$B = 4000$</td>
</tr>
<tr>
<td>Ratio of MC Error</td>
<td>0.0672</td>
<td>0.0209</td>
</tr>
<tr>
<td>Coverage of S.E.</td>
<td>0.8130</td>
<td>0.8690</td>
</tr>
<tr>
<td>Coverage of Corrected S.E.</td>
<td>0.8320</td>
<td>0.8760</td>
</tr>
<tr>
<td>Normalized RMSE of $Q_{n,11}^M$</td>
<td>0.1198</td>
<td>0.0916</td>
</tr>
</tbody>
</table>

We can observe that the Monte-Carlo error has non-trivial proportion in the actual mean squared error of QBE. Obviously, this problem is most pronounced when we have relatively small number of MCMC draws. By incorporating the Monte-Carlo error, we can achieve some improvement in coverage probabilities of confidence intervals. The estimation of $\hat{Q}_n^M$ is done by setting $S_T = 150$ for $B = 1000$ case and $S_T = 250$ for $B = 4000$ case. This is rather arbitrary, but considering the nature of geometric mixing, these numbers can be deemed to be reasonable. As for the kernel, the Bartlett kernel is employed. Despite relatively large number of MCMC draws, the accuracy of $\hat{Q}_n^M$ is not so impressive. This is largely due to bias in the estimation, potentially caused by low acceptance rate of Metropolis Hastings chain.

Sample size $n$ has two counteractive effects on the importance of the Monte-Carlo error. As sample size $n$ grows, the precision of QBE improves and thus the relative size of the Monte-Carlo error increases. However, growing sample size also shrinks the quasi-posterior by the rate of $\sqrt{n}$. Shrunken quasi-posterior reduces the Monte-Carlo error by construction of the Metropolis-Hastings chain, i.e. the chain will bounce within a narrower space so that the variance gets reduced. These counteractive effects make the Monte-Carlo error relevant for any sample size.

2.6. CONCLUSION

QBE provides a simple alternative to general class of extremum estimators when direct optimization is cumbersome or infeasible in practice. QBE utilizes MCMC to calculate mean of quasi-posterior and therefore numerical error or Monte-Carlo error arises. However, the
Monte-Carlo error is rarely accounted in practice. This paper quantifies the Monte-Carlo error and provides a method to incorporate such error in the statistical inference for quasi-Bayesian estimators. A central limit theorem is established for random walk Metropolis-Hastings algorithm for QBE problems and a consistent estimator for the Monte-Carlo error is provided.

It turns out that the Monte-Carlo error can be a substantial part of the actual standard error of quasi-Bayesian estimators. Adjusting standard errors to incorporate the Monte-Carlo error would result in better statistical inference, i.e. more accurate coverage of confidence intervals. Practical relevance of the Monte-Carlo error purely depends on the nature of the statistical model and one cannot be sure about it ex-ante. Unless a researcher can draw large number of MCMC draws cheaply, he or she should take a close look at the Monte-Carlo error and incorporate it into statistical inference procedures.
Bibliography


CHAPTER 3

Panel Data Models with Nonadditive Unobserved Heterogeneity: Estimation and Inference

Coauthored with Iván Fernández-Val

1This chapter is published in Quantitative Economics, Nov, 2013.
3.1. INTRODUCTION

This paper considers estimation and inference in linear and nonlinear panel data models with random coefficients and endogenous regressors. The quantities of interest are means, variances, and other moments of the distribution of the random coefficients. In a state level panel model of rational addiction, for example, we might be interested in the mean and variance of the distribution of the price effect on cigarette consumption across states, controlling for endogenous past and future consumptions. These models pose important challenges in estimation and inference if the relation between the regressors and random coefficients is left unrestricted. Fixed effects methods based on GMM estimators applied separately to the time series of each individual can be severely biased due to the incidental parameter problem. The source of the bias is the finite-sample bias of GMM if some of the regressors is endogenous or the model is nonlinear in parameters, or nonlinearities if the parameter of interest is the variance or other high order moment of the random coefficients. Neglecting the heterogeneity and imposing fixed coefficients does not solve the problem, because the resulting estimators are generally inconsistent for the mean of the random coefficients (Yitzhaki, 1996, and Angrist, Graddy and Imbens, 2000). Moreover, imposing fixed coefficients does not allow us to estimate other moments of the distribution of the random coefficients.

We introduce a class of bias-corrected panel fixed effects GMM estimators. Thus, instead of imposing fixed coefficients, we estimate different coefficients for each individual using the time series observations and correct for the resulting incidental parameter bias. For linear models, in addition to the bias correction, these estimators differ from the standard fixed effects estimators in that both the intercept and the slopes are different for each individual. Moreover, unlike for the classical random coefficient estimators, they do not rely on any restriction in the relationship between the regressors and random coefficients; see Hsiao and Pesaran (2004) for a recent survey on random coefficient models. This flexibility allows us to account for Roy (1951) type selection where the regressors are decision variables with levels determined by their returns. Linear models with Roy selection are commonly referred to as correlated random coefficient models in the panel data literature. In the presence of endogenous regressors, treating the random coefficients as fixed effects is also convenient to overcome the identification problems in these models pointed out by Kelejian (1974).

The most general models we consider are semiparametric in the sense that the distribution of the random coefficients is unspecified and the parameters are identified from moment conditions. These conditions can be nonlinear functions in parameters and variables, accommodating both linear and nonlinear random coefficient models, and allowing for the presence of time varying endogeneity in the regressors not captured by the random coefficients. We

\[\text{Heckman and Vytlacil (2000) and Angrist (2004) find sufficient conditions for fixed coefficient OLS and IV estimators to be consistent for the average coefficient.}\]
use the moment conditions to estimate the model parameters and other quantities of interest via GMM methods applied separately to the time series of each individual. The resulting estimates can be severely biased in short panels due to the incidental parameters problem, which in this case is a consequence of the finite-sample bias of GMM (Newey and Smith, 2004) and/or the nonlinearity of the quantities of interest in the random coefficients. We develop analytical corrections to reduce the bias.

To derive the bias corrections, we use higher-order expansions of the GMM estimators, extending the analysis in Newey and Smith (2004) for cross sectional estimators to panel data estimators with fixed effects and serial dependence. If \( n \) and \( T \) denote the cross sectional and time series dimensions of the panel, the corrections remove the leading term of the bias of order \( O(T^{-1}) \), and center the asymptotic distribution at the true parameter value under sequences where \( n \) and \( T \) grow at the same rate. This approach is aimed to perform well in econometric applications that use moderately long panels, where the most important part of the bias is captured by the first term of the expansion. Other previous studies that used a similar approach for the analysis of linear and nonlinear fixed effects estimators in panel data include, among others, Kiviet (1995), Phillips and Moon (1999), Alvarez and Arellano (2003), Hahn and Kuersteiner (2002), Lancaster (2002), Woutersen (2002), Hahn and Newey (2004), and Hahn and Kuersteiner (2011). See Arellano and Hahn (2007) for a survey of this literature and additional references.

A first distinctive feature of our corrections is that they can be used in overidentified models where the number of moment restrictions is greater than the dimension of the parameter vector. This situation is common in economic applications such as rational expectation models. Overidentification complicates the analysis by introducing an initial stage for estimating optimal weighting matrices to combine the moment conditions, and precludes the use of the existing methods. For example, Hahn and Newey’s (2004) and Hahn and Kuersteiner’s (2011) general bias reduction methods for nonlinear panel data models do not cover optimal two-step GMM estimators. A second distinctive feature is that our results are specifically developed for models with multidimensional nonadditive heterogeneity, whereas the previous studies focused mostly on models with additive heterogeneity captured by an scalar individual effect. Exceptions include Arellano and Hahn (2006) and Bester and Hansen (2008), which also considered multidimensional heterogeneity, but they focus on parametric likelihood-based panel models with exogenous regressors. Bai (2009) analyzed related linear panel models with exogenous regressors and multidimensional interactive individual effects. Bai’s nonadditive heterogeneity allows for interaction between individual effects and unobserved factors, whereas the nonadditive heterogeneity that we consider allows for interaction
between individual effects and observed regressors. A third distinctive feature of our analysis is the focus on moments of the distribution of the individual effects as one of the main quantities of interest.

We illustrate the applicability of our methods with empirical and numerical examples based on the cigarette demand application of Becker, Grossman and Murphy (1994). Here, we estimate a linear rational addictive demand model with state-specific coefficients for price and common parameters for the other regressors using a panel data set of U.S. states. We find that standard estimators that do not account for non-additive heterogeneity by imposing a constant coefficient for price can have important biases for the common parameters, mean of the price coefficient and demand elasticities. The analytical bias corrections are effective in removing the bias of the estimates of the mean and standard deviation of the price coefficient. Figure 1 gives a preview of the empirical results. It plots a normal approximation to the distribution of the price effect based on uncorrected and bias corrected estimates of the mean and standard deviation of the distribution of the price coefficient. The figure shows that there is important heterogeneity in the price effect across states. The bias correction reduces by more than 15% the absolute value of the estimate of the mean effect and by 30% the estimate of the standard deviation.

Some of the results for the linear model are related to the recent literature on correlated random coefficient panel models with fixed $T$. Graham and Powell (2008) gave identification and estimation results for average effects. Arellano and Bonhomme (2010) studied identification of the distributional characteristics of the random coefficients in exogenous linear models. None of these papers considered the case where some of the regressors have time varying endogeneity not captured by the random coefficients or the model is nonlinear. For nonlinear models, Chernozhukov, Fernández-Val, Hahn and Newey (2010) considered identification and estimation of average and quantile treatment effects. Their nonparametric and semiparametric bounds do not require large-$T$, but they do not cover models with continuous regressors and time varying endogeneity.

The rest of the paper is organized as follows. Section 2 illustrates the type of models considered and discusses the nature of the bias in two examples. Section 3 introduces the general model and fixed effects GMM estimators. Section 4 derives the asymptotic properties of the estimators. The bias corrections and their asymptotic properties are given in Section 5. Section 6 describes the empirical and numerical examples. Section 7 concludes with a summary of the main results. Additional numerical examples, proofs and other technical details are given in the online supplementary appendix Fernández-Val and Lee (2012).
3.2. Motivating Examples

In this section we describe in detail two simple examples to illustrate the nature of the bias problem. The first example is a linear correlated random coefficient model with endogenous regressors. We show that averaging IV estimators applied separately to the time series of each individual is biased for the mean of the random coefficients because of the finite-sample bias of IV. The second example considers estimation of the variance of the individual coefficients in a simple setting without endogeneity. Here the sample variance of the estimators of the individual coefficients is biased because of the non-linearity of the variance operator in the individual coefficients. The discussion in this section is heuristic leaving to Section 4 the specification of precise regularity conditions for the validity of the asymptotic expansions used.

3.2.1. Correlated random coefficient model with endogenous regressors.
Consider the following panel model:

\[(2.1) \quad y_{it} = \alpha_{0i} + \alpha_{1i} x_{it} + \epsilon_{it}, \quad (i = 1, \ldots, n; t = 1, \ldots, T);\]

where \(y_{it}\) is a response variable, \(x_{it}\) is an observable regressor, \(\epsilon_{it}\) is an unobservable error term, and \(i\) and \(t\) usually index individual and time period, respectively.\(^2\) This is a linear random coefficient model where the effect of the regressor is heterogenous across individuals, but no restriction is imposed on the distribution of the individual effect vector \(\alpha_i := (\alpha_{0i}, \alpha_{1i})'.\)

The regressor can be correlated with the error term and a valid instrument \((1, z_{it})\) is available for \((1, x_{it})\), that is \(E[\epsilon_{it} | \alpha_i] = 0, E[z_{it} \epsilon_{it} | \alpha_i] = 0\) and \(Cov[z_{it} x_{it} | \alpha_i] \neq 0\). An important example of this model is the panel version of the treatment-effect model (Wooldridge, 2002 Chapter 10.2.3, and Angrist and Hahn, 2004). Here, the objective is to evaluate the effect of a treatment \((D)\) on an outcome variable \((Y)\). The average causal effect for each level of treatment is defined as the difference between the potential outcome that the individual would obtain with and without the treatment, \(Y_D - Y_0\). If individuals can choose the level of treatment, potential outcomes and levels of treatment are generally correlated. An instrumental variable \(Z\) can be used to identify the causal effect. If potential outcomes are represented as the sum of permanent individual components and transitory individual-time specific shocks, that is \(Y_{jit} = Y_{ji} + \epsilon_{jit}\) for \(j \in \{0, 1\}\), then we can write this model as a special case of (2.1) with \(y_{it} = (1 - D_i)Y_{0it} + D_i Y_{1it}, \alpha_{0i} = Y_{0i}, \alpha_{1i} = Y_{1i} - Y_{0i}, x_{it} = D_i, z_{it} = Z_{it},\) and \(\epsilon_{it} = (1 - D_i) \epsilon_{0it} + D_i \epsilon_{1it}\).

Suppose that we are ultimately interested in \(\alpha_1 := E[\alpha_{1i}]\), the mean of the random slope coefficient. We could neglect the heterogeneity and run fixed effects OLS and IV regressions

\(^2\)More generally, \(i\) denotes a group index and \(t\) indexes the observations within the group. Examples of groups include individuals, states, households, schools, or twins.
\[ y_{it} = \alpha_{0i} + \alpha_1 x_{it} + u_{it}, \]

where \( u_{it} = x_{it}(\alpha_{1t} - \alpha_1) + \epsilon_{it} \) in terms of the model (2.1). In this case, OLS and IV estimate weighted means of the random coefficients in the population; see, for example, Yitzhaki (1996) and Angrist and Krueger (1999) for OLS, and Angrist, Graddy and Imbens (2000) for IV. OLS puts more weight on individuals with higher variances of the regressor because they give more information about the slope; whereas IV weighs individuals in proportion to the variance of the first stage fitted values because these variances reflect the amount of information that the individuals convey about the part of the slope affected by the instrument. These weighted means are generally different from the mean effect because the weights can be correlated with the individual effects.

To see how these implicit OLS and IV weighting schemes affect the estimand of the fixed-coefficient estimators, assume for simplicity that the relationship between \( x_{it} \) and \( z_{it} \) is linear, that is \( x_{it} = \pi_{0i} + \pi_{1i} z_{it} + \nu_{it} \), \((\epsilon_{it}, \nu_{it})\) is normal conditional on \((z_{it}, \alpha_i, \pi_i)\), \( z_{it} \) is independent of \((\alpha_i, \pi_i)\), and \((\alpha_i, \pi_i)\) is normal, for \( i := (\pi_0i, \pi_{1i})' \). Then, the probability limits of the OLS and IV estimators are\(^3\)

\[
\alpha_{1}^{OLS} = \alpha_1 + \{Cov[\epsilon_{it}, \nu_{it}] + 2E[\pi_{1i}]Var[z_{it}]Cov[\alpha_{1i}, \pi_{1i}]\}/Var[x_{it}],
\]
\[
\alpha_{1}^{IV} = \alpha_1 + Cov[\alpha_{1i}, \pi_{1i}]/E[\pi_{1i}].
\]

These expressions show that the OLS estimand differs from the average coefficient in presence of endogeneity, i.e. non zero correlation between the individual-time specific error terms, or whenever the random coefficients are correlated; while the IV estimand differs from the average coefficient only in the latter case.\(^4\) In the treatment-effects model, there exists correlation between the error terms in presence of endogeneity bias and correlation between the individual effects arises under Roy-type selection, i.e., when individuals who experience a higher permanent effect of the treatment are relatively more prone to accept the offer of treatment. Wooldridge (2005) and Murtazashvile and Wooldridge (2005) give sufficient conditions for consistency of standard OLS and IV fixed effects estimators. These conditions amount to \(Cov[\epsilon_{it}, \nu_{it}] = 0\) and \(Cov[x_{it}, \alpha_{1i} \alpha_{i0}] = 0\).

Our proposal is to estimate the mean coefficient from separate time series estimators for each individual. This strategy consists of running OLS or IV for each individual, and then estimating the population moment of interest by the corresponding sample moment

\(^3\)The limit of the IV estimator is obtained from a first stage equation that imposes also fixed coefficients, that is \( x_{it} = \pi_{0i} + \pi_1 z_{it} + w_{it} \), where \( w_{it} = z_{it}(\pi_{1i} - \pi_1) + \nu_{it} \). When the first stage equation is different for each individual, the limit of the IV estimator is

\[
\alpha_{1}^{IV} = \alpha_1 + 2E[\pi_{1i}]Cov[\alpha_{1i}, \pi_{1i}]/\{E[\pi_{1i}]^2 + Var[\pi_{1i}]\}.
\]

See Theorems 2 and 3 in Angrist and Imbens (1995) for a related discussion.

\(^4\)This feature of the IV estimator is also pointed out in Angrist, Graddy and Imbens (1999), p. 507.
of the individual estimators. For example, the mean of the random slope coefficient in the population is estimated by the sample average of the OLS or IV slopes. These sample moments converge to the population moments of interest as number of individuals \( n \) and time periods \( T \) grow. However, since a different coefficient is estimated for each individual, the asymptotic distribution of the sample moments can have asymptotic bias due to the incidental parameter problem (Neyman and Scott, 1948).

To illustrate the nature of this bias, consider the estimator of the mean coefficient \( \alpha_1 \) constructed from individual time series IV estimators. In this case the incidental parameter problem is caused by the finite-sample bias of IV. This can be explained using some expansions. Thus, assuming independence across \( t \), standard higher-order asymptotics gives (e.g. Rilstone et. al., 1996), as \( T \to \infty \)

\[
\sqrt{T}(\hat{\alpha}_{1i}^{IV} - \alpha_{1i}) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \psi_{it} + \frac{1}{\sqrt{T}} \beta_i + o_P(T^{-1/2}),
\]

where \( \psi_{it} = E[\hat{z}_{it} \hat{x}_{it} \mid \alpha_i, \pi_i]^{-1} \hat{z}_{it} \epsilon_{it} \) is the influence function of IV, \( \beta_i = -E[\hat{z}_{it} \hat{x}_{it} \mid \alpha_i, \pi_i]^{-2} E[\hat{z}_{it}^2 \hat{x}_{it} \epsilon_{it} \mid \alpha_i, \pi_i] \) is the higher-order bias of IV (see, e.g., Nagar, 1959, and Buse, 1992), and the variables with tilde are in deviation from their individual means, e.g., \( \tilde{z}_{it} = z_{it} - E[z_{it} \mid \alpha_i, \pi_i] \). In the previous expression the first order asymptotic distribution of the individual estimator is centered at the truth since \( \sqrt{T}(\hat{\alpha}_{1i}^{IV} - \alpha_{1i}) \to_d N(0, \sigma^2_i) \) as \( T \to \infty \), where \( \sigma^2_i = E[\hat{z}_{it} \hat{x}_{it} \mid \alpha_i, \pi_i]^{-2} E[\hat{z}_{it}^2 \epsilon_{it}^2 \mid \alpha_i, \pi_i] \).

Let \( \hat{\alpha}_1 = n^{-1} \sum_{i=1}^{n} \hat{\alpha}_{1i}^{IV} \), the sample average of the IV estimators. The asymptotic distribution of \( \hat{\alpha}_1 \) is not centered around \( \alpha_1 \) in short panels or more precisely under asymptotic sequences where \( T/n \to 0 \). To see this, consider the expansion for \( \hat{\alpha}_1 \)

\[
\sqrt{n}(\hat{\alpha}_1 - \alpha_1) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\alpha_{1i} - \alpha_1) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\hat{\alpha}_{1i}^{IV} - \alpha_{1i}).
\]

The first term is the standard influence function for a sample mean of known elements. The second term comes from the estimation of the individual elements inside the sample mean. Assuming independence across \( i \) and combining the previous expansions,

\[
\sqrt{n}(\hat{\alpha}_1 - \alpha_1) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\alpha_{1i} - \alpha_1) + \frac{1}{\sqrt{T/nT}} \sum_{i=1}^{n} \sum_{t=1}^{T} \psi_{it} + \frac{\sqrt{n}}{T} \sum_{i=1}^{n} \beta_i + o_P(1).
\]

This expression shows that the bias term dominates the asymptotic distribution of \( \hat{\alpha}_1 \) in short panels under sequences where \( T/\sqrt{n} \to 0 \). Averaging reduces the order of the variance of \( \hat{\alpha}_{1i}^{IV} \), without affecting the order of its bias. In this case the estimation of the random coefficients has no first order effect in the asymptotic variance of \( \hat{\alpha}_1 \) because the second term is of smaller order than the first term.
A potential drawback of the individual by individual time series estimation is that it might more be sensitive to weak identification problems than fixed coefficient pooled estimation. In the random coefficient model, for example, we require that \( E[\xi_{it} | \alpha_i, \tau_i] = \pi_{1i} \neq 0 \) with probability one, i.e., for all the individuals, whereas fixed coefficient IV only requires that this condition holds on average, i.e., \( E[\pi_{1i}] \neq 0 \). The individual estimators are therefore more sensitive than traditional pooled estimators to weak instruments problems. On the other hand, individual by individual estimation relaxes the exogeneity condition by conditioning on additive and non-additive time invariant heterogeneity, i.e, \( E[\xi_{it} \xi_{it}] = 0 \). Traditional fixed effects estimators only condition on additive time invariant heterogeneity. A formal treatment of these identification issues is beyond the scope of this paper.

3.2.2. Variance of individual coefficients. Consider the panel model:

\[
y_{it} = \alpha_i + \epsilon_{it}, \quad \epsilon_{it} | \alpha_i \sim (0, \sigma_{\epsilon}^2), \quad \alpha_i \sim (\alpha, \sigma_{\alpha}^2), \quad (t = 1, \ldots, T; i = 1, \ldots, n);
\]

where \( y_{it} \) is an outcome variable of interest, which can be decomposed in an individual effect \( \alpha_i \) with mean \( \alpha \) and variance \( \sigma_{\alpha}^2 \), and an error term \( \epsilon_{it} \) with zero mean and variance \( \sigma_{\epsilon}^2 \) conditional on \( \alpha_i \). The parameter of interest is \( \sigma_{\alpha}^2 = Var[\alpha_i] \) and its fixed effects estimator is

\[
\hat{\sigma}_{\alpha}^2 = (n - 1)^{-1} \sum_{i=1}^{n} (\hat{\alpha}_i - \hat{\alpha})^2,
\]

where \( \hat{\alpha}_i = T^{-1} \sum_{t=1}^{T} y_{it} \) and \( \hat{\alpha} = n^{-1} \sum_{i=1}^{n} \hat{\alpha}_i \).

Let \( \varphi_{\alpha_i} = (\alpha_i - \alpha)^2 - \sigma_{\alpha}^2 \) and \( \varphi_{\epsilon_{it}} = \epsilon_{it}^2 - \sigma_{\epsilon}^2 \). Assuming independence across \( i \) and \( t \), a standard asymptotic expansion gives, as \( n, T \to \infty \),

\[
\sqrt{n}(\hat{\sigma}_{\alpha}^2 - \sigma_{\alpha}^2) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi_{\alpha_i} + \frac{1}{\sqrt{T} \sqrt{nT}} \sum_{i=1}^{n} \sum_{t=1}^{T} \varphi_{\epsilon_{it}} + \frac{\sqrt{n}}{T} \sigma_{\epsilon}^2 + o_P(1).
\]

The first term corresponds to the influence function of the sample variance if the \( \alpha_i \)'s were known. The second term comes from the estimation of the \( \alpha_i \)'s. The third term is a bias term that comes from the nonlinearity of the variance in \( \hat{\alpha}_i \). The bias term dominates the expansion in short panels under sequences where \( T/\sqrt{n} \to 0 \). As in the previous example, the estimation of the \( \alpha_i \)'s has no first order effect in the asymptotic variance since the second term is of smaller order than the first term.

---

\(^5\)We thank a referee for pointing out this issue.
3.3. THE MODEL AND ESTIMATORS

We consider a general model with a finite number of moment conditions $d_g$. To describe it, let the data be denoted by $z_{it}$ ($i = 1, \ldots, n; t = 1, \ldots, T$). We assume that $z_{it}$ is independent over $i$ and stationary and strongly mixing over $t$. Also, let $\theta$ be a $d_\theta$-vector of common parameters, $\{\alpha_i : 1 \leq i \leq n\}$ be a sequence of $d_\alpha$-vectors with the realizations of the individual effects, and $g(z; \theta, \alpha_i)$ be an $d_g$-vector of functions, where $d_g \geq d_\theta + d_\alpha$. The model has true parameters $\theta_0$ and $\{\alpha_{i0} : 1 \leq i \leq n\}$, satisfying the moment conditions

$$E[g(z_{it}; \theta_0, \alpha_{i0})] = 0, \ (t = 1, \ldots, T; i = 1, \ldots, n),$$

where $E[\cdot]$ denotes conditional expectation with respect to the distribution of $z_{it}$ conditional on the individual effects.

