THREE ESSAYS ON PANEL DATA MODELS WITH INTERACTIVE AND UNOBSERVED EFFECTS

By

Nicholas Lynn Brown

A DISSERTATION

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

Economics – Doctor of Philosophy

2022

ABSTRACT

THREE ESSAYS ON PANEL DATA MODELS WITH INTERACTIVE AND UNOBSERVED EFFECTS

By

Nicholas Lynn Brown

Chapter 1: More Efficient Estimation of Multiplicative Panel Data Models in the Presence of Serial Correlation (with Jeffrey Wooldridge)

We provide a systematic approach in obtaining an estimator asymptotically more efficient than the popular fixed effects Poisson (FEP) estimator for panel data models with multiplicative heterogeneity in the conditional mean. In particular, we derive the optimal instrumental variables under appealing 'working' second moment assumptions that allow underdispersion, overdispersion, and general patterns of serial correlation. Because parameters in the optimal instruments must be estimated, we argue for combining our new moment conditions with those that define the FEP estimator to obtain a generalized method of moments (GMM) estimator no less efficient than the FEP estimator and the estimator using the new instruments. A simulation study shows that the GMM estimator behaves well in terms of bias, and it often delivers nontrivial efficiency gains – even when the working second-moment assumptions fail.

Chapter 2: Information equivalence among transformations of semiparametric nonlinear panel data models

I consider transformations of nonlinear semiparametric mean functions which yield moment conditions for estimation. Such transformations are said to be information equivalent if they yield the same asymptotic efficiency bound. I first derive a unified theory of algebraic equivalence for moment conditions created by a given linear transformation. The main equivalence result states that under standard regularity conditions, transformations which create conditional moment restrictions in a given empirical setting need only to have an equal rank to reach the same efficiency bound.

Example applications are considered, including nonlinear models with multiplicative heterogeneity and linear models with arbitrary unobserved factor structures.

Chapter 3: Moment-based Estimation of Linear Panel Data Models with Factor-augmented Errors

I consider linear panel data models with unobserved factor structures when the number of time periods is small relative to the number of cross-sectional units. I examine two popular methods of estimation: the first eliminates the factors with a parameterized quasi-long-differencing (QLD) transformation. The other, referred to as common correlated effects (CCE), uses the cross-sectional averages of the independent and response variables to project out the space spanned by the factors. I show that the classical CCE assumptions imply unused moment conditions which can be exploited by the QLD transformation to derive new linear estimators which weaken identifying assumptions and have desirable theoretical properties. I prove asymptotic normality of the linear QLD estimators under a heterogeneous slope model which allows for a tradeoff between identifying conditions. These estimators do not require the number of cross-sectional variables to be less than $T - 1$, a strong restriction in fixed-T CCE analysis. Finally, I investigate the effects of per-student expenditure on standardized test performance using data from the state of Michigan.

ACKNOWLEDGEMENTS

To my dissertation committee: Jeff, you have given me more of your time than I ever deserved. Thank you for all of your patience and guidance. Thank you Peter for seeing potential in me and helping me along my academic journey. Despite your protest, I can't help but think of you as my co-chair. To Ben: I have benefited greatly from having such a brilliant applied researcher on my committee, someone who quickly digested my work and showed me how to apply it in relevant cases. Finally, I want to thank Nicky, and I have enjoyed working with you through the AFRE tutoring program.

To my fellow graduate students: thank you for your friendship and support throughout these past five years. My qualifying exam study group, Sean, Andrew, Joffré, Elise, and Alex, to whom I would not be here without. To Mehmet and Taeyoon, the oddball macro and financial economists. And to my econometric mentor Alyssa and Akanksha. Finally, I want to give a special thanks to Bhavna: despite living halfway across the world, you were always available to jump on the phone and support me, especially during the job market. I look forward to our future collaborations.

To my family: my emotional bedrock. To my mom Kathi and dad Curt, I can never repay you for your love and support throughout my entire life. You have nurtured me into the person I am today, and I am forever grateful. Also to my bonus parents Lorraine and Kevin, who have become an integral part of my family. To my brothers Jack, Nicky, Mark, and Kian, four of my closest friends and partners in crime. To Katie, whose presence brightens my home. To my confidant and future sister Dana: I would not have made it through U of I without your friendship. I am elated you have joined our family.