Let $\bar{E}[\cdot]$ denote the expectation taken with respect to the distribution of the individual effects. In the previous model, the ultimate quantities of interest are smooth functions of parameters and observations, which in some cases could be the parameters themselves,

$$\zeta = \bar{E}[\zeta_i(z_{it}; \theta_0, \alpha_{i0})],$$

if $\bar{E}|\zeta_i(z_{it}; \theta_0, \alpha_{i0})| < \infty$, or moments or other smooth functions of the individual effects

$$\mu = \bar{E}[\mu(\alpha_{i0})],$$

if $\bar{E}|\mu(\alpha_{i0})| < \infty$. In the correlated random coefficient example, $g(z_{it}; \theta_0, \alpha_{i0}) = z_{it}(y_{it} - \alpha_{0i0} - \alpha_{1i0}x_{it})$, $\theta = \emptyset$, $d_\theta = 0$, $d_\alpha = 2$, and $\mu(\alpha_{i0}) = \alpha_{1i0}$. In the variance of the random coefficients example, $g(z_{it}; \theta_0, \alpha_{i0}) = (y_{it} - \alpha_{0i0})$, $\theta = \emptyset$, $d_\theta = 0$, $d_\alpha = 1$, and $\mu(\alpha_{i0}) = (\alpha_{1i0} - \bar{E}[\alpha_{1i0}])^2$.

Some more notation, which will be extensively used in the definition of the estimators and in the analysis of their asymptotic properties, is the following

$$\Omega_{ji}(\theta, \alpha_i) := E[g(z_{it}; \theta, \alpha_i)g(z_{it-j}; \theta, \alpha_i')], \ j \in \{0, 1, 2, \ldots\},$$

$$G_{\theta_i}(\theta, \alpha_i) := E[G_{\theta}(z_{it}; \theta, \alpha_i)] = E[\partial g(z_{it}; \theta, \alpha_i)/\partial \theta'],$$

$$G_{\alpha_i}(\theta, \alpha_i) := E[G_{\alpha}(z_{it}; \theta, \alpha_i)] = E[\partial g(z_{it}; \theta, \alpha_i)/\partial \alpha_i'],$$

where superscript ' denotes transpose and higher-order derivatives will be denoted by adding subscripts. Here $\Omega_{ji}$ is the covariance matrix between the moment conditions for individual $i$ at times $t$ and $t-j$, and $G_{\theta_i}$ and $G_{\alpha_i}$ are time series average derivatives of these conditions.

---

6We impose that some of the parameters are common for all the individuals to help preserve degrees of freedom in estimation of short panels with many regressors. An order condition for this model is that the number of individual specific parameters $d_\alpha$ has to be less than the time dimension $T$. 

58
Analogously, for sample moments

\[ \widehat{\Omega}_{ij}(\theta, \alpha_i) := T^{-1} \sum_{t=j+1}^{T} g(z_{it}; \theta, \alpha_i) g(z_{it-j}; \theta, \alpha_i)', \ j \in \{0, 1, ..., T - 1\}, \]

\[ \widehat{G}_{\theta_i}(\theta, \alpha_i) := T^{-1} \sum_{t=1}^{T} G_\theta(z_{it}; \theta, \alpha_i) = T^{-1} \sum_{t=1}^{T} \partial g(z_{it}; \theta, \alpha_i)/\partial \theta', \]

\[ \widehat{G}_{\alpha_i}(\theta, \alpha_i) := T^{-1} \sum_{t=1}^{T} G_\alpha(z_{it}; \theta, \alpha_i) = T^{-1} \sum_{t=1}^{T} \partial g(z_{it}; \theta, \alpha_i)/\partial \alpha_i'. \]

In the sequel, the arguments of the expressions will be omitted when the functions are evaluated at the true parameter values \((\theta_0', \alpha_0')', \ e.g., g(z_{it}) means g(z_{it}; \theta_0, \alpha_0). \)

In cross-section and time series models, parameters defined from moment conditions are usually estimated using the two-step GMM estimator of Hansen (1982). To describe how to adapt this method to panel models with fixed effects, let \( \widehat{\gamma}_i(\theta, \alpha_i) := T^{-1} \sum_{t=1}^{T} g(z_{it}; \theta, \alpha_i), \) and let \( (\hat{\theta}', \{\hat{\alpha}_i\}_{i=1}^{n})' \) be some preliminary one-step FE-GMM estimator, given by

\[ (\theta', \{\alpha_i\}_{i=1}^{n})' = \arg \inf \{(\theta', \{\alpha_i\}_{i=1}^{n})' \sum_{i=1}^{n} \widehat{\gamma}_i(\theta, \alpha_i) \mathbf{W}_i^{-1} \widehat{\gamma}_i(\theta, \alpha_i), \] where \( \mathcal{T} \subset \mathbb{R}^{d_\theta + d_\alpha} \) denotes the parameter space, and \( \{\mathbf{W}_i : 1 \leq i \leq n\} \) is a sequence of positive definite symmetric \( d_\theta \times d_\theta \) weighting matrices. The two-step FE-GMM estimator is the solution to the following program

\[ (\hat{\theta}', \{\hat{\alpha}_i\}_{i=1}^{n})' = \arg \inf \{\gamma(\theta, \{\alpha_i\}_{i=1}^{n})', \] where \( \widehat{\Omega}_i(\hat{\theta}, \hat{\alpha}_i) \) is an estimator of the optimal weighting matrix for individual \( i \)

\[ \Omega_i = \Omega_0 + \sum_{j=1}^{\infty} (\Omega_{ji} + \Omega_{ji}'). \]

To facilitate the asymptotic analysis, in the estimation of the optimal weighting matrix we assume that \( g(z_{it}; \theta_0, \alpha_0) \) is a martingale difference sequence with respect to the sigma algebra \( \sigma(\alpha_i, z_{i,t-1}, z_{i,t-2}, ...) \), so that \( \Omega_i = \Omega_0 \) and \( \widehat{\Omega}_i(\hat{\theta}, \hat{\alpha}_i) = \widehat{\Omega}_0(\hat{\theta}, \hat{\alpha}_i) \). This assumption holds in rational expectation models. We do not impose this assumption to derive the limiting distribution of the one-step FE-GMM estimator.

For the subsequent analysis of the asymptotic properties of the estimator, it is convenient to consider the concentrated or profile problem. This problem is a two-step procedure. In the first step the program is solved for the individual effects, given the value of the common parameter \( \theta \). The First Order Conditions (FOC) for this stage, reparametrized conveniently as in Newey and Smith (2004), are the following

\[ \hat{\epsilon}_i(\theta, \hat{\gamma}_i(\theta)) = - \left( \frac{\widehat{G}_{\alpha_i}(\theta, \hat{\alpha}_i(\theta)) \hat{\lambda}_i(\theta)}{\hat{\gamma}_i(\theta, \hat{\alpha}_i(\theta)) + \hat{\lambda}_i(\hat{\theta}, \hat{\alpha}_i) \hat{\lambda}_i(\theta)} \right) = 0, \ (i = 1, ..., n), \]
where $\lambda_i$ is a $d_\gamma$-vector of individual Lagrange multipliers for the moment conditions, and $\gamma_i := (\alpha_i', \lambda_i')'$ is an extended $(d_\alpha + d_\gamma)$-vector of individual effects. Then, the solutions to the previous equations are plugged into the original problem, leading to the following first order conditions for $\theta$, $\tilde{s}(\theta) = 0$, where

$$\tilde{s}(\theta) = n^{-1} \sum_{i=1}^{n} \tilde{s}_i(\theta, \tilde{\gamma}_i(\theta)) = -n^{-1} \sum_{i=1}^{n} \tilde{G}_i(\theta, \tilde{\alpha}_i(\theta))' \tilde{\lambda}_i(\theta),$$

is the profile score function for $\theta$.\footnote{In the original parametrization, the FOC can be written as}

Fixed effects estimators of smooth functions of parameters and observations are constructed using the plug-in principle, i.e. $\tilde{\zeta} = \tilde{\zeta}(\hat{\theta})$ where

$$\tilde{\zeta}(\theta) = (nT)^{-1} \sum_{i=1}^{n} \sum_{t=1}^{T} \zeta(z_{it}; \theta, \tilde{\alpha}_i(\theta)).$$

Similarly, moments of the individual effects are estimated by $\tilde{\mu} = \tilde{\mu}(\hat{\theta})$, where

$$\tilde{\mu}(\theta) = n^{-1} \sum_{i=1}^{n} \mu(\tilde{\alpha}_i(\theta)).$$

### 3.4. Asymptotic Theory for FE-GMM Estimators

In this section we analyze the properties of one-step and two-step FE-GMM estimators in large samples. We show consistency and derive the asymptotic distributions for estimators of individual effects, common parameters and other quantities of interest under sequences where both $n$ and $T$ pass to infinity with the sample size. We establish results separately for one-step and two-step estimators because the former are derived under less restrictive assumptions.

We make the following assumptions to show uniform consistency of the FE-GMM one-step estimator:

**Condition 1 (Sampling and asymptotics).** (i) For each $i$, conditional on $\alpha_i$, $z_i := \{z_{it} : 1 \leq t \leq T\}$ is a stationary mixing sequence of random vectors with strong mixing coefficients $a_i(l) = \sup_t \sup_{A \in A_i; D \in D^{i+1}} |P(A \cap D) - P(A)P(D)|$, where $A_i^t = \sigma(\alpha_i, z_{it}, z_{it-1}, ...)$ and $D_i^t = \sigma(\alpha_i, z_{it}, z_{i,t+1}, ...)$ such that $\sup_i |a_i(l)| \leq Ca^l$ for some $0 < a < 1$ and some $C > 0$; (ii) $\{(z_i, \alpha_i) : 1 \leq i \leq n\}$ are independent and identically distributed across $i$; (iii) $n, T \rightarrow \infty$ such that $n/T \rightarrow \kappa^2$, where $0 < \kappa^2 < \infty$; and (iv) $\dim [g(\cdot; \theta, \alpha_i)] = d_\gamma < \infty$.\footnote{In the original parametrization, the FOC can be written as}
For a matrix or vector $A$, let $|A|$ denote the Euclidean norm, that is $|A|^2 = \text{trace}[AA']$.

**Condition 2** (Regularity and identification). (i) The vector of moment functions $g(\cdot; \theta, \alpha) = (g_1(\cdot; \theta, \alpha), \ldots, g_d(\cdot; \theta, \alpha))'$ is continuous in $(\theta, \alpha) \in \mathcal{Y}$; (ii) the parameter space $\mathcal{Y}$ is a compact, convex subset of $\mathbb{R}^{d_\theta + d_\alpha}$; (iii) $\dim(\theta, \alpha) = d_\theta + d_\alpha \leq d_g$; (iv) there exists a function $M(z_{it})$ such that $|g_k(z_{it}; \theta, \alpha_i)| \leq M(z_{it})$, $|\partial g_k(z_{it}; \theta, \alpha_i) / \partial (\theta, \alpha_i)| \leq M(z_{it})$, for $k = 1, \ldots, d_g$, and $\sup_i E \left[ M(z_{it})^{1+\delta} \right] < \infty$ for some $\delta > 0$; and (v) there exists a deterministic sequence of symmetric finite positive definite matrices $\{W_i : 1 \leq i \leq n\}$ such that $\sup_{1 \leq i \leq n} |\hat{W}_i - W_i| \to_P 0$, and, for each $\eta > 0$

$$\inf_i \left[ Q_i^W (\theta_0, \alpha_0) - \sup_{\{(\theta, \alpha) : |(\theta, \alpha) - (\theta_0, \alpha_0)| > \eta \}} Q_i^W (\theta, \alpha) \right] > 0,$$

where

$$Q_i^W (\theta, \alpha_i) := -g_i(\theta, \alpha_i)' W_{i-1}^{-1} g_i(\theta, \alpha_i), \quad g_i(\theta, \alpha_i) := E \left[ g_i(\theta, \alpha_i) \right].$$

Conditions 1(i)-(ii) impose cross sectional independence, but allow for weak time series dependence as in Hahn and Kuersteiner (2011). Conditions 1(iii)-(iv) describe the asymptotic sequences that we consider where $T$ and $n$ grow at the same rate with the sample size, whereas the number of moments $d_g$ is fixed. Condition 2 adapts standard assumptions of the GMM literature to guarantee the identification of the parameters based on time series variation for all the individuals, see Newey and McFadden (1994). The dominance and moment conditions in 2(iv) are used to establish uniform consistency of the estimators of the individual effects.

**Theorem 1** (Uniform consistency of one-step estimators). Suppose that Conditions 1 and 2 hold. Then, for any $\eta > 0$

$$\Pr \left( \left| \hat{\theta} - \theta_0 \right| \geq \eta \right) = o(T^{-1}),$$

where $\hat{\theta} = \arg \max_{(\theta, \alpha) \in \mathcal{Y}} \frac{1}{n} \sum_{i=1}^{n} \hat{Q}_i^W (\theta, \alpha_i)$ and $\hat{Q}_i^W (\theta, \alpha_i) := -\hat{g}_i(\theta, \alpha_i)' \hat{W}_i^{-1} \hat{g}_i(\theta, \alpha_i)$. Also, for any $\eta > 0$

$$\Pr \left( \sup_{1 \leq i \leq n} |\hat{\alpha}_i - \alpha_0| \geq \eta \right) = o(T^{-1}) \quad \text{and} \quad \Pr \left( \sup_{1 \leq i \leq n} |\hat{\alpha}_i| \geq \eta \right) = o(T^{-1}),$$

where $\hat{\alpha}_i = \arg \max_{\alpha} \hat{Q}_i^W (\theta, \alpha)$ and $\hat{\alpha}_i = -\hat{W}_i^{-1} \hat{g}_i(\theta, \alpha_i)$.

Let $\Sigma_{\alpha}^W := (G_\alpha W_i^{-1} G_\alpha)' \Sigma_{\alpha} G_\alpha W_i^{-1} - W_i^{-1} G_\alpha H_{\alpha} W_i^{-1}$, $P_{\alpha}^W := W_i^{-1} - W_i^{-1} G_\alpha H_{\alpha} W_i^{-1}$, $J_{\alpha}^W := G_\alpha P_{\alpha}^W G_\alpha$, and $J_\alpha^W := E [J_{\alpha}^W]$. We use the following additional assumptions to derive the limiting distribution of the one-step estimator:

**Condition 3** (Regularity). (i) For each $i$, $(\theta_0, \alpha_0) \in \text{int } [\mathcal{Y}]$; and (ii) $J_{\alpha}^W$ is finite positive definite, and $\{G_\alpha W_i^{-1} G_\alpha : 1 \leq i \leq n\}$ is a sequence of finite positive definite matrices, where $\{W_i : 1 \leq i \leq n\}$ is the sequence of matrices of Condition 2(v).
Condition 4 (Smoothness). (i) There exists a function $M(z_{it})$ such that, for $k = 1, \ldots, d$, 
$$
|\partial^{d_1+d_2} g_k(z_{it}; \theta, \alpha_i) / \partial \theta^{d_1} \partial \alpha_i^{d_2}| \leq M(z_{it}), \quad 0 \leq d_1 + d_2 \leq 1, \ldots, 5,
$$
and $\sup_i E \left[ M(z_{it})^5 (d_1+d_2+\delta)/(1-10\omega)+\delta \right] < \infty$, for some $\delta > 0$ and $0 < \omega < 1/10$; and (ii) there exists $\xi_i(z_{it})$ such that $\hat{W}_i = W_i + \sum_{t=1}^T \xi_i(z_{it})/T + R_{W_i}^i/T$, where $\max_i |R_{W_i}^i| = o_p(T^{1/2})$, $E[\xi_i(z_{it})] = 0$, and $\sup_i E[|\xi_i(z_{it})|^{20}/(1-10\omega)+\delta] < \infty$, for some $\delta > 0$ and $0 < \omega < 1/10$.

Condition 3 is the panel data analog to the standard asymptotic normality condition for GMM with cross sectional data, see Newey and McFadden (1994). Condition 4 is similar to Condition 4 in Hahn and Kuersteiner (2011), and guarantees the existence of higher order expansions for the GMM estimators and the uniform convergence of their remainder terms.

Let $G_{\alpha_i} := (G'_{\alpha_{i,1}}, \ldots, G'_{\alpha_{i,q}})'$, where $G_{\alpha_{i,j}} = E[\partial G_{\alpha_i}(z_{it})/\partial \alpha_{i,j}]$, and $G_{\theta_{i,j}} := (G'_{\theta_{i,1}}, \ldots, G'_{\theta_{i,q}})'$, where $G_{\theta_{i,j}} = E[\partial G_{\theta_i}(z_{it})/\partial \alpha_{i,j}]$. The symbol $\otimes$ denotes kronecker product of matrices, $I_d$ a $d \times d$ identity matrix, $e_j$ a unitary $d_g$-vector with 1 in row $j$, and $P_{W_{i,j}}^i$ the $j$-th column of $P_{W_{i}}^i$. Recall that the extended individual effect is $\gamma_i = (\alpha_i', \lambda_i')$.

Lemma 1 (Asymptotic expansion for one-step estimators of individual effects). Under Conditions 1, 2, 3, and 4,

$$
\sqrt{T}(\hat{\gamma}_{i0} - \gamma_{i0}) = \tilde{\psi}_{t}^W + T^{-1/2} Q_{11}^W + T^{-1} R_{2t}^W,
$$

where $\tilde{\gamma}_{i0} := \tilde{\gamma}_{i}(\theta_0)$,

$$
\tilde{\psi}_{t}^W = - \begin{pmatrix} H_{W_{i}}^W \\ P_{W_{i}}^W \end{pmatrix} T^{-1/2} \sum_{t=1}^T g(z_{it}) d \rightarrow N(0, V_{i}^W),
$$

$$
n^{-1/2} \sum_{i=1}^n \tilde{\psi}_{i}^W \rightarrow N(0, E[V_{i}^W]), \quad n^{-1} \sum_{i=1}^n Q_{11}^W \rightarrow E[B_{W}^i], \quad B_{W}^i = B_{W,I}^i + B_{W,G}^i + B_{W,IS}^i,
$$

$$
\sup_{1 \leq i \leq n} R_{2t}^W = o_P(\sqrt{T}), \text{ for } \forall \lambda_{i},
$$

where

$$
V_{i}^W = \begin{pmatrix} H_{W_{i}}^W \\ P_{W_{i}}^W \end{pmatrix} \Omega_i \begin{pmatrix} H_{W_{i}}^W, P_{W_{i}}^W \end{pmatrix};
$$

$$
B_{W,I}^i = \begin{pmatrix} B_{W,I}^i \\ B_{W,G}^i \end{pmatrix} = \begin{pmatrix} H_{W_{i}}^W \\ P_{W_{i}}^W \end{pmatrix} \left( \sum_{j=-\infty}^{\infty} E[G_{\alpha_i}(z_{it}) H_{W_{i}}^W g(z_{i,t-j})] - \sum_{j=1}^{d_a} G_{\alpha_{i,j}} H_{W_{i}}^W \Omega_i H_{W_{i}}^W/2 \right),
$$

$$
B_{W,G}^i = \begin{pmatrix} B_{W,G}^i \\ B_{W,G}^i \end{pmatrix} = \begin{pmatrix} -\sum_{j=1}^{d_a} E[G_{\alpha_i}(z_{it}) P_{W_{i}}^W g(z_{i,t-j})] \\ \sum_{j=-\infty}^{\infty} E[G_{\alpha_i}(z_{it}) P_{W_{i}}^W g(z_{i,t-j})] \end{pmatrix} - \sum_{j=1}^{d_a} G_{\alpha_{i,j}}^W \Omega_i H_{W_{i}}^W/2 + \sum_{j=1}^{d_a} G_{\alpha_{i,j}}(I_{d_a} \otimes e_j) H_{W_{i}}^W \Omega_i P_{W_{i}}^W/2,
$$

$$
B_{W,IS}^i = \begin{pmatrix} B_{W,IS}^i \\ B_{W,IS}^i \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{d_a} E[G_{\alpha_{i,j}} P_{W_{i}}^W g(z_{i,t-j})] \\ \sum_{j=1}^{d_a} E[G_{\alpha_{i,j}} P_{W_{i}}^W g(z_{i,t-j})] \end{pmatrix} + \begin{pmatrix} H_{W_{i}}^W \\ P_{W_{i}}^W \end{pmatrix} \sum_{j=-\infty}^{\infty} E[\xi_i(z_{it}) P_{W_{i}}^W g(z_{i,t-j})].
$$

62
Theorem 2 (Limit distribution of one-step estimators of common parameters). Under Conditions 1, 2, 3 and 4,

\[ \sqrt{n T} (\hat{\theta} - \theta_0) \xrightarrow{d} - (J_s^W)^{-1} N \left( \kappa B_s^W, V_s^W \right), \]

where

\[ J_s^W = \bar{E} \left[ G_{\theta_i} P_{\alpha_i} G_{\theta_i} \right], V_s^W = \bar{E} \left[ G_{\theta_i} P_{\alpha_i} \Omega_i P_{\alpha_i} G_{\theta_i} \right], B_s^W = \bar{E} \left[ B_{st,i}^{W,B} + B_{st,i}^{W,C} + B_{st,i}^{W,V} \right], \]

and

\[ B_{st,i}^{W,B} = -G_{\theta_i} \left( B_{\lambda_i}^{W,I} + B_{\lambda_i}^{W,G} + B_{\lambda_i}^{W,1S} \right), B_{st,i}^{W,C} = \sum_{j=1}^{\infty} E \left[ G_{\theta_i}(z_{it})' P_{\alpha_i} g_{t}(z_{it,j}) \right], \]

\[ B_{st,i}^{W,V} = -\sum_{j=1}^{d_\alpha} G_{\theta_{\alpha,i,j}} P_{\alpha_i} H_{\alpha_i} P_{\alpha_i} / 2 - \sum_{j=1}^{d_\alpha} G_{\theta_{\alpha,i,j}} (I_{\alpha_i} \otimes e_j) H_{\alpha_i} P_{\alpha_i} / 2. \]

The expressions for \( B_{\lambda_i}^{W,I}, B_{\lambda_i}^{W,G}, \) and \( B_{\lambda_i}^{W,1S} \) are given in Lemma 1.

The source of the bias is the non-zero expectation of the profile score of \( \theta \) at the true parameter value, due to the substitution of the unobserved individual effects by sample estimators. These estimators converge to their true parameter value at a rate \( \sqrt{T} \), which is slower than \( \sqrt{n T} \), the rate of convergence of the estimator of the common parameter. Intuitively, the rate for \( \bar{\gamma}_{i0} \) is \( \sqrt{T} \) because only the \( T \) observations for individual \( i \) convey information about \( \gamma_{i0} \). In nonlinear and dynamic models, the slow convergence of the estimator of the individual effect introduces bias in the estimators of the rest of parameters. The expression of this bias can be explained with an expansion of the score around the true value of the individual effects\(^8\)

\[ E \left[ s_i^W (\theta_0, \bar{\gamma}_{i0}) \right] = E \left[ s_i^W \right] + E \left[ s_i^W \right]' E \left[ \bar{\gamma}_{i0} - \gamma_{i0} \right] + E \left[ (s_i^W - E \left[ s_i^W \right])' (\bar{\gamma}_{i0} - \gamma_{i0}) \right] \]

\[ + E \sum_{j=1}^{d_\alpha + d_\gamma} (\bar{\gamma}_{i0,j} - \gamma_{i0,j}) E \left[ s_{i,j}^W \right] (\bar{\gamma}_{i0} - \gamma_{i0}) \] /2 + o(T^{-1})

\[ = 0 + B_s^{W,B} / T + B_s^{W,C} / T + B_s^{W,V} / T + o(T^{-1}). \]

This expression shows that the bias has the same three components as in the MLE case, see Hahn and Newey (2004). The first component, \( B_s^{W,B} \), comes from the higher-order bias of the estimator of the individual effects. The second component, \( B_s^{W,C} \), is a correlation term and is present because individual effects and common parameters are estimated using the same

---

\(^8\)Using the notation introduced in Section 3, the score is

\[ s_i^W (\theta_0) = n^{-1} \sum_{t=1}^{n} s_i^W (\theta_0, \bar{\gamma}_{i0}) = -n^{-1} \sum_{t=1}^{n} \hat{G}_{\theta_i}(\theta_0, \bar{\alpha}_{i0})' \hat{\lambda}_{i0}, \]

where \( \bar{\gamma}_{i0} = (\bar{\alpha}_{i0}, \bar{\lambda}_{i0}) \) is the solution to

\[ \hat{\lambda}_{i}^W (\theta_0, \bar{\gamma}_{i0}) = - \left( \frac{\hat{G}_{\alpha_i}(\theta_0, \bar{\alpha}_{i0})' \bar{\lambda}_{i0}}{\hat{G}_{\theta_i}(\theta_0, \bar{\alpha}_{i0}) + W_i \bar{\lambda}_{i0}} \right) = 0. \]
observations. The third component, \( B_s^{VW} \), is a variance term. The bias of the individual effects, \( B_s^{WB} \), can be further decomposed in three terms corresponding to the asymptotic bias for a GMM estimator with the optimal score, \( B_s^{WI} \), when \( W \) is used as the weighting function; the bias arising from estimation of \( G_{\alpha_i} \), \( B_s^{WG} \); and the bias arising from not using an optimal weighting matrix, \( B_s^{W1S} \).

We use the following condition to show the consistency of the two-step FE-GMM estimator:

**Condition 5** (Smoothness, regularity, and martingale). (i) There exists a function \( M \left( z_{it} \right) \) such that \( \left| g_k \left( z_{it}; \theta, \alpha_i \right) \right| \leq M \left( z_{it} \right), \left| \partial g_k \left( z_{it}; \theta, \alpha_i \right) / \partial (\theta, \alpha_i) \right| \leq M \left( z_{it} \right) \) for \( k = 1, \ldots, d_g \), and
\[
\sup_i \mathbb{E} \left[ M \left( z_{it} \right)^{10(d_\theta+d_\alpha+6)/(1-10v)+\delta} \right] < \infty, \text{ for some } \delta > 0 \text{ and } 0 < v < 1/10; \text{ (ii) } \{ \Omega_i : 1 \leq i \leq n \} \text{ is a sequence of finite positive definite matrices; and (iii) for each } i, g(z_{it}; \theta, \alpha_i) \text{ is a martingale difference sequence with respect to } \sigma(\alpha_i, z_{i,t-1}, z_{i,t-2}, \ldots).}
\]

Conditions 5(i)-(ii) are used to establish the uniform consistency of the estimators of the individual weighting matrices. Condition 5(iii) is convenient to simplify the expressions of the optimal weighting matrices. It holds, for example, in rational expectation models that commonly arise in economic applications.

**Theorem 3** (Uniform consistency of two-step estimators). Suppose that Conditions 1, 2, 3 and 5 hold. Then, for any \( \eta > 0 \)
\[
\Pr \left( \left| \hat{\theta} - \theta_0 \right| \geq \eta \right) = o \left( T^{-1} \right),
\]
where \( \hat{\theta} = \arg \max_{(\theta', \alpha_i')_{i=1}^n} \sum_{i=1}^n \hat{Q}_i^0 \left( \theta, \alpha_i \right) \) and \( \hat{Q}_i^0 \left( \theta, \alpha_i \right) := -\hat{g}_i \left( \theta, \alpha_i \right) \hat{\Omega}_i \left( \hat{\theta}, \hat{\alpha}_i \right)^{-1} \hat{g}_i \left( \theta, \alpha_i \right). \) Also, for any \( \eta > 0 \)
\[
\Pr \left( \sup_{1 \leq i \leq n} \left| \hat{\alpha}_i - \alpha_0 \right| \geq \eta \right) = o \left( T^{-1} \right) \text{ and } \Pr \left( \sup_{1 \leq i \leq n} \left| \hat{\lambda}_i \right| \geq \eta \right) = o \left( T^{-1} \right),
\]
where \( \hat{\alpha}_i = \arg \max_{\alpha} \hat{Q}_i^0 \left( \hat{\theta}, \alpha \right) \) and \( \hat{g}_i \left( \hat{\theta}, \hat{\alpha}_i \right) + \hat{\Omega}_i \left( \hat{\theta}, \hat{\alpha}_i \right) \hat{\lambda}_i = 0. \)

We replace Condition 4 by the following condition to obtain the limit distribution of the two-step estimator:

**Condition 6** (Smoothness). There exists some \( M \left( z_{it} \right) \) such that, for \( k = 1, \ldots, d_g \)
\[
\left| \partial^{d_1+d_2} g_k \left( z_{it}; \theta, \alpha_i \right) / \partial \theta^{d_1} \partial \alpha_i^{d_2} \right| \leq M \left( z_{it} \right) \quad 0 \leq d_1 + d_2 \leq 1, \ldots, 5,
\]
and
\[
\sup_i \mathbb{E} \left[ M \left( z_{it} \right)^{10(d_\theta+d_\alpha+6)/(1-10v)+\delta} \right] < \infty, \text{ for some } \delta > 0 \text{ and } 0 < v < 1/10.
\]

Condition 6 guarantees the existence of higher order expansions for the estimators of the weighting matrices and uniform convergence of their remainder terms. Conditions 5 and 6 are stronger versions of conditions 2(iv), 2(v) and 4. They are presented separately because they are only needed when there is a first stage where the weighting matrices are estimated.
Let $\Sigma_{\alpha_i} := (G'_{\alpha_i} \Omega^{-1}_i G_{\alpha_i})^{-1}$, $H_{\alpha_i} := \Sigma_{\alpha_i} G'_{\alpha_i} \Omega^{-1}_i$, and $P_{\alpha_i} := \Omega^{-1}_i - \Omega^{-1}_i G_{\alpha_i} H_{\alpha_i}$.

**Lemma 2** (Asymptotic expansion for two-step estimators of individual effects). Under the Conditions 1, 2, 3, 4, and 5,

\[
\sqrt{T}(\hat{\gamma}_{i0} - \gamma_{i0}) = \tilde{\psi}_i + T^{-1/2}B_{\gamma_i} + T^{-1}R_{2i},
\]

where $\hat{\gamma}_{i0} := \gamma_i(\theta_0)$,

\[
\tilde{\psi}_i = - \left( \begin{array}{c} H_{\alpha_i} \\ P_{\alpha_i} \end{array} \right) T^{-1/2} \sum_{t=1}^{T} g(z_{it}) \xrightarrow{d} N(0, V_i),
\]

\[
n^{-1/2} \sum_{i=1}^{n} \tilde{\psi}_i \xrightarrow{d} N(0, \tilde{E}[V_i]), B_{\gamma_i} = B_{\gamma_i}^I + B_{\gamma_i}^G + B_{\gamma_i}^\Omega + B_{\gamma_i}^W, \sup_{1 \leq i \leq n} R_{2i} = O_p(\sqrt{T}), \text{ with, for } \Omega_{\alpha_i,j} = \partial \Omega_{\alpha_i}/\partial \alpha_{\alpha,j},
\]

\[
V_i = \text{diag}(\Sigma_{\alpha_i}, P_{\alpha_i}),
\]

\[
B_{\gamma_i}^I = \left( \begin{array}{c} B_{\alpha_i}^I \\ B_{\lambda_i}^I \end{array} \right) = \left( \begin{array}{c} H_{\alpha_i} \\ P_{\alpha_i} \end{array} \right) \left( \sum_{j=1}^{d_{\alpha_i}} \Sigma_{\alpha_i,j}/2 + E[G_{\alpha_i}(z_{it})H_{\alpha_i}g(z_{it-j})] \right),
\]

\[
B_{\gamma_i}^G = \left( \begin{array}{c} B_{\alpha_i}^G \\ B_{\lambda_i}^G \end{array} \right) = \left( \begin{array}{c} H_{\alpha_i} \\ P_{\alpha_i} \end{array} \right) \left( \sum_{j=1}^{\infty} E[G_{\alpha_i}(z_{it})P_{\alpha_i}g(z_{it-j})] \right),
\]

\[
B_{\gamma_i}^\Omega = \left( \begin{array}{c} B_{\alpha_i}^\Omega \\ B_{\lambda_i}^\Omega \end{array} \right) = \left( \begin{array}{c} H_{\alpha_i} \\ P_{\alpha_i} \end{array} \right) \left( \sum_{j=1}^{\infty} E[g(z_{it})g(z_{it})P_{\alpha_i}g(z_{it-j})] \right),
\]

\[
B_{\gamma_i}^W = \left( \begin{array}{c} B_{\alpha_i}^W \\ B_{\lambda_i}^W \end{array} \right) = \left( \begin{array}{c} H_{\alpha_i} \\ P_{\alpha_i} \end{array} \right) \sum_{j=1}^{d_{\alpha_i}} \Sigma_{\alpha_i,j} \left( H_{\alpha_i,j} + H_{\lambda_i,j} \right).
\]

**Theorem 4** (Limit distribution for two-step estimators of common parameters). Under the Conditions 1, 2, 3, 4, 5 and 6,

\[
\sqrt{nT}(\hat{\theta} - \theta_0) \xrightarrow{d} -J_s^{-1}N(\kappa B_s, J_s),
\]

where $J_s = \tilde{E} [G'_{\theta_i}P_{\alpha_i}G_{\theta_i}^2]$, $B_s = \tilde{E} [B_{s_i}^B + B_{s_i}^G]$, $B_{s_i}^B = -G'_{\theta_i} [B_{\lambda_i}^B + B_{\lambda_i}^G + B_{\lambda_i}^\Omega + B_{\lambda_i}^W]$, $B_{s_i}^G = \sum_{j=1}^{\infty} E[G_{\theta_i}(z_{it})P_{\alpha_i}g(z_{it-j})]$. The expressions for $B_{\alpha_i}^B$, $B_{\alpha_i}^G$, $B_{\alpha_i}^\Omega$, and $B_{\alpha_i}^W$ are given in Lemma 2.

Theorem 4 establishes that one iteration of the GMM procedure not only improves asymptotic efficiency by reducing the variance of the influence function, but also removes the variance and non-optimal weighting matrices components from the bias. The higher-order bias of the estimator of the individual effects, $B_{\lambda_i}$, now has four components, as in Newey and Smith (2004). These components correspond to the asymptotic bias for a GMM estimator with the optimal score, $B_{\lambda_i}^B$; the bias arising from estimation of $G_{\alpha_i}$, $B_{\lambda_i}^G$; the bias arising from estimation of $\Omega_i$, $B_{\lambda_i}^\Omega$; and the bias arising from the choice of the preliminary first step estimator, $B_{\lambda_i}^W$. An additional iteration of the GMM estimator removes the term $B_{\lambda_i}^W$. 

65
The general procedure for deriving the asymptotic distribution of the FE-GMM estimators consists of several expansions. First, we derive higher-order asymptotic expansions for the estimators of the individual effects, with the common parameter fixed at its true value $\theta_0$. Next, we obtain the asymptotic distribution for the profile score of the common parameter at $\theta_0$ using the expansions of the estimators of the individual effects. Finally, we derive the asymptotic distribution of estimator for the common parameter multiplying the asymptotic distribution of the score by the limit profile Jacobian matrix. This procedure is detailed in the online appendix Fernández-Val and Lee (2012). Here we characterize the asymptotic bias in a linear correlated random coefficient model with endogenous regressors. Motivated by the numerical and empirical examples that follow, we consider a model where only the variables with common parameter are endogenous and allow for the moment conditions not to be martingale difference sequences.

Example: Correlated random coefficient model with endogenous regressors. We consider a simplified version of the models in the empirical and numerical examples. The notation is the same as in the theorems discussed above. The moment condition is

$$g(z_{it}; \theta, \alpha_i) = w_{it}(y_{it} - x'_{1it}\alpha_i - x'_{2it}\theta),$$

where $w_{it} = (x'_{1it}, w'_{2it})'$ and $z_{it} = (x'_{1it}, x'_{2it}, w'_{2it}, y_{it})'$. That is, only the regressors with common coefficients are endogenous. Let $\epsilon_{it} = y_{it} - x'_{1it}\alpha_i - x'_{2it}\theta_0$. To simplify the expressions for the bias, we assume that $\epsilon_{it} | w_i, \alpha_i \sim i.i.d.(0, \sigma^2)$ and $E[x_{2it}\epsilon_{i,t-j} | w_i, \alpha_i] = E[x_{2it}\epsilon_{i,t-j}]$, for $w_i = (w_{i1}, ..., w_{iT})'$ and $j \in \{0, \pm 1, \ldots\}$. Under these conditions, the optimal weighted matrices are proportional to $E[w_{it}w'_{it}]$, which do not depend on $\theta_0$ and $\alpha_0$. We can therefore obtain the optimal GMM estimator in one step using the sample averages $T^{-1} \sum_{t=1}^T w_{it}w'_{it}$ to estimate the optimal weighting matrices.

In this model, it is straightforward to see that the estimators of the individual effects have no bias, that is $B_{\gamma_i}^{W,I} = B_{\gamma_i}^{W,C} = B_{\gamma_i}^{W,IS} = 0$. By linearity of the first order conditions in $\theta$ and $\alpha_i$, $B_{\alpha_i}^{W,V} = 0$. The only source of bias is the correlation between the estimators of $\theta$ and $\alpha_i$. After some straightforward but tedious algebra, this bias simplifies to

$$B_{\alpha_i}^{W,C} = - (d_\alpha - d_0) \sum_{j=-\infty}^{\infty} E[x_{2it}\epsilon_{i,t-j}].$$

For the limit Jacobian, we find

$$J_s^W = \hat{E} \left\{ E[\bar{x}_{2it}\bar{w}'_{2it}]E[\bar{w}_{2it}\bar{w}'_{2it}]^{-1}E[\bar{w}_{2it}\bar{x}'_{2it}] \right\},$$

where variables with tilde indicate residuals of population linear projections of the corresponding variable on $x_{1it}$, for example $\bar{x}_{2it} = x_{2it} - E[x_{2it}x'_{1it}]E[x_{1it}x'_{1it}]^{-1}x_{1it}$. The expression
In random coefficient models the ultimate quantities of interest are often functions of the data, model parameters and individual effects. The following corollaries characterize the asymptotic distributions of the fixed effects estimators of these quantities. The first corollary applies to averages of functions of the data and individual effects such as average partial effects and average derivatives in nonlinear models, and average elasticities in linear models with variables in levels. Section 6 gives an example of these elasticities. The second corollary applies to averages of smooth functions of the individual effects including means, variances and other moments of the distribution of these effects. Sections 2 and 6 give examples of these functions. We state the results only for estimators constructed from two-step estimators of the common parameters and individual effects. Similar results apply to estimators constructed from one-step estimators. Both corollaries follow from Lemma 2 and Theorem 4 by the delta method.

**Corollary 1** (Asymptotic distribution for fixed effects averages). Let \( \zeta(z; \theta, \alpha) \) be a twice continuously differentiable function in its second and third argument, such that \( \inf \text{Var} \{\zeta(z_t)\} > 0, \ E E[\zeta(z_t)^2] < \infty, \ E E[\alpha(z_t)]^2 < \infty, \) and \( E E[\theta(z_t)]^2 < \infty, \) where the subscripts on \( \zeta \) denote partial derivatives. Then, under the conditions of Theorem 4, for some deterministic sequence \( r_{nT} \to \infty \) such that \( r_{nT} = O(\sqrt{nT}), \)

\[
W_T \sim N(0, V_T),
\]

where \( V_T \) is defined by

\[
B_T = E E \left[ -\sum_{j=0}^{\infty} \zeta\alpha(z_{it})'H\alpha,g(z_{it},t-j) + \zeta\alpha(z_{it})'B\alpha + \sum_{j=1}^{d\alpha} \zeta\alpha\alpha_{j,r}(z_{it})\Sigma\alpha_{j}/2 - \zeta\theta(z_{it})'J_s^{-1}B_s \right],
\]

for \( B\alpha = B^L + B^G + B^D + B^W, \) and for \( r^2 = \lim_{nT \to \infty} r_{nT}^2/(nT), \)

\[
V_T = E \left\{ r^2 E \left[ \zeta\alpha(z_{it})\Sigma\alpha\zeta\alpha(z_{it}) + \zeta\theta(z_{it})'J_s^{-1}\zeta\theta(z_{it}) \right] + \lim_{nT \to \infty} \frac{r_{nT}^2}{nT} \left[ \left( \frac{1}{T} \sum_{t=1}^{T} \left( \zeta(z_{it}) - \zeta \right) \right) \right] \right\}.
\]

**Corollary 2** (Asymptotic distribution for smooth functions of individual effects). Let \( \mu(\alpha) \) be a twice differentiable function such that \( E[\mu(\alpha^2)] < \infty \) and \( E[\alpha(\alpha^2)] < \infty, \) where the subscripts on \( \mu \) denote partial derivatives. Then, under the conditions of Theorem 4

\[
\sqrt{n}(\hat{\mu} - \mu) \to N(\kappa B\mu, V\mu),
\]

67
where $\mu = \bar{E}[\mu(\alpha_{i0})]$

$$B_\mu = \bar{E} \left[ \mu_{\alpha_i}(\alpha_{i0})' B_{\alpha_i} + \sum_{j=1}^{d_{\alpha}} \mu_{\alpha_{i,j}}(\alpha_{i0})' \Sigma_{\alpha_i}/2 \right],$$

for $B_{\alpha_i} = B^I_{\alpha_i} + B^S_{\alpha_i} + B^O_{\alpha_i} + B^W_{\alpha_i}$, and $V_\mu = \bar{E}[(\mu(\alpha_{i0}) - \mu)^2].$

The convergence rate $r_{nT}$ in Corollary 1 depends on the function $\zeta(z; \theta, \alpha_i)$. For example, $r_{nT} = \sqrt{nT}$ for functions that do not depend on $\alpha_i$ such as $\zeta(z; \theta, \alpha_i) = c'\theta$, where $c$ is a known $d_\theta$ vector. In general, $r_{nT} = \sqrt{n}$ for functions that depend on $\alpha_i$. In this case $r^2 = 0$ and the first two terms of $V_\zeta$ drop out. Corollary 2 is an important special case of Corollary 1. We present it separately because the asymptotic bias and variance have simplified expressions.

3.5. BIAS CORRECTIONS

The FE-GMM estimators of common parameters, while consistent, have bias in the asymptotic distributions under sequences where $n$ and $T$ grow at the same rate. These sequences provide a good approximation to the finite sample behavior of the estimators in empirical applications where the time dimension is moderately large. The presence of bias invalidates any asymptotic inference because the bias is of the same order as the variance. In this section we describe bias correction methods to adjust the asymptotic distribution of the FE-GMM estimators of the common parameter and smooth functions of the data, model parameters and individual effects. All the corrections considered are analytical. Alternative corrections based on variations of Jackknife can be implemented using the approaches described in Hahn and Newey (2004) and Dhaene and Jochmans (2010).9

We consider three analytical methods that differ in whether the bias is corrected from the estimator or from the first order conditions, and in whether the correction is one-step or iterated for methods that correct the bias from the estimator. All these methods reduce the order of the asymptotic bias without increasing the asymptotic variance. They are based on analytical estimators of the bias of the profile score $B_s$ and the profile Jacobian matrix $J_s$. Since these quantities include cross sectional and time series means $\bar{E}$ and $E$ evaluated at the true parameter values for the common parameter and individual effects, they are estimated by the corresponding cross sectional and time series averages evaluated at the FE-GMM estimates. Thus, for any function of the data, common parameter and individual effects $f_{u_t}(\theta, \alpha_i)$, let $\hat{f}_{\mu_t}(\theta) = f_{u_t}(\theta, \bar{\alpha}_i(\theta)), \hat{f}_t(\theta) = \bar{E}[\hat{f}_{u_t}(\theta)] = T^{-1} \sum_{t=1}^{T} \hat{f}_{u_t}(\theta)$ and $\bar{f}(\theta) = \bar{E}[\hat{f}_t(\theta)] = n^{-1} \sum_{i=1}^{n} \hat{f}_t(\theta)$. Next, define $\hat{\Sigma}_{\alpha_i}(\theta) = [\hat{G}_{\alpha_i}(\theta)]^{-1} \hat{\Omega}_{\alpha_i}^{-1}[\hat{G}_{\alpha_i}(\theta)]^{-1}, \hat{H}_{\alpha_i}(\theta) = \hat{\Sigma}_{\alpha_i}(\theta) \hat{G}_{\alpha_i}(\theta)^{-1} \hat{\Omega}_{\alpha_i}^{-1}.$

9Hahn, Kuersteiner and Newey (2004) show that analytical, Bootstrap, and Jackknife bias corrections methods are asymptotically equivalent up to third order for MLE. We conjecture that the same result applies to GMM estimators, but the proof is beyond the scope of this paper.
and \( \hat{P}_{\alpha}(\theta) = \Omega_{i}^{-1} \hat{G}_{\alpha_{i}}(\theta) \hat{H}_{a_{i}}(\theta) \). To simplify the presentation, we only give explicit formulas for FE-GMM three-step estimators in the main text. We give the expressions for one and two-step estimators in the Supplementary Appendix. Let

\[
\hat{B}(\theta) = -\hat{J}_{s}(\theta)^{-1} \hat{B}_{s}(\theta), \quad \hat{B}_{s}(\theta) = \hat{E}[\hat{B}_{a_{i}}^{P}(\theta) + \hat{B}_{a_{i}}^{C}(\theta)], \quad \hat{J}_{s}(\theta) = \hat{E}[\hat{G}_{\theta_{i}}(\theta)^{\top} \hat{P}_{\alpha_{i}}(\theta) \hat{G}_{\theta_{i}}(\theta)],
\]

where \( \hat{B}_{a_{i}}^{P}(\theta) = -\hat{G}_{\theta_{i}}(\theta)^{\top} \left[ \hat{B}_{a_{i}}^{H}(\theta) + \hat{B}_{a_{i}}^{G}(\theta) + \hat{B}_{a_{i}}^{\Omega}(\theta) + \hat{B}_{a_{i}}^{W}(\theta) \right] \),

\[
\hat{B}_{a_{i}}^{H}(\theta) = \hat{H}_{a_{i}}(\theta)^{\top} \sum_{t=0}^{T} \sum_{t=j+1}^{T} \hat{G}_{\alpha_{i}}(\theta)^{\top} \hat{P}_{\alpha_{i}}(\theta) \hat{g}_{i,t-j}(\theta),
\]

\[
\hat{B}_{a_{i}}^{G}(\theta) = \hat{G}_{\alpha_{i}}(\theta)^{\top} \sum_{t=0}^{T} \sum_{t=j+1}^{T} \hat{G}_{\alpha_{i}}(\theta)^{\top} \hat{P}_{\alpha_{i}}(\theta) \hat{g}_{i,t-j}(\theta),
\]

\[
\hat{B}_{a_{i}}^{\Omega}(\theta) = \hat{P}_{\alpha_{i}}(\theta)^{\top} \sum_{t=0}^{T} \sum_{t=j+1}^{T} \hat{g}_{i,t}(\theta) \hat{g}_{i,t-j}(\theta)^{\top} \hat{P}_{\alpha_{i}}(\theta) \hat{g}_{i,t-j}(\theta),
\]

and \( \hat{B}_{a_{i}}^{W}(\theta) = T^{-1} \sum_{j=0}^{T} \sum_{t=j+1}^{T} \hat{G}_{\alpha_{i}}(\theta)^{\top} \hat{P}_{\alpha_{i}}(\theta) \hat{g}_{i,t-j}(\theta) \). In the previous expressions, the spectral time series averages that involve an infinite number of terms are trimmed. The trimming parameter \( \ell \) is a positive bandwidth that need to be chosen such that \( \ell \to \infty \) and \( \ell/T \to 0 \) as \( T \to \infty \) (Hahn and Kuersteiner, 2011).

The one-step correction of the estimator subtracts an estimator of the expression of the asymptotic bias from the estimator of the common parameter. Using the expressions defined above evaluated at \( \hat{\theta} \), the bias-corrected estimator is

\[
\hat{\theta}^{BC} = \hat{\theta} - \hat{B}(\hat{\theta})/T.
\]

This bias correction is straightforward to implement because it only requires one optimization. The iterated correction is equivalent to solving the nonlinear equation

\[
\hat{\theta}^{IBC} = \hat{\theta} - \hat{B}(\hat{\theta}^{IBC})/T.
\]

When \( \theta + \hat{B}(\theta) \) is invertible in \( \theta \), it is possible to obtain a closed-form solution to the previous equation.\(^{10} \) Otherwise, an iterative procedure is needed. The score bias-corrected estimator is the solution to the following estimating equation

\[
\hat{s}(\hat{\theta}^{SBC}) - \hat{B}_{s}(\hat{\theta}^{SBC})/T = 0.
\]

This procedure, while computationally more intensive, has the attractive feature that both estimator and bias are obtained simultaneously. Hahn and Newey (2004) show that fully iterated bias-corrected estimators solve approximated bias-corrected first order conditions. IBC and SBC are equivalent if the first order conditions are linear in \( \theta \).