Finally, to my flock: Griffin and Stark, my feathered friends. You drive me insane, but I could not imagine life without you two. Last but not least, Danielle. You are my best friend. You give me the strength to go on. You are my solace and my inspiration. Everything I do I do for you. I love you more than life. If I were a poet, I could fully articulate how much you mean to me, but unfortunately I'm only an economist, so you'll just have to take my word.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

MORE EFFICIENT ESTIMATION OF MULTIPLICATIVE PANEL DATA MODELS IN THE PRESENCE OF SERIAL CORRELATION

1.1 Introduction

The fixed effects Poisson (FEP) estimator was originally developed by Hausman, Hall, and Griliches (1984) (hereafter, HHG) in their study of the effects of firm-level R&D spending on patent filings. HHG used the method of conditional maximum likelihood estimation (CMLE) to estimate the parameters in the conditional mean. In deriving the CMLE, HHG assumed that, conditional on the unobserved heterogeneity and the history of the covariates, the outcome variable is independent over time with a Poisson distribution. HHG showed that, conditional on the covariates and the sum of the counts over time, the joint distribution of the counts is multinomial and does not depend on the heterogeneity. Therefore, standard maximum likelihood theory applies, and the asymptotic theory assuming a fixed number of time periods is standard. Hahn (1997) verified that the FEP estimator achieves the semiparametric efficiency bound under the full distributional and conditional independence assumptions.

Wooldridge (1999) showed that the consistency of the FEP estimator only requires correct specification of the conditional mean function up to a multiplicative heterogeneity term. In particular, any kind of variance is allowed along with any kind of serial dependence. In fact, the outcome variable need not even be a count variable: it can be any nonnegative outcome, including a continuous or corner solution response. Thus, the FEP estimator is to multiplicative panel data models what the linear FE estimator is to linear models with additive heterogeneity.

When the conditional mean function is differentiable in the parameters – by far the leading case – Wooldridge (1999) established Fisher consistency of the FEP very generally. Specifically, Wooldridge showed that the score has a zero conditional mean (evaluated at the true parameter value) when the structural conditional mean is correctly specified. In addition to establishing robustness of the FEP estimator, the zero conditional mean property of the score leads to additional moment conditions that can be exploited in generalized method of moments (GMM) estimation to obtain estimators asymptotically more efficient than the FEP estimator. Unfortunately, the extra moment conditions proposed by Wooldridge (1999) are essentially ad hoc: they are not based on any notion of optimality. Consequently, the GMM approach to estimating multiplicative panel data models has not caught on: FEP estimation with the fully robust standard errors derived in Wooldridge (1999) is much more common. Some recent examples include McCabe and Snyder (2014, 2015), Schlenker and Walker (2016), Krapf, Ursprung, and Zimmermann (2017), Castillo, Mejia, and Restrepo (2018), and Williams, Burnap, Javed, Liu, and Ozalp (2020).

Given that the FEP estimator is fully robust to distributional misspecification and serial independence, it is natural to wonder about its asymptotic efficiency under assumptions weaker than the full set of assumptions used by Hahn (1997). Recently, Verdier (2018) showed that the Poisson distributional assumption and conditional independence are not necessary for the FEP estimator to achieve Chamberlain's (1987, 1992) efficiency bound. In particular, Verdier (2018) showed that it is sufficient to impose the Poisson assumption that the variance equals the mean and that the outcomes are serially uncorrelated conditional on heterogeneity and the covariates. While weaker than the HHG assumptions, they are still restrictive. The assumption that the variance equals the mean, even after conditioning on unobserved heterogeneity, is very special. For example, the most common parameterization of the gamma distribution violates equality of the variance and mean. Moreover, serial correlation in the idiosyncratic errors of linear unobserved effects models is pervasive (which is why researchers now routinely compute standard errors robust to general serial correlation), and it is known how to exploit serial correlation in fixed effects versions of generalized least squares (GLS) to improve efficiency over the usual fixed effects estimator – see, for example, Im, Ahn, Schmidt, and Wooldridge (1999). It seems natural to search for analogous improvements over the FEP estimator in the presence of serial correlation and more flexible variance-mean relationships.

In this paper, we relax the second moment assumptions that are implied by the traditional HHG assumptions and derive the optimal instruments, thereby showing how to obtain an estimator that achieves Chamberlain's (1992) lower bound. Our efficiency result is new, and includes the Verdier (2018) result as a special case. The variance assumption we use to derive the optimal instruments is appealing because, conditional on the observed covariates and unobserved heterogeneity, it allows for underdispersion (relative to the Poisson) or overdispersion. In the spirit of the popular generalized estimating equations (GEE) approach – see Liang and Zeger (1986) – we assume constant conditional correlations, but allow for any pattern of serial correlation. One important difference from the GEE literature is that our assumptions are more "structural" in that we state the second moment assumptions conditional on the unobserved heterogeneity. This is analogous to the linear model with an additive, unobserved effect when the working correlation matrix of the idiosyncratic errors is assumed to be constant but is otherwise unrestricted.