\(^{10} \)See MacKinnon and Smith (1998) for a comparison of one-step and iterated bias correction methods.
Example: Correlated random coefficient model with endogenous regressors. The previous methods can be illustrated in the correlated random coefficient model example in Section 4. Here, the fixed effects GMM estimators have closed forms:

$$\tilde{\alpha}_i(\theta) = \left( \sum_{t=1}^{T} x_{1it} x'_{1it} \right)^{-1} \sum_{t=1}^{T} x_{1it} (y_{it} - x'_{2it} \theta),$$

and

$$\tilde{\theta} = (J_s)^{-1} \sum_{i=1}^{n} \left[ \sum_{t=1}^{T} \tilde{x}_{2it} \tilde{w}'_{2it} \left( \sum_{t=1}^{T} \tilde{w}_{2it} \tilde{w}'_{2it} \right)^{-1} \sum_{t=1}^{T} \tilde{w}_{2it} \tilde{y}_{it} \right],$$

where $J_s = \sum_{i=1}^{n} \sum_{t=1}^{T} \tilde{x}_{2it} \tilde{x}'_{2it} \left( \sum_{t=1}^{T} \tilde{w}_{2it} \tilde{w}'_{2it} \right)^{-1} \sum_{t=1}^{T} \tilde{w}_{2it} \tilde{x}'_{2it}$, and variables with tilde now indicate residuals of sample linear projections of the corresponding variable on $x_{1it}$, for example $\tilde{x}_{2it} = x_{2it} - \sum_{t=1}^{T} x_{2it} x'_{1it} (\sum_{t=1}^{T} x_{1it} x'_{1it})^{-1} x_{1it}$.

We can estimate the bias of $\tilde{\theta}$ from the analytic formula in expression (4.3) replacing population by sample moments and $\theta_0$ by $\tilde{\theta}$, and trimming the number of terms in the spectral expectation,

$$\hat{B}(\tilde{\theta}) = -(d_g - d_\alpha) (J_s)^{-1} \sum_{i=1}^{n} \sum_{t=1}^{T} \sum_{t'=\max(1,t+1)}^{\min(T,T+j)} \tilde{x}_{2it} (\tilde{y}_{i,t-j} - \tilde{x}'_{2i,t-j} \tilde{\theta}).$$

The one-step bias corrected estimates of the common parameter $\theta$ and the average of the individual parameter $\alpha := E[\alpha_i]$ are

$$\hat{\theta}^{BC} = \tilde{\theta} - \hat{B}(\tilde{\theta}) / T, \quad \hat{\alpha}^{BC} = n^{-1} \sum_{i=1}^{n} \hat{\alpha}_i(\hat{\theta}^{BC}).$$

The iterated bias correction estimator can be derived analytically by solving

$$\hat{\theta}^{IBC} = \tilde{\theta} - \hat{B}(\hat{\theta}^{BC}) / T,$$

which has closed-form solution

$$\hat{\theta}^{IBC} = \left[ I_{d_\alpha} + (d_g - d_\alpha) (J_s)^{-1} \sum_{i=1}^{n} \sum_{t=1}^{T} \sum_{t'=\max(1,t+1)}^{\min(T,T+j)} \tilde{x}_{2it} \tilde{x}'_{2i,t-j} / (nT^2) \right]^{-1} \times \left[ \tilde{\theta} + (d_g - d_\alpha) (J_s)^{-1} \sum_{i=1}^{n} \sum_{t=1}^{T} \sum_{t'=\max(1,t+1)}^{\min(T,T+j)} \tilde{x}_{2it} \tilde{y}_{i,t-j} / (nT^2) \right].$$

The score bias correction is the same as the iterated correction because the first order conditions are linear in $\theta$.

The bias correction methods described above yield normal asymptotic distributions centered at the true parameter value for panels where $n$ and $T$ grow at the same rate with
the sample size. This result is formally stated in Theorem 5, which establishes that all the methods are asymptotically equivalent, up to first order.

**Theorem 5** (Limit distribution of bias-corrected FE-GMM). Assume that \( \sqrt{nT}(\hat{\beta}_s(\hat{\theta}) - B_\star)/T \xrightarrow{p} 0 \) and \( \sqrt{nT}(\hat{J}_s(\hat{\theta}) - J_\star)/T \xrightarrow{p} 0 \), for some \( \hat{\theta} = \theta_0 + O_P((nT)^{-1/2}) \). Under Conditions 1, 2, 3, 4, 5 and 6, for \( C \in \{BC, SBC, IBC\} \)

\[
(5.4) \quad \sqrt{nT}(\hat{\theta}^C - \theta_0) \xrightarrow{d} N(0, J^{-1}_s),
\]

where \( \hat{\theta}^BC, \hat{\theta}^IBC \) and \( \hat{\theta}^{SBC} \) are defined in (5.1), (5.2) and (5.3), and \( J_s = E[G^\prime_\theta P_{\alpha_i} G_{\theta_i}] \).

The convergence condition for the estimators of \( B_\star \) and \( J_\star \) holds for sample analogs evaluated at the initial FE-GMM one-step or two-step estimators if the trimming sequence is chosen such that \( \ell \to \infty \) and \( \ell/T \to 0 \) as \( T \to \infty \). Theorem 5 also shows that all the bias-corrected estimators considered are first-order asymptotically efficient, since their variances achieve the semiparametric efficiency bound for the common parameters in this model, see Chamberlain (1992).

The following corollaries give bias corrected estimators for averages of the data and individual effects and for moments of the individual effects, together with the limit distributions of these estimators and consistent estimators of their asymptotic variances. To construct the corrections, we use bias corrected estimators of the common parameter. The corollaries then follow from Lemma 2 and Theorem 5 by the delta method. We use the same notation as in the estimation of the bias of the common parameters above to denote the estimators of the components of the bias and variance.

**Corollary 3** (Bias correction for fixed effects averages). Let \( \zeta(z; \theta, \alpha_i) \) be a twice continuously differentiable function in its second and third argument, such that \( \inf_\theta \text{Var}[\zeta(z_it)] > 0 \), \( E[E[\zeta(z_it)^2]] < \infty \), \( E[E[\zeta_\alpha(z_it)^2]] < \infty \), and \( E[E[\zeta_\theta(z_it)]^2] < \infty \). For \( C \in \{BC, SBC, IBC\} \), let \( \hat{\zeta}_C = \hat{\zeta}^C - \hat{B}_\zeta(\theta^C)/T \) where

\[
\hat{B}_\zeta(\theta) = \hat{E} \left[ \sum_{j=0}^{T} \frac{1}{T} \sum_{t=j+1}^{T} \tilde{\zeta}_{at}(\theta)\tilde{\psi}_{at,t-j}(\theta) + \tilde{\zeta}_{at}(\theta)\hat{B}_{\alpha_i}(\theta) + \sum_{j=1}^{d_\alpha} \tilde{\zeta}_{\alpha_{at}j}(\theta)^T\tilde{\zeta}_{\alpha_i}(\theta)/2 \right],
\]

where \( \ell \) is a positive bandwidth such that \( \ell \to \infty \) and \( \ell/T \to 0 \) as \( T \to \infty \). Then, under the conditions of Theorem 5

\[
r_{nT}(\hat{\zeta}_C - \zeta) \xrightarrow{d} N(0, V_\zeta),
\]

where \( r_{nT}, \zeta, \) and \( V_\zeta \) are defined in Corollary 1. Also, for any \( \tilde{\theta} = \theta_0 + O_P((nT)^{-1/2}) \) and \( \tilde{\zeta} = \zeta + O_P(r_{nT}^{-1}) \),

\[
\hat{V}_\zeta = \frac{r_{nT}^2}{nT} \hat{E} \left\{ \hat{E}[\tilde{\zeta}_{at}(\theta)^T\tilde{\zeta}_{at}(\theta)\tilde{\zeta}_{at}(\theta) + \tilde{\zeta}_{at}(\theta)^T\hat{J}_s(\tilde{\theta})^{-1}\tilde{\zeta}_{at}(\theta)] + T \left( \hat{E}[\tilde{\zeta}_{at}(\tilde{\theta}) - \zeta] \right)^2 \right\}
\]
is a consistent estimator for $V_c$.

**Corollary 4** (Bias correction for smooth functions of individual effects). Let $\mu(\alpha_i)$ be a twice differentiable function such that $E[\mu(\alpha_{i0})^2] < \infty$ and $E|\mu_{\alpha}(\alpha_{i0})|^2 < \infty$. For $C \in \{BC, SBC, IBC\}$, let $\hat{\mu}^C = \hat{E}[\mu(\theta^C)] - \hat{B}_\mu(\theta^C)/T$, where $\mu_i(\theta) = \mu(\hat{\alpha}_i(\theta))$, and $\hat{B}_\mu(\theta) = \hat{E}[\hat{\mu}_{\alpha_i}(\theta)'\hat{B}_{\alpha_i}(\theta)] + \sum_{j=1}^{d_u} \hat{\mu}_{\alpha_i,j}(\theta)\hat{\Sigma}_{\alpha_i}(\theta)/2$. Then, under the conditions of Theorem 5

$$\sqrt{n}(\hat{\mu}^C - \mu) \xrightarrow{d} N(0, V_c),$$

where $\mu = \hat{E}[\mu(\alpha_{i0})]$ and $V_\mu = \hat{E}[(\mu(\alpha_{i0}) - \mu)^2]$. Also, for any $\tilde{\theta} = \theta_0 + O_P((nT)^{-1/2})$ and $\mu = \mu + O_P(n^{-1/2})$,

$$\tilde{V}_\mu = \hat{E}\left[(\hat{\mu}_i(\tilde{\theta}) - \tilde{\mu})^2 + \hat{\mu}_{\alpha_i}(\tilde{\theta})\hat{\Sigma}_{\alpha_i}(\tilde{\theta})\hat{\mu}_{\alpha_i}(\theta)/T\right],$$

is a consistent estimator for $V_\mu$. The second term in (5.5) is included to improve the finite sample properties of the estimator in short panels.

### 3.6. EMPIRICAL EXAMPLE

We illustrate the new estimators with an empirical example based on the classical cigarette demand study of Becker, Grossman and Murphy (1994) (BGM hereafter). Cigarettes are addictive goods. To account for this addictive nature, early cigarette demand studies included lagged consumption as explanatory variables (e.g., Baltagi and Levin, 1986). This approach, however, ignores that rational or forward-looking consumers take into account the effect of today's consumption decision on future consumption decisions. Becker and Murphy (1988) developed a model of rational addiction where expected changes in future prices affect the current consumption. BGM empirically tested this model using a linear structural demand function based on quadratic utility assumptions. The demand function includes both future and past consumptions as determinants of current demand, and the future price affects the current demand only through the future consumption. They found that the effect of future consumption on current consumption is significant, what they took as evidence in favor of the rational model.

Most of the empirical studies in this literature use yearly state-level panel data sets. They include fixed effects to control for additive heterogeneity at the state-level and use leads and lags of cigarette prices and taxes as instruments for leads and lags of consumption. These studies, however, do not consider possible non-additive heterogeneity in price elasticities or sensitivities across states. There are multiple reasons why there may be heterogeneity in the price effects across states correlated with the price level. First, the considerable differences in income, industrial, ethnic and religious composition at inter-state level can translate into different tastes and policies toward cigarettes. Second, from the perspective of the theoretical model developed by Becker and Murphy (1988), the price effect is a function of the marginal
utility of wealth that varies across states and depends on cigarette prices. If the price
effect is heterogeneous and correlated with the price level, a fixed coefficient specification
may produce substantial bias in estimating the average elasticity of cigarette consumption
because the between variation of price is much larger than the within variation. Wangen
(2004) gives additional theoretical reasons against a fixed coefficient specification for the
demand function in this application.

We consider the following linear specification for the demand function

\[ C_{it} = \alpha_0 + \alpha_1 P_{it} + \theta_1 C_{i,t-1} + \theta_2 C_{i,t+1} + X_{it}' \delta + \epsilon_{it}, \]

where \( C_{it} \) is cigarette consumption in state \( i \) at time \( t \) measured by per capita sales in packs;
\( \alpha_0 \) is an additive state effect; \( \alpha_1 \) is a state specific price coefficient; \( P_{it} \) is the price in 1982-
1984 dollars; and \( X_{it} \) is a vector of covariates which includes income, various measures of
incentive for smuggling across states, and year dummies. We estimate the model parameters
using OLS and IV methods with both fixed coefficient for price and random coefficient for
price. The data set, consisting of an unbalanced panel of 51 U.S. states over the years 1957
to 1994, is the same as in Fenn, Antonovitz and Schroeter (2001). The set of instruments for
\( C_{i,t-1} \) and \( C_{i,t+1} \) in the IV estimators is the same as in specification 3 of BGM and includes
\( X_{it}, P_{it}, P_{it-1}, P_{it+1}, Tax_{it}, Tax_{i,t-1}, \) and \( Tax_{i,t+1} \), where \( Tax_{it} \) is the state excise tax for
cigarettes in 1982-1984 dollars.

Table 1 reports estimates of coefficients and demand elasticities. We focus on the coefficients
of the key variables, namely \( P_{it}, C_{i,t-1} \) and \( C_{i,t+1} \). Throughout the table, FC refers
to the fixed coefficient specification with \( \alpha_1 = \alpha_1 \) and RC refers to the random coefficient
specification in equation (6.1). BC and IBC refer to estimates after bias correction and iter-
ated bias correction, respectively. Demand elasticities are calculated using the expressions in
Appendix A of BGM. They are functions of \( C_{it}, P_{it}, \alpha_1, \theta_1 \) and \( \theta_2 \), linear in \( \alpha_1 \). For random
coefficient estimators, we report the mean of individual elasticities, i.e.

\[ \hat{\zeta}_h = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \zeta_h(z_{it}; \hat{\theta}, \hat{\alpha}_t), \]

where \( \zeta_h(z_{it}; \theta, \alpha_i) = \partial \log C_{it(h)}/\partial \log P_{it(h)} \) are price elasticities at different time horizons
\( h \). Standard errors for the elasticities are obtained by the delta method as described in
Corollaries 3 and 4. For bias-corrected RC estimators the standard errors use bias-corrected
estimates of \( \theta \) and \( \alpha_i \).

As BGM, we find that OLS estimates substantially differ from their IV counterparts.
IV-FC underestimates the elasticities relative to IV-RC. For example, the long-run elasticity estimate is \(-0.70\) with IV-FC, whereas it is \(-0.88\) with IV-RC. This difference is also pronounced for short-run elasticities, where the IV-RC estimates are more than 25 percent
larger than the IV-FC estimates. We observe the same pattern throughout the table for every elasticity. The bias comes from both the estimation of the common parameter $\theta_2$ and the mean of the individual specific parameter $E[\alpha_{1i}]$. The bias corrections increase the coefficient of future consumption $C_{t,t+1}$ and reduce the absolute value of the mean of the price coefficient. Moreover, they have significant impact on the estimator of dispersion of the price coefficient. The uncorrected estimates of the standard deviation are more than 20% larger than the bias corrected counterparts. In the online appendix Fernández-Val and Lee (2012), we show through a Monte-Carlo experiment calibrated to this empirical example, that the bias is generally large for dispersion parameters and the bias corrections are effective in reducing this bias. As a consequence of shrinking the estimates of the dispersion of $\alpha_{1i}$, we obtain smaller standard errors for the estimates of $E[\alpha_{1i}]$ throughout the table. In the Monte-Carlo experiment, we also find that this correction in the standard errors provides improved inference.

3.7. CONCLUSION

This paper introduces a new class of fixed effects GMM estimators for panel data models with unrestricted nonadditive heterogeneity and endogenous regressors. Bias correction methods are developed because these estimators suffer from the incidental parameters problem. Other estimators based on moment conditions, like the class of GEL estimators, can be analyzed using a similar methodology. An attractive alternative framework for estimation and inference in random coefficient models is a flexible Bayesian approach. It would be interesting to explore whether there are connections between moments of posterior distributions in the Bayesian approach and the fixed effects estimators considered in the paper. Another interesting extension would be to find bias reducing priors in the GMM framework similar to the ones characterized by Arellano and Bonhomme (2009) in the MLE framework. We leave these extensions to future research.

Bibliography


**Figure 1.** Normal approximation to the distribution of price effects using uncorrected (solid line) and bias corrected (dashed line) estimates of the mean and standard deviation of the distribution of price effects. Uncorrected estimates of the mean and standard deviation are -36 and 13, bias corrected estimates are -31 and 10.
Table 1: Estimates of Rational Addiction Model for Cigarette Demand

<table>
<thead>
<tr>
<th></th>
<th>OLS-FC</th>
<th>IV-FC</th>
<th>OLS-RC</th>
<th>IV-RC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NBC</td>
<td>BC</td>
<td>IBC</td>
<td>NBC</td>
</tr>
<tr>
<td><strong>Coefficients</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Std. Dev.) $P_t$</td>
<td>4.35</td>
<td>4.22</td>
<td>4.07</td>
<td>12.86</td>
</tr>
<tr>
<td>$C_{t-1}$</td>
<td>0.49</td>
<td>0.45</td>
<td>0.48</td>
<td>0.48</td>
</tr>
<tr>
<td></td>
<td>(0.01)</td>
<td>(0.06)</td>
<td>(0.04)</td>
<td>(0.04)</td>
</tr>
<tr>
<td>$C_{t+1}$</td>
<td>0.44</td>
<td>0.17</td>
<td>0.44</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>(0.01)</td>
<td>(0.07)</td>
<td>(0.04)</td>
<td>(0.04)</td>
</tr>
<tr>
<td><strong>Price elasticities</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Long-run</td>
<td>-1.05</td>
<td>-0.70</td>
<td>-1.30</td>
<td>-1.31</td>
</tr>
<tr>
<td></td>
<td>(0.24)</td>
<td>(0.12)</td>
<td>(0.28)</td>
<td>(0.28)</td>
</tr>
<tr>
<td>Own Price</td>
<td>-0.20</td>
<td>-0.32</td>
<td>-0.27</td>
<td>-0.27</td>
</tr>
<tr>
<td>(Anticipated)</td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.06)</td>
<td>(0.06)</td>
</tr>
<tr>
<td>Own Price</td>
<td>-0.11</td>
<td>-0.29</td>
<td>-0.15</td>
<td>-0.16</td>
</tr>
<tr>
<td>(Unanticipated)</td>
<td>(0.02)</td>
<td>(0.03)</td>
<td>(0.04)</td>
<td>(0.04)</td>
</tr>
<tr>
<td>Future Price</td>
<td>-0.07</td>
<td>-0.05</td>
<td>-0.10</td>
<td>-0.10</td>
</tr>
<tr>
<td>(Unanticipated)</td>
<td>(0.01)</td>
<td>(0.03)</td>
<td>(0.02)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>Past Price</td>
<td>-0.08</td>
<td>-0.14</td>
<td>-0.11</td>
<td>-0.11</td>
</tr>
<tr>
<td>(Unanticipated)</td>
<td>(0.01)</td>
<td>(0.02)</td>
<td>(0.03)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>Short-Run</td>
<td>-0.30</td>
<td>-0.35</td>
<td>-0.41</td>
<td>-0.41</td>
</tr>
<tr>
<td></td>
<td>(0.05)</td>
<td>(0.06)</td>
<td>(0.12)</td>
<td>(0.12)</td>
</tr>
</tbody>
</table>

RC/FC refers to random/fixed coefficient model. NBC/BC/IBC refers to no bias-correction/bias correction/iterated bias correction estimates.

Note: *Standard errors are in parenthesis.*
This supplement to the paper "Panel Data Models with Nonadditive Unobserved Heterogeneity: Estimation and Inference" provides additional numerical examples and the proofs of the main results. It is organized in seven appendices. Appendix A contains a Monte Carlo simulation calibrated to the empirical example of the paper. Appendix B gives the proofs of the consistency of the one-step and two-step FE-GMM estimators. Appendix C includes the derivations of the asymptotic distribution of one-step and two-step FE-GMM estimators. Appendix D provides the derivations of the asymptotic distribution of bias corrected FE-GMM estimators. Appendix E and Appendix F contain the characterization of the stochastic expansions for the estimators of the individual effects and the scores. Appendix G includes the expressions for the scores and their derivatives.

Throughout the appendices $O_{up}$ and $o_{up}$ will denote uniform orders in probability. For example, for a sequence of random variables $\{\xi_i : 1 \leq i \leq n\}$, $\xi_i = O_{up}(1)$ means $\sup_{1 \leq i \leq n} \xi_i = O_P(1)$ as $n \rightarrow \infty$, and $\xi_i = o_{up}(1)$ means $\sup_{1 \leq i \leq n} \xi_i = o_P(1)$ as $n \rightarrow \infty$. It can be shown that the usual algebraic properties for $O_P$ and $o_P$ orders also apply to the uniform orders $O_{up}$ and $o_{up}$.

Let $e_j$ denote a $1 \times d$ unitary vector with a one in position $j$. For a matrix $A$, $|A|$ denotes Euclidean norm, that is $|A|^2 = \text{trace}[AA']$. HK refers to Hahn and Kuersteiner (2011).

APPENDIX A. NUMERICAL EXAMPLE

We design a Monte Carlo experiment to closely match the cigarette demand empirical example in the paper. In particular, we consider the following linear model with common and individual specific parameters:

\[ C_{it} = \alpha_{0i} + \alpha_{1i} P_{it} + \theta_1 C_{i,t-1} + \theta_2 C_{i,t+1} + \psi_{\epsilon_{it}}, \]
\[ P_{it} = \eta_{0i} + \eta_{1i} T a x_{it} + \upsilon_{it}, \quad (i = 1, 2, \ldots, n, t = 1, 2, \ldots, T); \]

where $\{(\alpha_{0i}, \eta_{1i}) : 1 \leq i \leq n\}$ is i.i.d. bivariate normal with mean $(\mu_j, \mu_{\eta_j})$, variances $(\sigma_j^2, \sigma_{\eta_j}^2)$, and correlation $\rho_j$, for $j \in \{0, 1\}$; independent across $j$; $\{\upsilon_{it} : 1 \leq t \leq T, 1 \leq i \leq n\}$ is i.i.d $N(0, \sigma^2_{\upsilon})$; and $\{\epsilon_{it} : 1 \leq t \leq T, 1 \leq i \leq n\}$ is i.i.d. standard normal. We fix the values of $T a x_{it}$ to the values in the data set. All the parameters other than $\rho_1$ and $\psi$ are calibrated to the data set. Since the panel is balanced for only 1972 to 1994, we set $T = 23$ and generate balanced panels for the simulations. Specifically, we consider

\[ n = 51, \quad T = 23; \mu_0 = 72.86, \mu_1 = -31.26, \mu_{\eta_0} = 0.81, \mu_{\eta_1} = 0.13, \quad \sigma_0 = 18.54, \sigma_1 = 10.60, \sigma_{\eta_0} = 0.14, \]
\[ \sigma_{\eta_1} = 2.05, \sigma_{\upsilon} = 0.15, \theta_1 = 0.45, \theta_2 = 0.27, \quad \rho_0 = -0.17, \quad \rho_1 \in \{0, 0.3, 0.6, 0.9\}, \psi \in \{2, 4, 6\}. \]

In the empirical example, the estimated values of $\rho_1$ and $\psi$ are close to 0.3 and 5, respectively.

Since the model is dynamic with leads and lags of the dependent variable on the right hand side, we construct the series of $C_{it}$ by solving the difference equation following BGM. The stationary part of the solution is

\[ C_{it} = \frac{1}{\theta_1 \phi_1 (\phi_2 - \phi_1)} \sum_{s=1}^{\infty} \phi_{1}^s h_{1}(t + s) + \frac{1}{\theta_1 \phi_2 (\phi_2 - \phi_1)} \sum_{s=0}^{\infty} \phi_{2}^{-s} h_{1}(t - s) \]

APPENDIX A. NUMERICAL EXAMPLE

We design a Monte Carlo experiment to closely match the cigarette demand empirical example in the paper. In particular, we consider the following linear model with common and individual specific parameters:

\[ C_{it} = \alpha_{0i} + \alpha_{1i} P_{it} + \theta_1 C_{i,t-1} + \theta_2 C_{i,t+1} + \psi_{\epsilon_{it}}, \]
\[ P_{it} = \eta_{0i} + \eta_{1i} T a x_{it} + \upsilon_{it}, \quad (i = 1, 2, \ldots, n, t = 1, 2, \ldots, T); \]

where $\{(\alpha_{0i}, \eta_{1i}) : 1 \leq i \leq n\}$ is i.i.d. bivariate normal with mean $(\mu_j, \mu_{\eta_j})$, variances $(\sigma_j^2, \sigma_{\eta_j}^2)$, and correlation $\rho_j$, for $j \in \{0, 1\}$; independent across $j$; $\{\upsilon_{it} : 1 \leq t \leq T, 1 \leq i \leq n\}$ is i.i.d $N(0, \sigma^2_{\upsilon})$; and $\{\epsilon_{it} : 1 \leq t \leq T, 1 \leq i \leq n\}$ is i.i.d. standard normal. We fix the values of $T a x_{it}$ to the values in the data set. All the parameters other than $\rho_1$ and $\psi$ are calibrated to the data set. Since the panel is balanced for only 1972 to 1994, we set $T = 23$ and generate balanced panels for the simulations. Specifically, we consider

\[ n = 51, \quad T = 23; \mu_0 = 72.86, \mu_1 = -31.26, \mu_{\eta_0} = 0.81, \mu_{\eta_1} = 0.13, \quad \sigma_0 = 18.54, \sigma_1 = 10.60, \sigma_{\eta_0} = 0.14, \]
\[ \sigma_{\eta_1} = 2.05, \sigma_{\upsilon} = 0.15, \theta_1 = 0.45, \theta_2 = 0.27, \quad \rho_0 = -0.17, \quad \rho_1 \in \{0, 0.3, 0.6, 0.9\}, \psi \in \{2, 4, 6\}. \]

In the empirical example, the estimated values of $\rho_1$ and $\psi$ are close to 0.3 and 5, respectively.

Since the model is dynamic with leads and lags of the dependent variable on the right hand side, we construct the series of $C_{it}$ by solving the difference equation following BGM. The stationary part of the solution is

\[ C_{it} = \frac{1}{\theta_1 \phi_1 (\phi_2 - \phi_1)} \sum_{s=1}^{\infty} \phi_{1}^s h_{1}(t + s) + \frac{1}{\theta_1 \phi_2 (\phi_2 - \phi_1)} \sum_{s=0}^{\infty} \phi_{2}^{-s} h_{1}(t - s) \]
where
\[ h_i(t) = \alpha_{0i} + \alpha_{1i} P_{i,t-1} + \psi \epsilon_{i,t-1}, \quad \phi_1 = \frac{1 - (1 - 4\theta_1 \theta_2)^{1/2}}{2\theta_1}, \quad \phi_2 = \frac{1 + (1 - 4\theta_1 \theta_2)^{1/2}}{2\theta_1}. \]

In our specification, these values are \( \phi_1 = 0.31 \) and \( \phi_2 = 1.91 \). The parameters that we vary across the experiments are \( \rho_1 \) and \( \psi \). The parameter \( \rho_1 \) controls the degree of correlation between \( \alpha_{1i} \) and \( P_{it} \) and determines the bias caused by using fixed coefficient estimators. The parameter \( \psi \) controls the degree of endogeneity in \( C_{i,t-1} \) and \( C_{i,t+1} \), which determines the bias of OLS and the incidental parameter bias of random coefficient IV estimators. Although \( \psi \) is not an ideal experimental parameter because it is the variance of the error, it is the only free parameter that affects the endogeneity of \( C_{i,t-1} \) and \( C_{i,t+1} \). In this design we cannot fully remove the endogeneity of \( C_{i,t-1} \) and \( C_{i,t+1} \) because of the dynamics.

In each simulation, we estimate the parameters with standard fixed coefficient OLS and IV with additive individual effects (FC), and the FE-GMM OLS and IV estimators with the individual specific coefficients (RC). For IV, we use the same set of instruments as in the empirical example. We report results only for the common coefficient \( \theta_2 \), and the mean and standard deviation of the individual-specific coefficient \( \alpha_{1i} \). Throughout the tables, Bias refers to the mean of the bias across simulations; SD refers to the standard deviation of the estimates; SE/SD denotes the ratio of the average standard error to the standard deviation; and \( \phi_{0.05} \) is the rejection frequency of a two-sided test with nominal level of 0.05 that the parameter is equal to its true value. For bias-corrected RC estimators the standard errors are calculated using bias corrected estimates of the common parameter and individual effects.

Table A.1 reports the results for the estimators of \( \theta_2 \). We find significant biases in all the OLS estimators relative to the standard deviations of these estimators. The bias of OLS grows with \( \psi \). The IV-RC estimator has bias unless \( \rho_1 = 0 \), that is unless there is no correlation between \( \alpha_{1i} \) and \( P_{it} \), and its test shows size distortions due to the bias and underestimation in the standard errors. IV-RC estimators have no bias in every configuration and their tests display much smaller size distortions than for the other estimators. The bias corrections preserve the bias and inference properties of the RC-IV estimator.

Table A2 reports similar results for the estimators of the mean of the individual specific coefficient \( \mu_1 = \bar{E}[\alpha_{1i}] \). We find substantial biases for OLS and IV-FC estimators. RC-IV displays some bias, which is removed by the corrections in some configurations. The bias corrections provide significant improvements in the estimation of standard errors. IV-RC standard errors overestimate the dispersion by more than 15% when \( \psi \) is greater than 2, whereas IV-BC or IV-IBC estimators have SE/SD ratios close to 1. As a result bias corrected estimators show smaller size distortions. This improvement comes from the bias correction in the estimates of the dispersion of \( \alpha_{1i} \) that we use to construct the standard errors. The bias of the estimator of the dispersion is generally large, and is effectively removed by the correction. We can see more evidence on this phenomenon in Table A3.

Table A3 shows the results for the estimators of the standard deviation of the individual specific coefficient \( \sigma_1 = \bar{E}[(\alpha_{1i} - \mu_1)^2]^{1/2} \). As noted above, the bias corrections are relevant in this case. As \( \psi \) increases, the bias grows in orders of \( \psi \). Most of bias is removed by the correction even when \( \psi \) is large. For example, when \( \psi = 6 \), the bias of IV-RC estimator is about 4 which is larger than two times its standard deviation. The correction reduces the bias to about 0.5, which is small relative to the standard deviation. Moreover, despite the overestimation in the standard errors, there are important size distortions for IV-RC estimators for tests on \( \sigma_1 \) when \( \psi \) is large. The bias corrections bring the rejection frequencies close to their nominal levels.
Overall, the calibrated Monte-Carlo experiment confirms that the IV-RC estimator with bias correction provides improved estimation and inference for all the parameters of interest for the model considered in the empirical example.

**APPENDIX B. CONSISTENCY OF ONE-STEP AND TWO-STEP FE-GMM ESTIMATOR**

**Lemma 3.** Suppose that the Conditions 1 and 2 hold. Then, for every $\eta > 0$
\[
\Pr \left\{ \sup_{1 \leq i \leq n} \sup_{(\theta, \alpha) \in T} \left| Q_i^W(\theta, \alpha) - Q_i^W(\theta, \alpha') \right| \geq \eta \right\} = o(T^{-1}),
\]
and
\[
\sup_{\alpha} \left| Q_i^W(\theta, \alpha) - Q_i^W(\theta', \alpha) \right| \leq C \cdot E[M(z_t)]^2 |\theta - \theta'|
\]
for some constant $C > 0$.

**Proof.** First, note that
\[
\left| Q_i^W(\theta, \alpha) - Q_i^W(\theta', \alpha) \right| \leq \left| \tilde{g}_i(\theta, \alpha) W_i^{-1} \tilde{g}_i(\theta, \alpha) - g_i(\theta, \alpha)' W_i^{-1} g_i(\theta, \alpha) \right| + \left| \tilde{g}_i(\theta, \alpha)' (W_i^{-1} - W_i^{-1}) g_i(\theta, \alpha) \right|
\]
\[
+ \left| g_i(\theta, \alpha)' (W_i^{-1} - W_i^{-1}) g_i(\theta, \alpha) \right| + \left| g_i(\theta, \alpha)' (W_i^{-1} - W_i^{-1}) g_i(\theta, \alpha) \right|
\]
\[
+ \left| g_i(\theta, \alpha)' (W_i^{-1} - W_i^{-1}) g_i(\theta, \alpha) \right| \leq d_2^2 \max_{1 \leq k \leq d_2} |\tilde{g}_{k,i}(\theta, \alpha) - g_{k,i}(\theta, \alpha)|^2 |W_i|^{-1}
\]
\[
+ 2d_2^2 \sup_{1 \leq i \leq n} E[M(z_t)] |W_i|^{-1} \max_{1 \leq k \leq d_2} |\tilde{g}_{k,i}(\theta, \alpha) - g_{k,i}(\theta, \alpha)| + o_P \left( \max_{1 \leq k \leq d_2} |\tilde{g}_{k,i}(\theta, \alpha) - g_{k,i}(\theta, \alpha)| \right),
\]
where we use that $\sup_{1 \leq i \leq n} |\tilde{W}_i - W_i| = o_P(1)$. Then, by Condition 2, we can apply Lemma 4 of HK to $|\tilde{g}_{k,i}(\theta, \alpha) - g_{k,i}(\theta, \alpha)|$ to obtain the first part.

The second part follows from
\[
\left| Q_i^W(\theta, \alpha) - Q_i^W(\theta', \alpha) \right| \leq \left| g_i(\theta, \alpha)' W_i^{-1} [g_i(\theta, \alpha) - g_i(\theta', \alpha)] \right| + \left| g_i(\theta, \alpha) - g_i(\theta', \alpha)' W_i^{-1} g_i(\theta', \alpha) \right|
\]
\[
\leq 2 \cdot d_2^2 E[M(z_t)]^2 |W_i|^{-1} |\theta - \theta'|
\]
\[
\square
\]

**B.1. Proof of Theorem 1.**

**Proof.** Part I: Consistency of $\tilde{\theta}$. For any $\eta > 0$, let $\varepsilon := \inf_{(\theta, \alpha)}[Q_i^W(\theta_0, \alpha_0) - \sup_{(\theta, \alpha) \neq (\theta_0, \alpha_0)} |Q_i^W(\theta, \alpha)|] > 0$ as defined in Condition 2. Using the standard argument for consistency of extremum estimator, as in Newey and McFadden (1994), with probability $1 - o(T^{-1})$
\[
\max_{|\theta - \theta_0| > \eta, \alpha_1, \ldots, \alpha_n} n^{-1} \sum_{i=1}^{n} \tilde{Q}_i^W(\theta, \alpha_i) < n^{-1} \sum_{i=1}^{n} Q_i^W(\theta_0, \alpha_0) - \frac{1}{3}\varepsilon,
\]
by definition of $\varepsilon$ and Lemma 3. Thus, by continuity of $\tilde{Q}_i^W$ and the definition of the lefthand side above, we conclude that $\Pr \left[ |\tilde{\theta} - \theta_0| \geq \eta \right] = o(T^{-1})$.

Part II: Consistency of $\tilde{a}_i$. By Part I and Lemma 3,
\[
\Pr \left[ \sup_{1 \leq i \leq n} \sup_{\alpha} \left| \tilde{Q}_i^W(\tilde{\theta}, \alpha) - Q_i^W(\theta_0, \alpha) \right| \geq \eta \right] = o(T^{-1})
\]
for any \( \eta > 0 \). Let
\[
\varepsilon := \inf_i \left[ Q_i^W (\theta_0, \alpha_{i0}) - \sup_{(\alpha_i|\alpha_i - \alpha_{i0}| > \eta)} Q_i^W (\theta_0, \alpha_i) \right] > 0.
\]
Condition on the event
\[
\left\{ \sup_{1 \leq i \leq n} \sup_{\alpha_i} \left| Q_i^W (\bar{\theta}, \alpha_i) - Q_i^W (\theta_0, \alpha_i) \right| \leq \frac{1}{3} \varepsilon \right\},
\]
which has a probability equal to \( 1 - o (T^{-1}) \) by (B.1). Then
\[
\max_{|\alpha_i - \alpha_{i0}| < \eta} Q_i^W (\bar{\theta}, \alpha_i) < \max_{|\alpha_i - \alpha_{i0}| < \eta} Q_i^W (\theta_0, \alpha_i) + \frac{1}{3} \varepsilon < Q_i^W (\theta_0, \alpha_{i0}) - \frac{2}{3} \varepsilon < Q_i^W (\bar{\theta}, \alpha_{i0}) - \frac{1}{3} \varepsilon.
\]
This is inconsistent with \( Q_i^W (\bar{\theta}, \bar{\alpha}_i) \geq Q_i^W (\bar{\theta}, \alpha_{i0}) \), and therefore, \( |\bar{\alpha}_i - \alpha_{i0}| \leq \eta \) with probability \( 1 - o(T^{-1}) \) for every \( i \).

**Part III: Consistency of \( \bar{\lambda}_i \).** First, note that
\[
|\bar{\lambda}_i| = |\bar{W}_i^{-1} \bar{g}_i(\bar{\theta}, \bar{\alpha}_i)| \leq d_g |\bar{W}_i|^{-1} \max_{1 \leq k \leq d_g} \left( |\bar{g}_{k,i}(\bar{\theta}, \bar{\alpha}_i) - g_{k,i}(\bar{\theta}, \bar{\alpha}_i)| + |g_{k,i}(\bar{\theta}, \bar{\alpha}_i)| \right)
\]
\[
\leq d_g |\bar{W}_i|^{-1} \max_{1 \leq k \leq d_g} \sup_{(\theta, \alpha_i) \in T} |\bar{g}_{k,i}(\theta, \alpha_i) - g_{k,i}(\theta, \alpha_i)|
\]
\[
+ d_g |\bar{W}_i|^{-1} M(z_{it}) |\bar{\theta} - \theta_0| + d_g |\bar{W}_i|^{-1} M(z_{it}) |\bar{\alpha}_i - \alpha_{i0}|.
\]
Then, the result follows because \( \sup_{1 \leq i \leq n} |\bar{W}_i - W_i| = o_P(1) \) and \( \{ W_i : 1 \leq i \leq n \} \) are positive definite by Condition 2, \( \max_{1 \leq k \leq d_g} \sup_{(\theta, \alpha_i) \in T} |\bar{g}_{k,i}(\theta, \alpha_i) - g_{k,i}(\theta, \alpha_i)| = o_P(1) \) by Lemma 4 in HK, and \( |\bar{\theta} - \theta_0| = o_P(1) \) and \( \sup_{1 \leq i \leq n} |\bar{\alpha}_i - \alpha_{i0}| = o_P(1) \) by Parts I and II.

**B.2. Proof of Theorem 3.**

**Proof.** First, assume that Conditions 1, 2, 3 and 5 hold. The proofs are exactly the same as that of Theorem 1 using the uniform convergence of the criterion function.

To establish the uniform convergence of the criterion function as in Lemma 3, we need
\[
\sup_{1 \leq i \leq n} \left| \Omega_i(\bar{\theta}, \bar{\alpha}_i) - \Omega_i(\theta_0, \alpha_{i0}) \right| = o_P(1),
\]
along with an extended version of the continuous mapping theorem for \( o_P \). This can be shown by noting that
\[
|\bar{\lambda}_i - \Omega_i(\theta_0, \alpha_{i0})| \leq \Omega_i(\bar{\theta}, \bar{\alpha}_i) - \Omega_i(\bar{\theta}, \bar{\alpha}_i) + |\Omega_i(\bar{\theta}, \bar{\alpha}_i) - \Omega_i(\theta_0, \alpha_{i0})|
\]
\[
\leq \left[ |\Omega_i(\bar{\theta}, \bar{\alpha}_i) - \Omega_i(\bar{\theta}, \bar{\alpha}_i)| + d_g^2 E \left[ M(z_{it})^2 \right] \right] |\bar{\theta} - \theta_0|.
\]
The convergence follows by the consistency of \( \bar{\theta} \) and \( \bar{\alpha}_i \)’s, and the application of Lemma 2 of HK to \( g_{k}(z_{it}; \theta, \alpha_i)g_{l}(z_{it}; \theta, \alpha_i) \) using that \( |g_{k}(z_{it}; \theta, \alpha_i)g_{l}(z_{it}; \theta, \alpha_i)| \leq M(z_{it})^2 \).

**APPENDIX C. ASYMPTOTIC DISTRIBUTION OF ONE-STEP AND TWO-STEP FE-GMM ESTIMATOR**

**C.1. Some Lemmas.**

**Lemma 4.** Assume that Condition 1 holds. Let \( h(z_{it}; \theta, \alpha_i) \) be a function such that (i) \( h(z_{it}; \theta, \alpha_i) \) is continuously differentiable in \( (\theta, \alpha_i) \in T \subset \mathbb{R}^{d_{\theta} + d_{\alpha_i}} \); (ii) \( T \) is convex; (iii) there exists a function \( M(z_{it}) \) such that \( |h(z_{it}; \theta, \alpha_i)| \leq M(z_{it}) \) and \( |\partial h(z_{it}; \theta, \alpha_i)/\partial \theta(\theta, \alpha_i)| \leq M(z_{it}) \) with \( E \left[ M(z_{it})^{5(d_{\theta} + d_{\alpha_i} + 6)/(1 - 10\alpha + \delta)} \right] < \infty \).
for some \( \delta > 0 \) and \( 0 < \nu < 1/10 \). Define \( \tilde{H}_i(\theta, \alpha_i) := T^{-1} \sum_{t=1}^{T} h(z_{it}; \theta, \alpha_i) \), and \( H_i(\theta, \alpha_i) := E \left[ \tilde{H}_i(\theta, \alpha_i) \right] \).

Let
\[
\alpha^*_i = \arg \max_{\alpha_i} \tilde{Q}_i^W(\theta^*, \alpha_i),
\]
such that \( \alpha^*_i - \alpha_{i0} = o_P(T^{\alpha_0}) \) and \( \theta^* - \theta_0 = o_P(T^{\alpha_0}) \), with \(-2/5 \leq a \leq 0\), for \( a = \max(\alpha_\alpha, a_\theta) \). Then, for any \( \bar{\theta} \) between \( \theta^* \) and \( \theta_0 \), and \( \bar{\alpha}_i \) between \( \alpha^*_i \) and \( \alpha_{i0} \),
\[
\sqrt{T} \left[ \tilde{H}_i(\bar{\theta}, \bar{\alpha}_i) - H_i(\bar{\theta}, \bar{\alpha}_i) \right] = o_u(T^{1/10}), \quad \tilde{H}_i(\bar{\theta}, \bar{\alpha}_i) - H_i(\theta_0, \alpha_{i0}) = o_u(T^a).
\]

**Proof.** The first statement follows from Lemma 2 in HK. The second statement follows by the first statement and the conditions of the Lemma by a mean value expansion since
\[
\left| \tilde{H}_i(\bar{\theta}, \bar{\alpha}_i) - H_i(\theta_0, \alpha_{i0}) \right| \leq \frac{1}{\alpha_{i0} - \alpha^*_i} \left| \sum_{t=1}^{T} M(z_{it}) \right| \left( \frac{1}{T} \sum_{t=1}^{T} M(z_{it}) \right) = o_u(P(T^a))
\]
\[
+ \frac{1}{\alpha_{i0} - \alpha^*_i} \left| \tilde{H}_i(\theta_0, \alpha_{i0}) - H_i(\theta_0, \alpha_{i0}) \right| = o_u(T^{a/2})
\]

**Lemma 5.** Assume that Conditions 1, 2, 3 and 4 hold. Let \( \tilde{\gamma}_i^W(\theta, \gamma_i) \) denote the first stage GMM score of the fixed effects, that is
\[
\tilde{\gamma}_i^W(\theta, \gamma_i) = - \left( \tilde{G}_i(\theta, \alpha_i) \gamma_i + \tilde{W}_i \gamma_i \right),
\]
where \( \gamma_i = (\alpha^*_i, \gamma_i') \), \( \tilde{s}_i^W(\theta, \gamma_i) \) denote the one-step GMM score for the common parameter, that is
\[
\tilde{s}_i^W(\theta, \gamma_i) = - \tilde{G}_i(\theta, \alpha_i) \gamma_i,
\]
and \( \tilde{\gamma}_i(\theta) \) be such that \( \tilde{\gamma}_i^W(\theta, \tilde{\gamma}_i(\theta)) = 0 \).

Let \( \tilde{T}_{i,j}^W(\theta, \gamma_i) \) denote \( \partial \tilde{\gamma}_i^W(\theta, \gamma_i)/\partial \gamma'_i \partial \gamma_{i,j} \), and \( \tilde{S}_{i,j}^W(\theta, \gamma_i) \) denote \( \partial \tilde{s}_i^W(\theta, \gamma_i)/\partial \gamma'_i \partial \gamma_{i,j} \), for some \( 0 \leq j \leq d_2 + d_{\alpha} \), where \( \gamma_{i,j} \) is the \( j \)th element of \( \gamma_i \) and \( j = 0 \) denotes no second derivative. Let \( \tilde{N}_i^W(\theta, \gamma_i) \) denote \( \partial \tilde{\gamma}_i^W(\theta, \gamma_i)/\partial \theta' \) and \( \tilde{S}_i^W(\theta, \gamma_i) \) denote \( \partial \tilde{s}_i^W(\theta, \gamma_i)/\partial \theta' \). Let \( \bar{\theta}, \bar{\gamma}_1, \ldots, \bar{\gamma}_n \) be the one-stage GMM estimator. Then, for any \( \bar{\theta} \) between \( \theta \) and \( \theta_0 \), and \( \bar{\gamma}_1 \) between \( \gamma_1 \) and \( \gamma_0 \),
\[
\tilde{T}_{i,j}^W(\bar{\theta}, \bar{\gamma}_i) - T_{i,j}^W = o_u(1), \quad M_{i,j}^W(\bar{\theta}, \bar{\gamma}_i) - M_{i,j}^W = o_u(1), \quad N_i^W(\bar{\theta}, \bar{\gamma}_i) - N_i^W = o_u(1), \quad S_i^W(\bar{\theta}, \bar{\gamma}_i) - S_i^W = o_u(1).
\]

Also, for any \( \bar{\gamma}_i0 \) between \( \gamma_0 \) and \( \gamma_0 = \bar{\gamma}_i(\theta_0) \),
\[
\sqrt{T} \tilde{T}_{i,j}^W(\theta_0, \bar{\gamma}_i0) = o_u(T^{1/10}), \quad \sqrt{T} \left( \tilde{T}_{i,j}^W(\theta_0, \bar{\gamma}_i0) - T_{i,j}^W \right) = o_u(T^{1/10}), \quad \sqrt{T} \left( M_{i,j}^W(\theta_0, \bar{\gamma}_i0) - M_{i,j}^W \right) = o_u(T^{1/10}), \quad \sqrt{T} \left( N_i^W(\theta_0, \bar{\gamma}_i0) - N_i^W \right) = o_u(T^{1/10}), \quad \sqrt{T} \left( S_i^W(\theta_0, \bar{\gamma}_i0) - S_i^W \right) = o_u(T^{1/10}).
\]

**Proof.** The first set of results follows by inspection of the scores and their derivatives (the expressions are given in Appendix G), uniform consistency of \( \bar{\gamma}_i \) by Theorem 1 and application of the first part of Lemma 4 to \( \theta^* = \bar{\theta} \) and \( \alpha^*_i = \bar{\alpha}_i \) with \( a = 0 \).

The following steps are used to prove the second set of result. By Lemma 4,
\[
\sqrt{T} \tilde{T}_{i,j}^W = o_u(T^{1/10}), \quad \tilde{T}_{i,j}^W(\theta_0, \bar{\gamma}_i0) - T_{i,j}^W = o_u(1).
\]
where \( \tilde{\gamma}_{i0} \) is between \( \tilde{\gamma}_{i0} \) and \( \gamma_{i0} \). Then, a mean value expansion of the FOC of \( \tilde{\gamma}_{i0} \), \( \tilde{T}_i^W(\theta_0, \tilde{\gamma}_{i0}) = 0 \), around \( \tilde{\gamma}_{i0} = \gamma_{i0} \) gives
\[
\sqrt{T(\gamma_{i0} - \gamma_{i0})} = -(T_i^W)^{-1} \sqrt{T_i^W} - (T_i^W)^{-1} (\tilde{T}_i^W(\theta_0, \tilde{\gamma}_{i0}) - T_i^W) \sqrt{T(\gamma_{i0} - \gamma_{i0})} = O_u(T_i^W(\gamma_{i0} - \gamma_{i0})) \]
by Condition 3 and the previous result. Therefore,
\[
(1 + o_u(1)) \sqrt{T(\gamma_{i0} - \gamma_{i0})} = o_u(T_i^W(1/10)) \Rightarrow \sqrt{T(\gamma_{i0} - \gamma_{i0})} = o_u(T_i^W(1/10)).
\]
Given this uniform rate for \( \tilde{\gamma}_{i0} \), the desired result can be obtained by applying the second part of Lemma 4 to \( \theta^* = \theta_0 \) and \( \alpha^* = \alpha_{i0} \) with \( a = -2/5 \).