In order to obtain parametric forms for the optimal instruments, we supplement the flexible second moment assumptions for the response variable with moment assumptions about the multiplicative heterogeneity. These parametric assumptions are fairly flexible and are commonly used in the literature, particularly in traditional and correlated random effects environments when one needs to impose distributional assumptions on the heterogeneity in order to obtain consistent estimators. Here, we impose first and second moment assumptions in order to obtain the optimal instruments.

We must emphasize that the estimator based on the optimal instruments – which we refer to as the "generalized FEP (GFEP) estimator" – does not require any assumptions for consistency and asymptotic normality beyond those used by the FEP estimator. That our new estimator is just as robust as the FEP estimator in terms of consistency is important, as it is unfair to claim efficiency improvements if the new estimator is not as robust as the popular, robust FEP estimator. In order to emphasize the robustness of our estimator, we use the term "working" assumptions. The key is that, under these parametric "working" assumptions we obtain the optimal instruments. If the working assumptions are correct, then we have a just identified estimator that is more efficient than the FEP estimator.

If any of the working assumptions are incorrect, the "optimal" instrumental variables (IVs) are no longer optimal, and so the GFEP no longer achieves Chamberlain's lower bound. Therefore, we have two estimators that are consistent under the same assumptions but efficient under different working assumptions. To ensure that we have an estimator that is at least as efficient than both the FEP estimator and the GFEP estimator, and usually more efficient, we combine the two sets of moment conditions. With K parameters this gives K overidentifying restrictions. The overidentifying restrictions are useful for testing the conditional mean specification – not the working assumptions, as those are not being used for consistency.

To summarize, this paper has three primary contributions. First, we relax the second moment assumptions implied by the traditional fixed effects Poisson setting and obtain the optimal instruments under an appealing set of second moment working assumptions, including allowing for general patterns of serial correlation. Second, we operationalize the estimator by imposing additional working assumptions on moments of the heterogeneity distribution, resulting in a GMM estimator that is computationally simple and is guaranteed to be asymptotically more efficient than both the FEP estimator and the GFEP estimator. Third, we significantly relax the conditions under which the FEP estimator achieves the asymptotic variance lower bound, allowing for both underdispersion and overdispersion in the variance conditional on observed covariates and unobserved heterogeneity.

The underlying asymptotic theory in this paper is for the microeconometric setting that treats the number of time periods, T, as fixed, and lets the cross section dimension, N, increase without bound. We assume random sampling in the cross section dimension but impose no restrictions on the time series dependence. We do not provide formal regularity conditions because the asymptotic theory is standard, and follow as in hundreds of panel data papers that impose random sampling in the cross section. We do assume smoothness so that certain derivatives – in particular, that of the conditional mean function – exist and are continuous.

The rest of the paper is organized as follows. Section 1.2 presents the conditional mean model and summarizes the consistency result for the FEP estimator. Section 1.3 derives the optimal instruments under two working variance assumptions, including an unrestricted (but constant) conditional correlation matrix. Section 1.4 shows how to implement the GFEP estimator and the

GMM estimator that combines the two sets of moment conditions. Section 1.5 provides promising simulation evidence comparing the FEP, GFEP, and GMM estimators under serial correlation with both underdispersion and overdispersion in the structural variance. Section 1.6 contains concluding remarks.

1.2 Model and Background

We consider a balanced panel data setting where, for each i, $\{(y_{it}, \mathbf{x}_{it}, c_i) : t = 1, 2, ..., T\}$ is a random draw from the population. We observe the nonnegative response variable $y_{it} \geq 0$ and $\mathbf{x}_{it}$, a $1 \times K$ vector. The scalar c_i is the unobserved heterogeneity. As is usual in fixed effects environments, the elements of $\mathbf{x}_{it}$ must have variation across t for at least some population units. Typically, these would include dummy variables indicating different time periods to allow for flexible aggregate time effects. The entire observed history of the covariates is $\mathbf{x}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{iT})$. As mentioned in the introduction, we are treating T as fixed in the asymptotic analysis. Therefore, because we assume random sampling in the cross section, relevant assumptions can be stated for a random draw i from the population.

The substantive assumptions that we make throughout the paper are that the model of the conditional mean is correctly specified, the heterogeneity is multiplicative, and the covariates are strictly exogenous conditional on c_i. These are all captured by the following.