C.2. Proof of Theorem 2.

Proof. By a mean value expansion of the FOC for \( \tilde{\theta} \) around \( \tilde{\theta} = \theta_0 \),
\[
0 = s^W(\tilde{\theta}) = s^W(\theta_0) + \frac{d s^W(\tilde{\theta})}{d\theta'}(\tilde{\theta} - \theta_0),
\]
where \( \tilde{\theta} \) lies between \( \tilde{\theta} \) and \( \theta_0 \).

Part I: Asymptotic limit of \( d s^W(\tilde{\theta})/d\theta' \). Note that
\[
\frac{d s^W(\tilde{\theta})}{d\theta'} = \frac{1}{n} \sum_{i=1}^{n} \frac{d s_i^W(\tilde{\theta}, \tilde{\gamma}_i(\tilde{\theta}))}{d\theta'}, \quad \text{(C.1)}
\]
By Lemma 5,
\[
\frac{d s_i^W(\tilde{\theta}, \tilde{\gamma}_i(\tilde{\theta}))}{d\theta'} = s_i^W(\tilde{\theta}) + o_u(1), \quad \text{and} \quad \frac{d s_i^W(\tilde{\theta}, \tilde{\gamma}_i(\tilde{\theta}))}{d\theta'} = m_i^W + o_u(1).
\]
Then, differentiation of the FOC for \( \tilde{\gamma}_i(\tilde{\theta}) \), \( \tilde{T}_i^W(\tilde{\theta}, \tilde{\gamma}_i(\tilde{\theta})) = 0 \), with respect to \( \theta \) and \( \tilde{\gamma}_i \) gives
\[
\tilde{T}_i^W(\tilde{\theta}, \tilde{\gamma}_i(\tilde{\theta}))(\frac{\partial \tilde{\gamma}_i(\tilde{\theta})}{\partial \theta'} + \frac{\partial \tilde{\gamma}_i(\tilde{\theta})}{\partial \tilde{\gamma}_i(\tilde{\theta}))} = 0,
\]
By repeated application of Lemma 5 and Condition 3,
\[
\frac{\partial \tilde{\gamma}_i(\tilde{\theta})}{\partial \theta'} = -(T_i^W)^{-1} N_i^W + o_u(1).
\]
Finally, replacing the expressions for the components in \( \text{(C.1)} \) and using the formulae for the derivatives, which are provided in the Appendix G,
\[
\frac{d s^W(\tilde{\theta})}{d\theta'} = \frac{1}{n} \sum_{i=1}^{n} G^W_{\alpha} + \sum_{i=1}^{n} G^W_{\alpha} + o_u(1) = J^W + o_u(1), \quad J^W = E[G^W_{\alpha} + o_u(1)], \quad \text{(C.2)}
\]
Part II: Asymptotic Expansion for \( \tilde{\theta} - \theta_0 \). By \( \text{(C.2)} \) and Lemma 22, which states the stochastic expansion of \( \sqrt{nT} s^W(\theta_0) \),
\[
0 = \frac{\sqrt{nT} s^W(\theta_0)}{O_u(1)} + \frac{\frac{d s^W(\tilde{\theta})}{d\theta'}}{O_u(1)} \sqrt{nT}(\tilde{\theta} - \theta_0).
\]
Therefore, $\sqrt{n}T(\hat{\theta} - \theta_0) = O_P(1)$, and by part I, Lemma 22 and Condition 3,

$$\sqrt{n}T(\hat{\theta} - \theta_0) \overset{d}{\to} -(J^*)^{-1}N(\kappa B_s^W, V_s^W).$$

\[\Box\]


Proof. Applying Lemma 4 with a minor modification, along with Condition 4, we can prove an exact counterpart to Lemma 5 for the two-step GMM score for the fixed effects

$$\hat{I}_i(\theta, \gamma_i) = \hat{I}_i^F(\theta, \gamma_i) + \hat{I}_i^R(\theta, \gamma_i),$$

where the expressions of $\hat{I}_i^F$ and $\hat{I}_i^R$ are given in the Appendix G, and for the two-step score of the common parameter

$$\hat{a}_i(\theta, \gamma_i) = -\hat{G}_{a_i}(\theta, \gamma_i)^\top \hat{a}_i(\theta),$$

The only difference arises due to the term $\hat{I}_i^F(\theta, \gamma_i)$, which involves $\hat{I}_i(\theta, \gamma_i) - \Omega_i$. Lemma 8 shows that $\sqrt{T}(\hat{I}_i(\theta, \gamma_i) - \Omega_i) = o_u(T^{1/10})$, so that a result similar to Lemma 5 holds for the two-step scores.

Thus, we can make the same argument as in the proof of Theorem 2 using the stochastic expansion of $\sqrt{n}T\hat{a}(\theta_0)$ given in Lemma 23. \[\Box\]

APPENDIX D. ASYMPTOTIC DISTRIBUTION OF BIAS-CORRECTED TWO-STEP GMM ESTIMATOR

D.1. Some Lemmas.

Lemma 6. Assume that Conditions 1, 2, 3, 4 and 5 hold. Let $\hat{I}_i(\theta, \gamma_i)$ denote the two-step GMM score for the fixed effects, $\hat{a}_i(\theta, \gamma_i)$ denote the two-step GMM score for the common parameter, and $\hat{\gamma}_i(\theta)$ be such that $\hat{I}_i(\theta, \gamma_i(\theta)) = 0$. Let $\bar{\hat{I}}_{i,j}(\theta, \gamma_i)$ denote $\partial \hat{I}_{i,j}(\theta, \gamma_i)/\partial \gamma_{i,j}$, for some $0 \leq j \leq d_\theta + d_\gamma$, where $\gamma_{i,j}$ is the $j$th component of $\gamma_i$ and $j = 0$ denotes no second derivative. Let $\bar{\hat{N}}_{i}(\theta, \gamma_i)$ denote $\partial \hat{a}_i(\theta, \gamma_i)/\partial \theta'$. Let $\bar{\hat{M}}_{i,j}(\theta, \gamma_i)$ denote $\partial \bar{\hat{a}}_i(\theta, \gamma_i)/\partial \gamma_{i,j}$, for some $0 \leq j \leq d_\theta + d_\gamma$. Let $\bar{\hat{S}}_{i}(\theta, \gamma_i)$ denote $\partial \bar{\hat{a}}_i(\theta, \gamma_i)/\partial \theta'$. Let $\hat{\theta}(\gamma_{i}), \{\gamma_{i}\}_{i=1}^{n}$ be the two-step GMM estimators.

Then, for any $\bar{\theta}$ between $\hat{\theta}$ and $\theta_0$, and $\gamma_{i}$ between $\gamma_{i}$ and $\gamma_{i0}$,

$$\sqrt{T}(\bar{\hat{I}}_{i,d}(\bar{\theta}, \gamma_{i}) - \hat{T}_{i,d}) = o_u(T^{1/10}), \quad \sqrt{T}(\bar{\hat{M}}_{i,j}(\bar{\theta}, \gamma_{i}) - \hat{M}_{i,j}) = o_u(T^{1/10}),$$
$$\sqrt{T}(\bar{\hat{N}}_{i}(\bar{\theta}, \gamma_{i}) - \hat{N}_{i}) = o_u(T^{1/10}), \quad \sqrt{T}(\bar{\hat{S}}_{i}(\bar{\theta}, \gamma_{i}) - \hat{S}_{i}) = o_u(T^{1/10}).$$

Proof. Let $\hat{\gamma}_{i} = \hat{\gamma}_{i}(\bar{\theta})$ and $\gamma_{i0} = \hat{\gamma}_{i}(\theta_0)$. First, note that

$$\sqrt{T}(\hat{\gamma}_{i} - \gamma_{i0}) = \frac{\partial \hat{\gamma}_{i}(\bar{\theta})}{\partial \theta'} \sqrt{T}(\bar{\theta} - \theta_0) = -(T_{i}^{-1})^{-1} N_{i} \sqrt{T}(\bar{\theta} - \theta_0) + o_u(T^{(n^{-1/2})})$$

where the second equality follows from the proof of Theorem 2 and 4. Thus, by the same argument used in the proof of Lemma 5,

$$\sqrt{T}(\hat{\gamma}_{i} - \gamma_{i0}) = \sqrt{T}(\hat{\gamma}_{i} - \gamma_{i0}) + \sqrt{T}(\hat{\gamma}_{i0} - \gamma_{i0}) = o_u(T^{1/10}).$$

Given this result and inspection of the scores and their derivatives (see the Appendix G), the proof is similar to the proof of the second part of Lemma 5. \[\Box\]
Lemma 7. Assume that Condition 1 holds. Let \( h_j(z_{it}; \theta, \alpha_i) \), \( j = 1, 2 \) be two functions such that (i) \( h_j(z_{it}; \theta, \alpha_i) \) is continuously differentiable in \((\theta, \alpha_i) \in T \subset \mathbb{R}^{d_\theta + d_\alpha} \); (ii) \( T \) is convex; (iii) there exists a function \( M(z_{it}) \) such that \(|h_j(z_{it}; \theta, \alpha_i)| \leq M(z_{it}) \) and \(|\partial h_j(z_{it}; \theta, \alpha_i)/\partial (\theta, \alpha_i)| \leq M(z_{it}) \) with \( E \{M(z_{it})^{10(d_\theta + d_\alpha + 6)/(1 - 10v) + 4} \} \). Define \( \hat{F}_i(\theta, \alpha_i) := T^{-1} \sum_{t=1}^T h_1(z_{it}; \theta, \alpha_i)h_2(z_{it}; \theta, \alpha_i) \), and \( F_i(\theta, \alpha_i) := E \{\hat{F}_i(\theta, \alpha_i)\} \). Let \( \alpha^*_i = \arg \sup_{\alpha_i} \hat{Q}_i^W(\alpha^*_i, \alpha) \), such that \( \alpha^*_i - \alpha_{i0} = o_P(T^{-a}) \) and \( \theta^* - \theta_0 = o_P(T^{-a/2}) \), with \(-2/5 \leq \alpha \leq 0\), for \( \alpha = \max(\alpha_a, \alpha_\theta) \). Then, for any \( \bar{\theta} \) between \( \theta^* \) and \( \theta_0 \), and \( \bar{\alpha}_i \) between \( \alpha^*_i \) and \( \alpha_{i0} \),

\[
\sqrt{T} \left( \Omega_{\alpha^*_i, \theta^*_i}(\bar{\theta}, \bar{\alpha}_i) - \Omega_{\alpha^*_i, \theta^*_i} \right) = o_p \left( T^{1/10} \right).
\]

Proof. Same as for Lemma 4, replacing \( H_i \) by \( F_i \), and \( M(z_{it}) \) by \( M(z_{it})^2 \). \( \square \)

Lemma 8. Assume that Conditions 1, 2, 3, 4, 5, and \( \delta \) hold. Let \( \hat{\Omega}_i(\bar{\theta}, \bar{\alpha}_i) = T^{-1} \sum_{t=1}^T g(z_{it}; \bar{\theta}, \bar{\alpha}_i)g(z_{it}; \bar{\theta}, \bar{\alpha}_i)' \) be an estimator of the covariance function \( \Omega_i = E[g(z_{it}; \bar{\theta}, \bar{\alpha}_i)g(z_{it}; \bar{\theta}, \bar{\alpha}_i)'] \), where \( \bar{\theta} = \theta_0 + o_P(T^{-2/5}) \) and \( \bar{\alpha}_i = \alpha_{i0} + o_p(T^{-2/5}) \). Let \( \hat{\Omega}_{\alpha^*_i, \theta^*_i}(\bar{\theta}, \bar{\alpha}_i) = \partial_{\alpha^*_i, \theta^*_i} \hat{\Omega}(\bar{\theta}, \bar{\alpha}_i)/\partial_{\alpha^*_i, \theta^*_i} \), for \( 0 \leq d_1 + d_2 \leq 2 \). Then,

\[
\sqrt{T} \left( \Omega_{\alpha^*_i, \theta^*_i}(\bar{\theta}, \bar{\alpha}_i) - \Omega_{\alpha^*_i, \theta^*_i} \right) = o_p \left( T^{1/10} \right).
\]

Proof. Note that

\[
|g(z_{it}; \bar{\theta}, \bar{\alpha}_i)g(z_{it}; \bar{\theta}, \bar{\alpha}_i)' - E[g(z_{it}; \bar{\theta}, \bar{\alpha}_i)g(z_{it}; \bar{\theta}, \bar{\alpha}_i)']| \\
\leq d_2 \left( \max_{1 \leq k \leq d_2} \left| g_k(z_{it}; \bar{\theta}, \bar{\alpha}_i)g(z_{it}; \bar{\theta}, \bar{\alpha}_i)' - E[g_k(z_{it}; \bar{\theta}, \bar{\alpha}_i)g(z_{it}; \bar{\theta}, \bar{\alpha}_i)'] \right| \right).
\]

Then we can apply Lemma 7 to \( h_1 = g_1 \) and \( h_2 = g_1 \) with \( \alpha = -2/5 \). A similar argument applies to the derivatives, since they are sums of products of elements that satisfy the assumption of Lemma 7. \( \square \)

Lemma 9. Assume that Conditions 1, 2, 3, 4, 5, and \( \delta \) hold, and \( \ell \to \infty \) such that \( \ell/T \to 0 \) as \( T \to \infty \). For any \( \bar{\theta} \) between \( \bar{\theta} \) and \( \theta_0 \), let \( \tilde{\Sigma}_\alpha(\bar{\theta}) = \left[ \hat{G}_{\alpha_t}(\bar{\theta}) \hat{G}_{\alpha_t}(\bar{\theta})' \right]^{-1}, \hat{H}_{\alpha_t}(\bar{\theta}) = \tilde{\Sigma}_\alpha(\bar{\theta}) \hat{G}_{\alpha_t}(\bar{\theta})' \hat{G}_{\alpha_t}(\bar{\theta})^{-1}, \hat{P}_{\alpha_t}(\bar{\theta}) = \hat{G}_{\alpha_t}(\bar{\theta})^{-1} - \hat{G}_{\alpha_t}(\bar{\theta}) \hat{H}_{\alpha_t}(\bar{\theta}) \hat{G}_{\alpha_t}(\bar{\theta})^{-1}, \tilde{\Sigma}_{\alpha_t}^W(\bar{\theta}) = \left[ \hat{G}_{\alpha_t}(\bar{\theta}) \hat{H}_{\alpha_t}(\bar{\theta}) \hat{G}_{\alpha_t}(\bar{\theta})' \right]^{-1}, \hat{H}_{\alpha_t}^W(\bar{\theta}) = \tilde{\Sigma}_{\alpha_t}^W(\bar{\theta}) \hat{G}_{\alpha_t}(\bar{\theta})' \hat{W}_{\alpha_t}(\bar{\theta})^{-1}, \hat{P}_{\alpha_t}^W(\bar{\theta}) = \hat{G}_{\alpha_t}(\bar{\theta})' \hat{P}_{\alpha_t}(\bar{\theta}) \hat{G}_{\alpha_t}(\bar{\theta})^{-1} - \hat{G}_{\alpha_t}(\bar{\theta})'[\hat{B}_{\alpha_t}(\bar{\theta}) + \hat{B}_{\alpha_t}^W(\bar{\theta}) + \hat{B}_{\alpha_t}^W(\bar{\theta})], \) where

\[
\hat{B}_{\alpha_t}^W(\bar{\theta}) = -\hat{P}_{\alpha_t}(\theta) \sum_{j=1}^{d_\alpha} \hat{G}_{\alpha_t,j}(\theta) \hat{H}_{\alpha_t}(\theta) \hat{G}_{\alpha_t,j}(\theta) \hat{G}_{\alpha_t,j}(\theta),
\]

\[
\hat{B}_{\alpha_t}^W(\bar{\theta}) = \hat{H}_{\alpha_t}(\theta) \sum_{j=1}^{d_\alpha} \hat{G}_{\alpha_t,j}(\theta) \hat{P}_{\alpha_t}(\theta) \hat{G}_{\alpha_t,j}(\theta),
\]

\[
\hat{B}_{\alpha_t}^W(\bar{\theta}) = \hat{P}_{\alpha_t}(\theta) \sum_{j=1}^{d_\alpha} \hat{G}_{\alpha_t,j}(\theta) \hat{G}_{\alpha_t,j}(\theta) \hat{H}_{\alpha_t}(\theta) \hat{G}_{\alpha_t,j}(\theta),
\]

\[
\hat{B}_{\alpha_t}^W(\theta) = \hat{P}_{\alpha_t}(\theta) \sum_{j=1}^{d_\alpha} \hat{G}_{\alpha_t,j}(\theta) \hat{P}_{\alpha_t}(\theta) \hat{H}_{\alpha_t}(\theta) \hat{G}_{\alpha_t,j}(\theta),
\]

\[
\hat{B}_{\alpha_t}^W(\bar{\theta}) = \hat{P}_{\alpha_t}(\theta) \sum_{j=1}^{d_\alpha} \hat{G}_{\alpha_t,j}(\theta) \hat{P}_{\alpha_t}(\theta) \hat{H}_{\alpha_t}(\theta) \hat{G}_{\alpha_t,j}(\theta),
\]

86
be estimators of $E$, $H$, $P$, $E_i$, $HW$, $J$, $Bq$ and $Bf$. Let $\hat{F}_{\alpha_1, \theta_2}(\theta, \hat{\alpha}_i(\theta))$ and $F_{\alpha_1, \theta_2}(\theta, \alpha_i)$, with $F \in \{\Sigma, H, P, \Sigma^W, H^W, J_{\beta}, B_{\beta_1}, B_{\beta}^W\}$ denote their derivatives for $0 \leq d_1 + d_2 \leq 1$. Then,

$$\sqrt{T} \left( \hat{F}_{\alpha_1, \theta_2}(\theta, \hat{\alpha}_i(\theta)) - F_{\alpha_1, \theta_2}(\theta, \alpha_i) \right) = o_u P \left( T^{1/10} \right),$$

where $F_{\alpha_1, \theta_2}(\theta, \alpha_i) := F$ if $d_1 + d_2 = 0$.

**Proof.** The results follow by Theorem 3 and Lemma 6, using the algebraic properties of the $o_u P$ orders and Lemma 12 of HK to show the properties of the estimators of the spectral expectations. \hfill \Box

**Lemma 10.** Assume that Conditions 1, 2, 3, 4, 5, and 6 hold. Then, for any $\bar{\theta}$ between $\hat{\theta}$ and $\theta_0$,

$$\hat{J}_s(\bar{\theta}) = J_s + o_P(T^{-2/5}).$$

**Proof.** Note that

$$\sqrt{T} \left[ \hat{G}_{\theta}(\bar{\theta})' \hat{P}_{\alpha}(\theta) \hat{G}_{\theta}(\bar{\theta}) - G_{\theta} \alpha_{\alpha} G_{\theta} \right] = o_u P \left( T^{1/10} \right),$$

by Theorem 3 and Lemmas 6 and 9, using the algebraic properties of the $o_u P$ orders. The result then follows by a CLT for independent sequences since

$$\hat{J}_s(\bar{\theta}) - J_s = E \left[ (\hat{G}_{\theta}(\bar{\theta})' \hat{P}_{\alpha}(\theta) \hat{G}_{\theta}(\bar{\theta}) - G_{\theta} \alpha_{\alpha} G_{\theta} ) \right] = n^{-1} \sum_{i=1}^n (G_{\theta} \alpha_{\alpha} G_{\theta} - E) + o_u P \left( T^{-2/5} \right).$$

**Lemma 11.** Assume that Conditions 1, 2, 3, 4, 5, and 6 hold. Then, for any $\bar{\theta}$ between $\hat{\theta}$ and $\theta_0$,

$$\hat{B}_s(\bar{\theta}) = B_s + o_P(T^{-2/5}).$$

**Proof.** Analogous to the proof of Lemma 10 replacing $J_s$ by $B_s$. \hfill \Box

**Lemma 12.** Assume that Conditions 1, 2, 3, 4, 5, and 6 hold. Then, for any $\bar{\theta}$ between $\hat{\theta}$ and $\theta_0$, and $B = -J^{-1}B_s$,

$$\hat{B}(\bar{\theta}) = -\hat{J}_s(\bar{\theta})^{-1} \hat{B}_s(\bar{\theta}) = B + o_P(T^{-2/5}).$$

**Proof.** The result follows from Lemmas 10 and 11, using a Taylor expansion argument. \hfill \Box

**D.2. Proof of Theorem 5.**

**Proof.** Case I: $C = BC$. By Lemmas 10 and 25

$$\sqrt{nT} \left( \hat{\theta} - \theta_0 \right) = -\hat{J}_s(\theta_0)^{-1} \hat{s}(\theta_0) = -J_s^{-1} \hat{s}(\theta_0) + o_P(T^{-2/5})O_P \left( \sqrt{\frac{n}{T}} \right) = -J_s^{-1} \hat{s}(\theta_0) + o_P(1).$$

Then, by Lemmas 12 and 25

$$\sqrt{nT} \left( \hat{\theta}^{BC} - \theta_0 \right) = \sqrt{nT} \left( \hat{\theta} - \theta_0 \right) - \sqrt{nT} \frac{1}{J_s} \hat{B}_s(\theta) = -J_s^{-1} \hat{s}(\theta_0) + \sqrt{\frac{n}{T}} J_s^{-1} B_s + o_P(1)$$

$$= -J_s^{-1} \left[ \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_{\alpha i} + \sqrt{\frac{n}{T}} B_s - \sqrt{\frac{n}{T}} B_s \right] + o_P(1) \xrightarrow{\text{d}} N(0, J_s^{-1}).$$

Case II: $C = SBC$. First, note that since the correction of the score is of order $O_P(T^{-1})$, $\hat{\theta}^{SBC} - \hat{\theta} = O_P(T^{-1})$. Then, by a Taylor expansion of the corrected FOC around $\hat{\theta}^{SBC} = \theta_0$

$$0 = \hat{s}(\hat{\theta}^{SBC}) - T^{-1} \hat{B}_s(\hat{\theta}^{SBC}) = \hat{s}(\theta_0) + \hat{J}_s(\hat{\theta})(\hat{\theta}^{SBC} - \theta_0) - T^{-1} B_s + o_P(T^{-2}),$$

87
where $\bar{\theta}$ lies between $\hat{\theta}^{SBC}$ and $\theta_0$. Then by Lemma 25
\[
\sqrt{nT} \left( \hat{\theta}^{SBC} - \theta_0 \right) = -J_s(\bar{\theta})^{-1} \left[ \sqrt{nT} \hat{s}(\theta_0) - n^{1/2}T^{-1/2}B_s \right] + o_P(1)
\]
\[
= -J_s(\bar{\theta})^{-1} \left[ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \bar{\psi}_i + \sqrt{\frac{n}{T}} B_s - \sqrt{\frac{n}{T}} B_s \right] + o_P(1) \xrightarrow{d} N(0, J_s^{-1}).
\]

Case III: $C = IBC$. A similar argument applies to the estimating equation (5.2), since $\hat{\theta}^{IBC}$ is in a $O(T^{-1})$ neighborhood of $\theta_0$. 

APPENDIX E. STOCHASTIC EXPANSION FOR $\hat{\gamma}_{i0} = \hat{\gamma}_i(\theta_0)$ AND $\hat{\gamma}_{i0} = \hat{\gamma}_i(\theta_0)$

We characterize the stochastic expansions up to second order for one-step and two-step estimators of the individual effects given the true common parameter. We only provide detailed proofs of the results for the two-step estimator $\hat{\gamma}_{i0}$, because the proofs the one-step estimator $\hat{\gamma}_0$ follow by similar arguments. Lemmas 1 and 2 in the main text are corollaries of these expansions. The expressions for the scores and their derivatives in the components of the expansions are given in Appendix G.

Lemma 13. Suppose that Conditions 1, 2, 3, and 4 hold. Then
\[
\sqrt{T}(\hat{\gamma}_{i0} - \gamma_{i0}) = \bar{\psi}_i^W + T^{-1/2}R^W_{1i} \xrightarrow{d} N(0, V_i^W),
\]
where
\[
\bar{\psi}_i^W = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \psi_{it}^W = -(T_i^W)^{-1} \sqrt{T_i^W} = o_u(T^{1/10}), \quad R^W_{1i} = o_u(T^{1/5}), \quad V_i^W = E[\bar{\psi}_i^W \bar{\psi}_i^W'].
\]

Also
\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \bar{\psi}_i^W = O_P(1).
\]

Proof. We just show the part of the remainder term because the rest of the proof is similar to the proof of Lemma 16. By the proof of Lemma 5, $\sqrt{T}(\hat{\gamma}_{i0} - \gamma_{i0}) = o_u(T^{1/10})$ and
\[
R^W_{1i} = -(T_i^W)^{-1} \left( \tilde{T}_i^W(\theta_0, \tilde{\gamma}_{i0}) - T_i^W \right) \sqrt{T}(\tilde{\gamma}_{i0} - \gamma_{i0}) = o_u(T^{1/5}).
\]

Lemma 14. Suppose that Conditions 1, 2, 3, and 4 hold. Then,
\[
\sqrt{T}(\hat{\gamma}_{i0} - \gamma_{i0}) = \bar{\psi}_i^W + T^{-1/2}Q^W_{1i} + T^{-1}R^W_{2i},
\]
where
\[
Q^W_{1i} = -(T_i^W)^{-1} \left[ \tilde{A}_i^W \bar{\psi}_i^W + \frac{1}{2} \sum_{j=1}^{d_i + d_o} \tilde{\psi}_{ij}^W \tilde{T}_i^W \tilde{\psi}_{ij}^W \bar{\psi}_i^W \right] = o_u(T^{1/5}),
\]
\[
\tilde{A}_i^W = \sqrt{T}(\tilde{T}_i^W - T_i^W) = o_u(T^{1/10}), \quad R^W_{2i} = o_u(T^{3/10}).
\]

Also,
\[
\frac{1}{n} \sum_{i=1}^{n} Q^W_{1i} = O_P(1).
\]
Proof. Similar to the proof of Lemma 18.

Lemma 15. Suppose that Conditions 1, 2, 3, and 4 hold. Then,

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_i^W \rightarrow N(0, E[V_i^W]), \quad \frac{1}{n} \sum_{i=1}^{n} (\alpha_i)^W P_{\alpha_i}^W \rightarrow E[B_{\alpha_i}^W + B_{\alpha_i}^{W,G} + B_{\alpha_i}^{W,IS}] =: E_i^W,
\]

where

\[
V_i^W = \left( \frac{H_{\alpha_i}^W}{P_{\alpha_i}^W} \right) \Omega_i \left( \frac{H_{\alpha_i}^W}{P_{\alpha_i}^W} \right),
\]

\[
B_{\alpha_i}^{W,I} = \left( \frac{B_{\alpha_i}^{W,I}}{B_{\alpha_i}^{W,G}} \right) = \left( \frac{B_{\alpha_i}^W}{P_{\alpha_i}^W} \right) \left( \sum_{j=-\infty}^{\infty} E \left[ G_{\alpha_i}(z_{it})H_{\alpha_i}^W g(z_{it-j}) \right] - \frac{d_\alpha}{\sum_{j=1}^{d_\alpha} G_{\alpha_i,j}^W H_{\alpha_i}^W \Omega_i \alpha_i H_{\alpha_i}^W / 2 \right),
\]

\[
B_{\alpha_i}^{W,G} = \left( \frac{B_{\alpha_i}^{W,G}}{B_{\alpha_i}^{W,IS}} \right) = \left( \frac{-\Sigma_{\alpha_i}^W}{H_{\alpha_i}^W} \right) \sum_{j=-\infty}^{\infty} E \left[ G_{\alpha_i}(z_{it}) P_{\alpha_i}^W g(z_{it-j}) \right],
\]

\[
B_{\alpha_i}^{W,IS} = \left( \frac{B_{\alpha_i}^{W,IS}}{B_{\alpha_i}^{W,IS}} \right) = \left( \frac{-\Sigma_{\alpha_i}^W}{H_{\alpha_i}^W} \right) \sum_{j=-\infty}^{\infty} E \left[ G_{\alpha_i}(z_{it}) P_{\alpha_i}^W g(z_{it-j}) \right] + \left( \frac{H_{\alpha_i}^W}{P_{\alpha_i}^W} \right) \sum_{j=-\infty}^{\infty} E \left[ \xi_i(z_{it}) P_{\alpha_i}^W g(z_{it-j}) \right],
\]

for \( \Sigma_{\alpha_i}^W = (G_{\alpha_i}^W W_{\alpha_i}^{-1} G_{\alpha_i})^{-1}, H_{\alpha_i}^W = \Sigma_{\alpha_i}^W G_{\alpha_i}^W W_{\alpha_i}^{-1}, \) and \( P_{\alpha_i}^W = W_{\alpha_i}^{-1} - W_{\alpha_i}^{-1} G_{\alpha_i}^W H_{\alpha_i}^W. \)

Proof. The results follow from Lemmas 13 and 14, noting that

\[
(T_t^W)^{-1} = - \left( \left( -\sum_{\alpha_i}^W H_{\alpha_i}^W \right), \left( \frac{H_{\alpha_i}^W}{P_{\alpha_i}^W} \right) \right), \quad \psi_i^W = - \left( \frac{H_{\alpha_i}^W}{P_{\alpha_i}^W} \right) \hat{g}(z_{it}),
\]

\[
E \left[ \hat{\psi}_t^W \hat{\psi}_t^W \right] = \left( \frac{H_{\alpha_i}^W}{P_{\alpha_i}^W} \right) \Omega_i \left( \frac{H_{\alpha_i}^W}{P_{\alpha_i}^W} \right),
\]

\[
E \left[ A_t^W \hat{\psi}_t^W \right] = \sum_{j=-\infty}^{\infty} \left( E \left[ G_{\alpha_i}(z_{it}) H_{\alpha_i}^W g(z_{it-j}) \right] + E \left[ \xi_i(z_{it}) P_{\alpha_i}^W g(z_{it-j}) \right] \right),
\]

\[
E \left[ \hat{\psi}_t^W T_{t_t}^W \hat{\psi}_t^W \right] = \left\{ \begin{array}{ll}
G_{\alpha_i,j}^W P_{\alpha_i}^W \Omega_i H_{\alpha_i}^W, & \text{if } j \leq d_\alpha; \\
G_{\alpha_i,j}^W (I_{d_\alpha} \otimes e_{j-d_\alpha}) H_{\alpha_i}^W \Omega_i P_{\alpha_i,j}^W, & \text{if } j > d_\alpha.
\end{array} \right.
\]

Lemma 16. Suppose that Conditions 1, 2, 3, 4, 5, and 6 hold. Then,

\[
\sqrt{T}(\gamma_{t0} - \gamma_{t0}) = \tilde{\psi}_t + T^{-1/2} R_{1t} \rightarrow N(0, V_t),
\]

where

\[
\tilde{\psi}_t = \frac{1}{\sqrt{T}} \sum_{i=1}^{T} \psi_{it} = -(T_t^Q)^{-1} \sqrt{T_{t_t}^Q} = o_p \left( T^{1/10} \right), \quad R_{1t} = o_p \left( T^{1/5} \right), \quad V_t = E[\tilde{\psi}_t \tilde{\psi}_t^T].
\]

Also

\[
\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \tilde{\psi}_t = O_p(1)
\]

89
Proof. The statements about $\psi_i$ follow by the proof of Lemma 5 applied to the second stage, and the CLT in Lemma 3 of HK. From a similar argument to the proof of Lemma 5,

$$R_{1i} = - (T_i^{\Omega})^{-1} \sqrt{T} (\tilde{T}_i^{\Omega} (\theta_0, \bar{\gamma}_i) - T_i^{\Omega}) \sqrt{T} (\bar{\gamma}_0 - \gamma_0) - (T_i^{\Omega})^{-1} \sqrt{T} (\tilde{T}_i^{\Omega} (\theta_0, \bar{\gamma}_i) - T_i^{\Omega}) \sqrt{T} (\bar{\gamma}_0 - \gamma_0)$$

= $o_P(T^{1/10})$,

by Conditions 3 and 4.

Lemma 17. Assume that Conditions 1, 2, 3, 4 and 5 hold. Then,

$$\hat{\Omega}_i (\hat{\theta}, \tilde{\alpha}_i) = \Omega_i + T^{-1/2} \tilde{\psi}_i + T^{-1} R_{1ii}^{\Omega},$$

where

$$\tilde{\psi}_i = \sqrt{T} (\hat{\Omega}_i - \Omega_i) + \sum_{j=1}^{d_d} \Omega_{\alpha_{i,j}} \tilde{\psi}_{ij} = o_P(T^{1/10}), \quad R_{1ii}^{\Omega} = o_P(T^{1/5}),$$

and $\tilde{\psi}_{i,j}$ is the $j$th element of $\tilde{\psi}_i$.

Proof. By a mean value expansion around $(\theta_0, \alpha_0)$,

$$\hat{\Omega}_i (\hat{\theta}, \tilde{\alpha}_i) = \hat{\Omega}_i + \sum_{j=1}^{d_d} \hat{\Omega}_{\alpha_{i,j}} (\hat{\theta}, \tilde{\alpha}_i) (\tilde{\alpha}_{i,j} - \alpha_{i0,j}) + \sum_{j=1}^{d_d} \hat{\Omega}_{\theta_j} (\hat{\theta}, \tilde{\alpha}_i) (\tilde{\theta}_j - \theta_{0,j}),$$

where $(\theta, \alpha_i)$ lies between $(\hat{\theta}, \tilde{\alpha}_i)$ and $(\theta_0, \alpha_0)$. The expressions for $\tilde{\psi}_i^{\Omega}$ can be obtained using the expansions for $\tilde{\gamma}_i$ in Lemma 13 since $\tilde{\gamma}_i - \tilde{\gamma}_0 = o_P(T^{-3/10})$. The order of this term follows from Lemma 13 and the CLT for independent sequences. The remainder term is

$$R_{1ii}^{\Omega} = \sum_{j=1}^{d_d} [\Omega_{\alpha_{i,j}} R_{1i,j}^{\Omega} + \sqrt{T} (\hat{\Omega}_{\alpha_{i,j}} (\hat{\theta}, \tilde{\alpha}_i) - \Omega_{\alpha_{i,j}} ) \sqrt{T} (\tilde{\alpha}_{i,j} - \alpha_{i0,j} ) ] + \sum_{j=1}^{d_d} \hat{\Omega}_{\theta_j} (\hat{\theta}, \tilde{\alpha}_i) T (\tilde{\theta}_j - \theta_{0,j}).$$

The uniform rate of convergence then follows by Lemmas 8 and 13, and Theorem 1.

Lemma 18. Suppose that Conditions 1, 2, 3, 4, and 5 hold. Then,

$$(E.1) \quad \sqrt{T} (\tilde{\gamma}_i - \gamma_0) = \tilde{\psi}_i + T^{-1/2} Q_{1i} + T^{-1} R_{2i},$$

where

$$Q_{1i} (\tilde{\psi}_i, \tilde{\alpha}_i) = - (T_i^{\Omega})^{-1} \left[ A_i^{\Omega} \tilde{\psi}_i + \frac{1}{2} \sum_{j=1}^{d_d + d_n} \tilde{\psi}_{i,j} T_i^{\Omega} \tilde{\psi}_i + \text{diag}[0, \tilde{\psi}_{G_1}] \tilde{\psi}_i \right] = o_P \left( T^{1/5} \right),$$

$$A_i^{\Omega} = \sqrt{T} (\tilde{T}_i^{\Omega} - T_i^{\Omega}) = o_P \left( T^{1/10} \right), \quad R_{2i} = o_P \left( T^{3/10} \right).$$

Also,

$$\frac{1}{n} \sum_{i=1}^{n} Q_{1i} = O_P(1).$$

Proof. By a second order Taylor expansion of the FOC for $\tilde{\gamma}_i$, we have

$$0 = \tilde{\gamma}_i (\theta_0, \gamma_0) = \tilde{T}_i^{\Omega} + \tilde{T}_i (\tilde{\gamma}_0 - \gamma_0) + \frac{1}{2} \sum_{j=1}^{d_d} (\tilde{\gamma}_{0,j} - \gamma_{0,j}) \tilde{\Omega}_{i,j} (\theta_0, \tilde{\gamma}_i) (\tilde{\gamma}_0 - \gamma_0),$$

where $\tilde{\gamma}_i (\theta_0, \gamma_0) = \tilde{T}_i^{\Omega} + \tilde{T}_i (\tilde{\gamma}_0 - \gamma_0) + \frac{1}{2} \sum_{j=1}^{d_d} (\tilde{\gamma}_{0,j} - \gamma_{0,j}) \tilde{\Omega}_{i,j} (\theta_0, \tilde{\gamma}_i) (\tilde{\gamma}_0 - \gamma_0),$
where \( \bar{\gamma}_i \) is between \( \hat{\gamma}_{i0} \) and \( \gamma_{i0} \). The expression for \( Q_{1i} \) can be obtained in a similar fashion as in Lemma A4 in Newey and Smith (2004). The rest of the properties for \( Q_{1i} \) follow by Lemma 5 applied to the second stage, Lemma 16, and an argument similar to the proof of Theorem 1 in HK that uses Corollary A.2 of Hall and Heide (1980, p. 278) and Lemma 1 of Andrews (1991). The remainder term is

\[
R_{2i} = - (T_i^\Omega)^{-1} \left[ \bar{A}_i^\Omega R_{i1} + \sum_{j=1}^{d_i} \left[ R_{1i,j} T_{i,j}^\Omega \sqrt{T} (\bar{\gamma}_{i0} - \gamma_{i0}) + \bar{\psi}_{i,j} T_{i,j}^\Omega R_{i1} \right] / 2 \right] 
- (T_i^\Omega)^{-1} \sum_{j=1}^{d_i} \sqrt{T} (\bar{\gamma}_{i0,j} - \gamma_{i0,j}) \sqrt{T} (T_{i,j}^\Omega (\theta_0, \bar{\gamma}_i) - T_{i,j}^\Omega (\theta_0, \bar{\gamma}_i)) / 2
- (T_i^\Omega)^{-1} \left[ \text{diag}[0, R_i^W] \sqrt{T} (\bar{\gamma}_{i0} - \gamma_{i0}) + \text{diag}[0, \bar{\psi}_{i1}^W] R_{i1} \right].
\]

The uniform rate of convergence then follows by Lemmas 5 and 16, and Conditions 3 and 4.

Lemma 19. Suppose that Conditions 1, 2, 3, 4, 5, and 6 hold. Then,

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{\psi}_i \overset{d}{\rightarrow} N(0, \bar{E}[V_i]), \quad \frac{1}{n} \sum_{i=1}^{n} Q_{1i} \overset{p}{\rightarrow} \bar{E}[B_{i1}^L + B_{i1}^G + B_{i1}^Q + B_{i1}^W] =: B_i^*.
\]

where

\[
V_i = \text{diag}(\Sigma_{i1}, \Sigma_{i2}),
\]

\[
B_{i1}^L = \begin{pmatrix} B_{i1}^L \\ B_{i1}^L \end{pmatrix} = \begin{pmatrix} H_{i1} \\ P_{i1} \end{pmatrix} \begin{pmatrix} - \sum_{j=1}^{d_i} G_{i1,j} \Sigma_{i1} / 2 + E[G_{i1}(z_{i1})H_{i1}g(z_{i1})] \\ 0 \end{pmatrix},
\]

\[
B_{i1}^G = \begin{pmatrix} B_{i1}^G \\ B_{i1}^G \end{pmatrix} = \begin{pmatrix} - \Sigma_{i1} \\ H_{i1} \end{pmatrix} \sum_{j=0}^{\infty} E[G_{i1}(z_{i1})'P_{i1}g(z_{i1})],
\]

\[
B_{i1}^Q = \begin{pmatrix} B_{i1}^Q \\ B_{i1}^Q \end{pmatrix} = \begin{pmatrix} H_{i1} \\ P_{i1} \end{pmatrix} \sum_{j=0}^{\infty} E[g(z_{i1})g(z_{i1})'P_{i1}g(z_{i1})],
\]

\[
B_{i1}^W = \begin{pmatrix} B_{i1}^W \\ B_{i1}^W \end{pmatrix} = \begin{pmatrix} H_{i1} \\ P_{i1} \end{pmatrix} \sum_{j=1}^{d_i} \Omega_{i1,j} \left(H_{i1,j} - H_{i1,j} \right),
\]

for \( \Sigma_{i1} = (G_{i1}'\Omega_i^{-1}G_{i1})^{-1} \), \( H_{i1} = \Sigma_{i1}G_{i1}'\Omega_i^{-1} \), and \( P_{i1} = \Omega_i^{-1} - \Sigma_i^{-1}G_{i1}H_{i1} \).

Proof. The results follow by Lemmas 16 and 18, noting that

\[
(T_i^\Omega)^{-1} = - \begin{pmatrix} - \Sigma_{i1} & H_{i1} \\ H_{i1}' & P_{i1} \end{pmatrix}, \quad \psi_i = - \begin{pmatrix} H_{i1} \\ P_{i1} \end{pmatrix} g(z_{i1}),
\]

\[
E[\tilde{\psi}_i' \tilde{\psi}_i] = \begin{pmatrix} \Sigma_{i1} & 0 \\ 0 & P_{i1} \end{pmatrix}, \quad E[\bar{A}_i^\Omega \bar{\psi}_i] = \sum_{j=0}^{\infty} \begin{pmatrix} E[G_{i1}(z_{i1})'P_{i1}g(z_{i1})] \\ E[G_{i1}(z_{i1})'H_{i1}g(z_{i1})] \end{pmatrix},
\]

\[
E[\tilde{\psi}_{i,j} T_{i,j}^\Omega \bar{\psi}_i] = \begin{cases} 0, & \text{if } j \leq d_i; \\ G_{i1,j} \Sigma_{i1}, & \text{if } j > d_i. \end{cases}
\]

\[
E[\text{diag}[0, \bar{\psi}_{i1}^W] \bar{\psi}_i] = \begin{pmatrix} 0 \\ \sum_{j=0}^{d_i} E[g(z_{i1})g(z_{i1})'P_{i1}g(z_{i1})] + \sum_{j=1}^{d_i} \Omega_{i1,j} (H_{i1,j} - H_{i1,j}) \end{pmatrix}.
\]
APPENDIX F. STOCHASTIC EXPANSION FOR $\tilde{s}^W_i(\theta, \gamma_{i0})$ AND $\tilde{s}_i(\theta, \gamma_{i0})$

We characterize stochastic expansions up to second order for one-step and two-step profile scores of the common parameter evaluated at the true value of the common parameter. The expressions for the scores and their derivatives in the components of the expansions are given in Appendix G.

Lemma 20. Suppose that Conditions 1, 2, 3, and 4 hold. Then,

$$\tilde{s}^W_i(\theta, \gamma_{i0}) = T^{-1/2} \tilde{\psi}_{si}^W + T^{-1/2} Q_{isi}^W + T^{-3/2} R_{2si}^W,$$

where

$$\tilde{\psi}_{si}^W = M_i^W \tilde{\psi}_{si}^W + o_u(T^{1/1}) \quad Q_{isi}^W = M_i^W Q_{isi}^W + \tilde{C}_{si}^W \tilde{\psi}_{si}^W + \frac{1}{2} \sum_{j=1}^{d_s+d_a} \tilde{\psi}_{ij}^W M_{ij}^W \tilde{\psi}_{ij}^W + o_u(T^{1/2}),$$

$$\tilde{C}_{si}^W = \sqrt{T}(M_i^W - M_i^W) = o_u(T^{1/10}), \quad R_{2si}^W = o_u(T^{2/5}).$$

Also,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{\psi}_{si}^W = O_P(1), \quad \frac{1}{n} \sum_{i=1}^{n} Q_{isi}^W = O_P(1).$$

Proof. By a second order Taylor expansion of $\tilde{s}^W_i(\theta, \gamma_{i0})$ around $\gamma_{i0} = \gamma_{i0}$,

$$\tilde{s}^W_i(\theta, \gamma_{i0}) = \tilde{s}^W_i + M_i^W(\gamma_{i0} - \gamma_{i0}) + \frac{1}{2} \sum_{i=1}^{d_s+d_a} (\tilde{\psi}_{ij}^W - \gamma_{ij}^W) \tilde{M}_{ij}^W(\theta, \gamma_i)(\gamma_{i0} - \gamma_{i0}),$$

where $\gamma_i$ is between $\gamma_{i0}$ and $\gamma_{i0}$. Noting that $\tilde{s}^W_i(\theta, \gamma_{i0}) = 0$ and using the expansion for $\gamma_{i0}$ in Lemma 14, we can obtain the expressions for $\tilde{\psi}_{si}^W$ and $Q_{isi}^W$, after some algebra. The rest of the properties for these terms follow by the properties of $\tilde{\psi}_{si}^W$ and $Q_{isi}^W$. The remainder term is

$$R_{2si}^W = M_i^W R_{2i}^W + \tilde{C}_{si}^W R_{2i}^W + \frac{1}{2} \sum_{i=1}^{d_s+d_a} \left[ R_{1i,j}^W M_{ij}^W \sqrt{T}(\gamma_{i0} - \gamma_{i0}) + \tilde{\psi}_{ij}^W M_{ij}^W R_{1i}^W \right]$$

$$+ \frac{1}{2} \sum_{i=1}^{d_s+d_a} \sqrt{T}(\gamma_{i0,j} - \gamma_{i0,j}) \sqrt{T}(M_{ij}^W(\theta, \gamma_i) - M_{ij}^W(\theta, \gamma_i)) \sqrt{T}(\gamma_{i0} - \gamma_{i0}).$$

The uniform order of $R_{2si}^W$ follows by the properties of the components in the expansion of $\gamma_{i0}$, Lemma 5, and Conditions 3 and 4.

Lemma 21. Suppose that Conditions 1, 2, 3, and 4 hold. We then have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{\psi}_{si}^W \rightarrow_d N(0, V_s^W), \quad V_s^W = \tilde{E}\left[ G_{\theta_i}^W \tilde{P}_{\alpha_{i}}^W \tilde{O}_{i}^W \tilde{P}_{\alpha_{i}}^W G_{\theta_i}^W \right],$$

$$\frac{1}{n} \sum_{i=1}^{n} Q_{isi}^W \rightarrow_{a} \tilde{E}\left[ Q_{isi}^W \right] = \tilde{E}\left[ B_{si}^{W,B} + B_{si}^{W,C} + B_{si}^{W,V} \right]:= B_{s}^W,$$

where $B_{si}^{W,B} = -G_{\theta_i}^W B_{\lambda_i} = \tilde{E}\left[ G_{\theta_i}^W \tilde{P}_{\alpha_{i}}^W \tilde{O}_{i}^W \tilde{P}_{\alpha_{i}}^W G_{\theta_i}^W \right], \quad B_{si}^{W,C} = \tilde{E}\left[ B_{si}^{W,L} + B_{si}^{W,C} + B_{si}^{W,L,S} \right], \quad B_{si}^{W,V} = \sum_{j=1}^{d_s} \tilde{E}\left[ G_{\alpha_{i}} W_{i}^W H_{i}^W / 2 - \sum_{j=1}^{d_s} (\tilde{G}_{\theta_i}^W (I_{d_s} \otimes e_j) H_{i}^W P_{i}^W / 2, H_{i}^W = \Sigma_{\alpha_i} G_{\alpha_i} W_{i}^W, \Sigma_{\alpha_i} = (G_{\alpha_i} W_{i}^W)^{-1} \tilde{G}_{\alpha_i}, \tilde{P}_{\tilde{W}_{i}^W} = W_{i}^{-1} - W_{i}^{-1} G_{\alpha_i} H_{\alpha_i}.\right]
Proof. The results follow by Lemmas 20 and 15, noting that
\[
E \left[ \psi_{st}^W \psi_{st}^W \right] = M_s^W \left( \begin{array}{cc} H_s^W \Omega_s H_s^W & H_s^W \Omega_s P_s^W \\ P_s^W \Omega_s H_s^W & P_s^W \Omega_s P_s^W \end{array} \right) M_s^W,
\]
\[
E \left[ \psi_{st}^W \psi_{st}^W \right] = \sum_{j=-\infty}^{\infty} E \left[ G_{\theta_s}(z_{st})' P_{\alpha_s} g(z_{st-j}) \right],
\]
\[
E \left[ \psi_{st}^W \psi_{st}^W \right] = \left\{ \begin{array}{ll}
-G_{\theta_s}(z_{st-j})' P_{\alpha_s} H_s^W, & \text{if } j \leq d_s; \\
-G_{\theta_s}(I_{d_s} \otimes e_{j-d_s}) H_s^W \Omega_s P_s^W, & \text{if } j > d_s.
\end{array} \right.
\]

Lemma 22. Suppose that Conditions 1, 2, 3, and 4 hold. Then, for \( S^W(\theta_0) = n^{-1} \sum_{i=1}^{n} \tilde{s}^W_i(\theta_0, \gamma_0) \),
\[
\sqrt{nT} S^W(\theta_0) \xrightarrow{d} N \left( \kappa B_s^W, V_s^W \right),
\]
where \( B_s^W \) and \( V_s^W \) are defined in Lemma 21.

Proof. By Lemma 20,
\[
\sqrt{nT} S^W(\theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{\psi}_{si} + \frac{\sqrt{n}}{T} \frac{1}{n} \sum_{i=1}^{n} Q_{1st}^W + \frac{\sqrt{n}}{T^2} \frac{1}{n} \sum_{i=1}^{n} R_{2st}^W = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{\psi}_{si} + \frac{\sqrt{n}}{T} \frac{1}{n} \sum_{i=1}^{n} Q_{1st}^W + o_p(1).
\]
Then, the result follows by Lemma 21.

Lemma 23. Suppose that Conditions 1, 2, 3, 4, 5, and 6 hold. Then,
\[
\tilde{\psi}_{i}(\theta_0, \gamma_0) = T^{-1/2} \tilde{\psi}_{si} + T^{-1} Q_{1st}^W + T^{-3/2} R_{2st}^W,
\]
where all the terms are identical to that of Lemma 20 after replacing \( W \) by \( \Omega \). Also, the properties of all the terms of the expansion are the analogous to those of Lemma 20.

Proof. The proof is similar to the proof of Lemma 20.

Lemma 24. Suppose that Conditions 1, 2, 3, 4, 5, and 6 hold. Then,
\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{\psi}_{si} \xrightarrow{d} N(0, J_s), \quad J_s = E[G_{\theta_s}' P_{\alpha_s} G_{\theta_s}]
\]
\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Q_{1st}^W \xrightarrow{p} EE[Q_{1st}^W] = E[B_{si}^B + B_{si}^G] =: B_s,
\]
where \( B_{si}^B = -G_{\theta_s}' (B_{si}^L + B_{si}^G + B_{si}^\alpha + B_{si}^W), B_{si}^G = \sum_{j=0}^{\infty} E[G_{\theta_s}(z_{st})' P_{\alpha_s} g(z_{st-j})], P_{\alpha_s} = \Omega_s^{-1} - \Omega_s^{-1} G_{\alpha_s} H_{\alpha_s}, H_{\alpha_s} = \Sigma_{\alpha, G_{\alpha_s}' \Omega_s^{-1} G_{\alpha_s}}, \) and \( \Sigma_{\alpha} = (G_{\alpha_s}' \Omega_s^{-1} G_{\alpha_s})^{-1}. \)

Proof. The results follow by Lemmas 16, 18, 19 and 23, noting that
\[
E \left[ \psi_{st}^W \psi_{st}^W \right] = M_s^\Omega \left( \begin{array}{cc} \Sigma_{\alpha_s} & 0 \\ 0 & P_{\alpha_s} \end{array} \right) M_s^\Omega', \quad E \left[ \psi_{st}^W \psi_{st}^W \right] = \sum_{j=0}^{\infty} E[G_{\theta_s}(z_{st})' P_{\alpha_s} g(z_{st-j})], \quad E \left[ \psi_{st}^W \psi_{st}^W \right] = 0.
\]
Lemma 25. Suppose that Conditions 1, 2, 3, 5, and 4 hold. Then, or \( \hat{\mathcal{S}}(\theta_0) = n^{-1} \sum_{i=1}^{n} \hat{\mathcal{S}}_i(\theta_0, \hat{\gamma}_0) \),

\[
\sqrt{n \hat{\mathcal{S}}(\theta_0)} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{\psi}_i + \sqrt{\frac{n}{T}} B \sigma + o_p(1) \Rightarrow N(0B_s, \mathcal{J}_s),
\]

where \( \tilde{\psi}_i \) and \( B \) are defined in Lemmas 23 and 24, respectively.

Proof. Using the expansion form obtained in Lemma 23, we can get the result by examining each term with Lemma 24.

\[\square\]

APPENDIX G. SCORES AND DERIVATIVES

G.1. One-Step Score and Derivatives: Individual Effects. We denote dimensions of \( \mathcal{G}(z_{it}), \alpha_i \), and \( \theta \) by \( d_g, d_\alpha \), and \( d_\theta \). The symbol \( \otimes \) denotes kronecker product of matrices, and \( I_{d_\alpha} \) denotes a \( d_\alpha \)-order identity matrix. Let \( \mathcal{G}_{a\alpha,i}(z_{it}; \theta, \alpha_i) := (\mathcal{G}_{a\alpha,i,1}(z_{it}; \theta, \alpha_i), \ldots, \mathcal{G}_{a\alpha,i,d_\alpha}(z_{it}; \theta, \alpha_i))' \), where

\[
\mathcal{G}_{a\alpha,i,j}(z_{it}; \theta, \alpha_i) = \frac{\partial \mathcal{G}_{\alpha,i}(z_{it}; \theta, \alpha_i)}{\partial \alpha_i}.
\]

We denote derivatives of \( \mathcal{G}_{a\alpha,i}(z_{it}; \theta, \alpha_i) \) with respect to \( \alpha_i \) by \( \mathcal{G}_{a\alpha,i}(z_{it}; \theta, \alpha_i) \), and use additional subscripts for higher order derivatives.

G.1.1. Score.

\[
\widehat{t}_i^W(\gamma_i) = -\frac{1}{T} \sum_{t=1}^{T} \left( \mathcal{G}_{\alpha,i}(z_{it}; \theta, \alpha_i)' \gamma_i \right) = \left( \hat{\mathcal{G}}_{\alpha,i}(\theta, \alpha_i)' \gamma_i \right).\]

G.1.2. Derivatives with respect to the fixed effects.

First Derivatives

\[
\widehat{T}_i^W(\gamma_i) = \frac{\partial \widehat{t}_i^W(\gamma_i, \theta)}{\partial \gamma_i} = \left( \hat{\mathcal{G}}_{\alpha,i}(\theta, \alpha_i)'(I_{d_\alpha} \otimes \lambda_i) \right) \left( \hat{\mathcal{G}}_{\alpha,i}(\theta, \alpha_i)'(I_{d_\alpha} \otimes \lambda_i) \right)^{-1}.
\]

\[
T_i^W = E[\hat{t}_i^W] = \left( \begin{array}{c}
0 \\
G_{\alpha,i} W_i
\end{array} \right).
\]

\[
(T_i^W)^{-1} = \left( \begin{array}{c}
-H_{\alpha,i} W_i \\
-P_{\alpha,i} W_i
\end{array} \right).
\]

Second Derivatives

\[
\widehat{T}_{i,j}^W(\gamma_i) = \frac{\partial^2 \hat{t}_i^W(\gamma_i, \theta)}{\partial \gamma_{ij} \partial \gamma_{k}} = \left( \begin{array}{c}
-\hat{G}_{a\alpha,i,j}(\theta, \alpha_i)'(I_{d_\alpha} \otimes \lambda_i) \hat{G}_{a\alpha,i,j}(\theta, \alpha_i)'(I_{d_\alpha} \otimes \lambda_i) \\
0
\end{array} \right), \quad \text{if } j \leq d_\alpha;
\]

\[
\widehat{T}_{i,j}^W(\gamma_i) = \left( \begin{array}{c}
-\hat{G}_{a\alpha,i,j}(\theta, \alpha_i)'(I_{d_\alpha} \otimes \lambda_i) \\
0
\end{array} \right), \quad \text{if } j > d_\alpha.
\]

\[
T_{i,j}^W = E[\hat{t}_{i,j}^W(\gamma_0; \theta_0)] = \left( \begin{array}{c}
0 \\
G_{a\alpha,i,j}
\end{array} \right), \quad \text{if } j \leq d_\alpha;
\]

\[
T_{i,j}^W = \left( \begin{array}{c}
0 \\
G_{a\alpha,i,j}(I_{d_\alpha} \otimes e_{j-d_\alpha})
\end{array} \right), \quad \text{if } j > d_\alpha.
\]

94
Third Derivatives

\[
\tilde{T}_{i,j,k}^W(\theta, \gamma_i) = \frac{\partial^2 \tilde{T}_{i,j}^W(\theta, \gamma_i)}{\partial \gamma_{i,k} \partial \gamma_{i,j} \partial \gamma_{i}} = \begin{cases}
- \left( \begin{array}{c}
\widehat{G}_{aa,aa,jk}(\theta, \alpha_i)' (I_{d_a} \otimes \lambda_i) \\
\widehat{G}_{aa,aa,jk}(\theta, \alpha_i)' (I_{d_a} \otimes e_{k-d_a}) \\
\widehat{G}_{aa,aa,jk}(\theta, \alpha_i)' (I_{d_a} \otimes e_{j-d_a}) \\
0
\end{array} \right), & \text{if } j \leq d_a, k \leq d_a; \\
- \left( \begin{array}{c}
0 \\
0 \\
0 \\
0
\end{array} \right), & \text{if } j > d_a, k > d_a;
\end{cases}
\]

\[
T_{i,j,k}^W = E \left[ \tilde{T}_{i,j,k}^W \right] = \begin{cases}
\left( \begin{array}{c}
0 \\
0 \\
0 \\
0
\end{array} \right), & \text{if } j \leq d_a, k \leq d_a; \\
\left( \begin{array}{c}
\widehat{G}_{aa,aa,jk} \\
0 \\
\widehat{G}_{aa,aa,jk}' (I_{d_a} \otimes e_{k-d_a}) \\
\widehat{G}_{aa,aa,jk}' (I_{d_a} \otimes e_{j-d_a}) \\
0
\end{array} \right), & \text{if } j \leq d_a, k > d_a; \\
\left( \begin{array}{c}
0 \\
0 \\
0 \\
0
\end{array} \right), & \text{if } j > d_a, k \leq d_a; \\
\left( \begin{array}{c}
0 \\
0 \\
0 \\
0
\end{array} \right), & \text{if } j > d_a, k > d_a.
\end{cases}
\]

G.1.3. Derivatives with respect to the common parameter.

First Derivatives

\[
\tilde{N}_{i,j}^W(\theta, \gamma_i) = \frac{\partial \tilde{N}_{i,j}^W(\gamma_i, \theta)}{\partial \theta_j} = - \left( \begin{array}{c}
\widehat{G}_{\theta_j,\alpha_i}(\theta, \alpha_i)' \lambda_i \\
\widehat{G}_{\theta_j,\alpha_i}(\theta, \alpha_i)' \lambda_i
\end{array} \right),
\]

\[
N_{i,j}^W = E \left[ \tilde{N}_{i,j}^W \right] = \begin{cases}
0, & \text{if } j \leq d_a, k \leq d_a; \\
\left( \begin{array}{c}
0 \\
0
\end{array} \right), & \text{if } j \leq d_a, k > d_a; \\
\left( \begin{array}{c}
0 \\
0
\end{array} \right), & \text{if } j > d_a, k \leq d_a; \\
\left( \begin{array}{c}
0 \\
0
\end{array} \right), & \text{if } j > d_a, k > d_a.
\end{cases}
\]

G.2. One-Step Score and Derivatives: Common Parameters. Let \( G_{\theta,\alpha_i}(z_{it}; \theta, \alpha_i) := (G_{\theta,\alpha_i,1}(z_{it}; \theta, \alpha_i)', \ldots, G_{\theta,\alpha_{i,a}}(z_{it}; \theta, \alpha_i))' \), where

\[
G_{\theta,\alpha_i,1}(z_{it}; \theta, \alpha_i) = \frac{\partial G_{\theta}(z_{it}; \theta, \alpha_i)}{\partial \alpha_i}.
\]

We denote the derivatives of \( G_{\theta,\alpha_i}(z_{it}; \theta, \alpha_i) \) with respect to \( \alpha_{i,j} \) by \( G_{\theta,\alpha_{i,j}}(z_{it}; \theta, \alpha_i) \), and use additional subscripts for higher order derivatives.

G.2.1. Score.

\[
\tilde{z}_{i,j}^W(\theta, \gamma_i) = - \frac{1}{T} \sum_{t=1}^{T} G_{\theta}(z_{it}; \theta, \alpha_i)' \lambda_i = - \widehat{G}_{\theta}(\theta, \alpha_i)' \lambda_i.
\]

G.2.2. Derivatives with respect to the fixed effects.

First Derivatives
\[
\bar{M}_i^W(\theta, \gamma_i) = \frac{\partial \hat{\theta}_i^W(\theta, \gamma_i)}{\partial \gamma_i} = - \left( \begin{array}{c}
\hat{G}_{\theta_0}(\theta, \alpha_i)'(I_{d_\alpha}\otimes \lambda_i) \\
\hat{G}_{\theta_1}(\theta, \alpha_i)'
\end{array} \right).
\]
\[
M_i^W = E \left[ \bar{M}_i^W \right] = \left( \begin{array}{c} 0 \\
G_{\theta_0}'
\end{array} \right).
\]

Second Derivatives

\[
\bar{M}_{i,j}^W(\theta, \gamma_i) = \frac{\partial^2 \hat{\theta}_i^W(\theta, \gamma_i)}{\partial \gamma_i \partial \gamma_j} = \left\{ \begin{array}{l}
- \left( \begin{array}{c}
\hat{G}_{\theta_0,\alpha_{i,j}}(\theta, \alpha_i)'(I_{d_\alpha}\otimes \lambda_i) \\
\hat{G}_{\theta_0,\alpha_{i,j}}(\theta, \alpha_i)'
\end{array} \right), \\
- \left( \begin{array}{c}
\hat{G}_{\theta_0,\alpha_{i,j}}(\theta, \alpha_i)'(I_{d_\alpha}\otimes e_{j-d_\alpha}) \\
0
\end{array} \right), \\
- \left( \begin{array}{c}
\hat{G}_{\theta_0,\alpha_{i,j}}(\theta, \alpha_i)'(I_{d_\alpha}\otimes e_{j-d_\alpha}) \\
0
\end{array} \right), \\
\left( \begin{array}{c} 0 \\
0
\end{array} \right),
\end{array} \right\},
\]

\[
M_{i,j}^W = E \left[ \bar{M}_{i,j}^W(\theta_0, \gamma_0) \right] = \left\{ \begin{array}{l}
- \left( \begin{array}{c}
0 \\
G_{\theta_0,\alpha_{i,j}}'
\end{array} \right), \\
- \left( \begin{array}{c}
G_{\theta_0,\alpha_{i,j}}(I_{d_\alpha}\otimes e_{j-d_\alpha}) \\
0
\end{array} \right), \\
- \left( \begin{array}{c}
G_{\theta_0,\alpha_{i,j}}(I_{d_\alpha}\otimes e_{j-d_\alpha}) \\
0
\end{array} \right), \\
\left( \begin{array}{c} 0 \\
0
\end{array} \right),
\end{array} \right\},
\]

Third Derivatives

\[
\bar{M}_{i,j,k}^W(\theta, \gamma_i) = \frac{\partial^3 \hat{\theta}_i^W(\theta, \gamma_i)}{\partial \gamma_{i,k} \partial \gamma_{i,j} \partial \gamma_i} = \left\{ \begin{array}{l}
- \left( \begin{array}{c}
\hat{G}_{\theta_0,\alpha_{i,j,k}}(\theta, \alpha_i)'(I_{d_\alpha}\otimes \lambda_i) \\
\hat{G}_{\theta_0,\alpha_{i,j,k}}(\theta, \alpha_i)'
\end{array} \right), \\
- \left( \begin{array}{c}
\hat{G}_{\theta_0,\alpha_{i,j,k}}(\theta, \alpha_i)'(I_{d_\alpha}\otimes e_{k-d_\alpha}) \\
0
\end{array} \right), \\
- \left( \begin{array}{c}
\hat{G}_{\theta_0,\alpha_{i,j,k}}(\theta, \alpha_i)'(I_{d_\alpha}\otimes e_{j-d_\alpha}) \\
0
\end{array} \right), \\
\left( \begin{array}{c} 0 \\
0
\end{array} \right),
\end{array} \right\},
\]

\[
M_{i,j,k}^W = E \left[ \bar{M}_{i,j,k}^W \right] = \left\{ \begin{array}{l}
- \left( \begin{array}{c}
0 \\
G_{\theta_0,\alpha_{i,j,k}}'
\end{array} \right), \\
- \left( \begin{array}{c}
G_{\theta_0,\alpha_{i,j,k}}(I_{d_\alpha}\otimes e_{k-d_\alpha}) \\
0
\end{array} \right), \\
- \left( \begin{array}{c}
G_{\theta_0,\alpha_{i,j,k}}(I_{d_\alpha}\otimes e_{j-d_\alpha}) \\
0
\end{array} \right), \\
\left( \begin{array}{c} 0 \\
0
\end{array} \right),
\end{array} \right\},
\]

G.2.3. Derivatives with respect to the common parameters.

First Derivatives

\[
\hat{S}_i^W(\theta, \gamma_i) = \frac{\partial \hat{\theta}_i^W(\theta, \gamma_i)}{\partial \theta_j} = -\hat{G}_{\theta_0,i}(\theta, \alpha_i)'\lambda_i.
\]
\[
S_i^W = E \left[ \hat{S}_i^W \right] = 0.
\]

G.3. Two-Step Score and Derivatives: Fixed Effects.

G.3.1. Score.

\[
\hat{e}_i(\theta, \gamma_i) = -\frac{1}{T} \sum_{t=1}^{T} \left( \begin{array}{c}
g(z_{it}; \theta, \alpha_i) + \hat{\lambda}_i(\theta, \alpha_i)\lambda_i \\
\hat{G}_{\alpha_{i,t}}(\theta, \alpha_i)'\lambda_i \\
\hat{G}_{\alpha_{i,t}}(\theta, \alpha_i) + \Omega_{i,t}\lambda_i \\
0
\end{array} \right) = \hat{\mathcal{M}}_i^W(\theta, \gamma_i) + \hat{\mathcal{P}}_i^R(\theta, \gamma_i).
\]

Note that the formulae for the derivatives of Appendix G.1 apply for \( \hat{\mathcal{M}}_i^W \) replacing \( \hat{\mathcal{W}} \) by \( \Omega \). Hence, we only need to derive the derivatives for \( \hat{\mathcal{P}}_i^R \).
G.3.2. Derivatives with respect to the fixed effects.

First Derivatives

\[
\mathcal{T}_i^R(\theta, \gamma_i) = \frac{\partial \hat{R}(\theta, \gamma_i)}{\partial \gamma_i} = -\begin{pmatrix} 0 & 0 \\ 0 & \hat{\Omega}_i(\theta, \bar{a}_i) - \Omega_i \end{pmatrix}.
\]

\[
T_i^R = E[\mathcal{T}_i^R] = -\begin{pmatrix} 0 & 0 \\ 0 & E[\hat{\Omega}_i - \Omega_i] \end{pmatrix}.
\]

Second and Third Derivatives

Since \( \mathcal{T}_i^R(\gamma_i, \theta) \) does not depend on \( \gamma_i \), the derivatives (and its expectation) of order greater than one are zero.

G.3.3. Derivatives with respect to the common parameters.

First Derivatives

\[
\mathcal{N}_i^R(\theta, \gamma_i) = \frac{\partial \hat{N}(\theta, \gamma_i)}{\partial \theta'} = 0.
\]

G.4. Two-Step Score and Derivatives: Common Parameters.

G.4.1. Score.

\[
\hat{\gamma}_i(\theta, \gamma_i) = -\frac{1}{T} \sum_{t=1}^{T} G_\theta(z_{it}; \theta, \alpha_i)' \lambda_i = -\hat{G}_\theta(\theta, \alpha_i)' \lambda_i.
\]

Since this score does not depend explicitly on \( \hat{\Omega}_i(\theta, \bar{a}_i) \), the formulae for the derivatives are the same as in Appendix G.2.

References


<table>
<thead>
<tr>
<th>Estimator</th>
<th>Bias</th>
<th>SD</th>
<th>SE/SD</th>
<th>p; .05</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS-FC</td>
<td>0.06</td>
<td>0.01</td>
<td>0.84</td>
<td>0.06</td>
</tr>
<tr>
<td>IV-FC</td>
<td>0.00</td>
<td>0.01</td>
<td>0.90</td>
<td>0.08</td>
</tr>
<tr>
<td>OLS-RC</td>
<td>0.04</td>
<td>0.01</td>
<td>0.97</td>
<td>1.00</td>
</tr>
<tr>
<td>BC</td>
<td>0.04</td>
<td>0.01</td>
<td>0.97</td>
<td>1.00</td>
</tr>
<tr>
<td>IBC</td>
<td>0.04</td>
<td>0.01</td>
<td>0.97</td>
<td>1.00</td>
</tr>
<tr>
<td>IV-RC</td>
<td>0.00</td>
<td>0.01</td>
<td>1.00</td>
<td>0.05</td>
</tr>
<tr>
<td>BC-IV</td>
<td>0.00</td>
<td>0.01</td>
<td>1.00</td>
<td>0.05</td>
</tr>
<tr>
<td>IBC-IV</td>
<td>0.00</td>
<td>0.01</td>
<td>1.00</td>
<td>0.05</td>
</tr>
<tr>
<td>OLS-FC</td>
<td>0.12</td>
<td>0.01</td>
<td>1.09</td>
<td>1.00</td>
</tr>
<tr>
<td>IV-FC</td>
<td>0.00</td>
<td>0.02</td>
<td>0.94</td>
<td>0.07</td>
</tr>
<tr>
<td>OLS-RC</td>
<td>0.10</td>
<td>0.01</td>
<td>1.06</td>
<td>1.00</td>
</tr>
<tr>
<td>BC</td>
<td>0.10</td>
<td>0.01</td>
<td>1.06</td>
<td>1.00</td>
</tr>
<tr>
<td>IBC</td>
<td>0.10</td>
<td>0.01</td>
<td>1.06</td>
<td>1.00</td>
</tr>
<tr>
<td>IV-RC</td>
<td>0.00</td>
<td>0.02</td>
<td>0.98</td>
<td>0.06</td>
</tr>
<tr>
<td>BC-IV</td>
<td>0.00</td>
<td>0.02</td>
<td>0.97</td>
<td>0.05</td>
</tr>
<tr>
<td>IBC-IV</td>
<td>0.00</td>
<td>0.02</td>
<td>0.97</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Note: 1,000 repetitions. RC/FC refers to random/fixed coefficient model. BC/IBC refers to bias corrected/iterated bias corrected estimates.

p; = \phi

Table A1: Common Parameter \theta
Note: 1,000 repetitions. RC/FC refers to random/fixed coefficient model. BC/IBC refers to bias corrected/iterated bias corrected estimates.

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean</th>
<th>SD</th>
<th>SE/SD</th>
<th>p &gt; .05</th>
<th>Bias</th>
<th>SE</th>
<th>SD</th>
<th>SE/SD</th>
<th>p &gt; .05</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBC - IV</td>
<td>0.94</td>
<td>0.06</td>
<td>0.09</td>
<td>0.06</td>
<td>0.94</td>
<td>0.06</td>
<td>0.09</td>
<td>0.06</td>
<td>0.94</td>
</tr>
<tr>
<td>BC - IV</td>
<td>1.07</td>
<td>0.12</td>
<td>0.15</td>
<td>0.12</td>
<td>1.07</td>
<td>0.12</td>
<td>0.15</td>
<td>0.12</td>
<td>1.07</td>
</tr>
<tr>
<td>IVC - IV</td>
<td>0.94</td>
<td>0.06</td>
<td>0.09</td>
<td>0.06</td>
<td>0.94</td>
<td>0.06</td>
<td>0.09</td>
<td>0.06</td>
<td>0.94</td>
</tr>
</tbody>
</table>

Table A2: Mean of Individual Specific Parameter, P = 0.3
### Table A3: Standard Deviation of the Individual Specific Parameter

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Bias SD SE/SD p105 0.4</th>
<th>Bias SD SE/SD p105 0.4</th>
<th>Bias SD SE/SD p105 0.4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: 1,000 repetitions. RC/FC refers to random/fixed coefficient model. BC/IBC refers to bias corrected/iterated bias corrected estimates.