Assumption Conditional Mean (CM): For $t = 1, ..., T$ and some $\boldsymbol{\beta}_0 \in \mathbb{R}^P$,

$$\mathrm{E}\left(y_{it}|\mathbf{x}_i, c_i\right) = \mathrm{E}\left(y_{it}|\mathbf{x}_{it}, c_i\right) = c_i m_t\left(\mathbf{x}_{it}, \boldsymbol{\beta}_0\right) \tag{1.2.1}$$

where $m_t\left(\mathbf{x}_t, \cdot\right) \geq 0$ is continuously differentiable on $\mathbb{R}^P$ for all $\mathbf{x}_t \in \mathcal{X}_t$, the support of $\mathbf{x}_{it}$. ■

As discussed in Wooldridge (1999), for consistency of the FEP estimator one can get by with assuming continuity over the parameter space, but we impose assumptions that imply asymptotic normality and easy calculation of asymptotic efficiency bounds. See Newey and McFadden (1994) or Wooldridge (2010, Chapter 12) for formal regularity conditions. In terms of smoothness, assuming $m_t\left(\mathbf{x}_{it}, \cdot\right)$ is twice continuously differentiable is sufficient and is almost always true in practice.

By far the leading case of the conditional mean function is

$$\mathrm{E}\left(y_{it}|\mathbf{x}_{it}, c_i\right) = c_i \exp\left(\mathbf{x}_{it}\boldsymbol{\beta}_0\right) \tag{1.2.2}$$

where $\mathbf{x}_{it}$ can include time period dummies to allow different intercepts inside the exponential function. Naturally, $\mathbf{x}_{it}$ can also include nonlinear functions of underlying explanatory variables, including squares and interactions. Given the choice in (1.2.2), $P = K$, but we also allow more general mean functions. Because we want to allow arbitrary dependence between c_i and $\mathbf{x}_{it}$, we need time variation in the latter for at least some units in the population. This permits, for example, interactions among variables that have some time variation and others that do not.

Strict exogeneity conditional on the unobserved effect c_i is implied by the first equality in (1.2.1). This assumption is restrictive – for example, it rules out lagged dependent variables – but it is much less restrictive than the strict exogeneity assumption typically used in the GEE literature because of conditioning on c_i. In the typical GEE approach the strict exogeneity assumption is stated as $\mathrm{E}\left(y_{it}|\mathbf{x}_i\right) = \mathrm{E}\left(y_{it}|\mathbf{x}_{it}\right)$. [For a discussion of GEE from an econometrics perspective, see Wooldridge (2010, Section 13.11.4).] Using iterated expectations, if (1.2.1) holds then

$$\mathrm{E}\left(y_{it}|\mathbf{x}_i\right) = \mathrm{E}\left(c_i|\mathbf{x}_i\right) m_t\left(\mathbf{x}_{it}, \boldsymbol{\beta}_0\right)$$

and the latter expression is not $\mathrm{E}\left(y_{it}|\mathbf{x}_{it}\right)$ if $\mathrm{E}\left(c_i|\mathbf{x}_i\right) \neq \mathrm{E}\left(c_i\right)$.

The multiplicative formulation using the exponential function in (1.2.2) can be obtained from

$$\mathrm{E}\left(y_{it}|\mathbf{x}_{it}, a_i\right) = \exp\left(a_i + \mathbf{x}_{it}\boldsymbol{\beta}_0\right)$$

where $c_i \equiv \exp\left(a_i\right)$. In applications where $\mathrm{P}(y_{it} = 0) > 0$, it is important to use (1.2.2) to allow for the possibility that $c_i = 0$, which then implies $y_{it} = 0$, $t = 1, 2, ..., T$. Often in count data and corner solution applications one sees some units with $y_{it} = 0$ for all t. Remember, we are only assuming $y_{it} \geq 0$; no other restrictions are imposed on the support of y_{it}. In most cases, a model such as (1.2.2) is appealing when y_{it} has no natural upper bound.

In FEP estimation, the following residual function, first studied by HHG, plays an important role:

$$u_{it}(\boldsymbol{\beta}) \equiv y_{it} - n_i p_t(\mathbf{x}_i, \boldsymbol{\beta}) \tag{1.2.3}$$

where $n_i \equiv \sum_{r=1}^{T} y_{ir}$ and

$$p_t(\mathbf{x}_i, \boldsymbol{\beta}) \equiv \frac{m_t(\mathbf{x}_{it}, \boldsymbol{\beta})}{\sum_{r=1}^{T} m_r(\mathbf{x}_{ir}, \boldsymbol{\beta})} \tag{1.2.4}$$

As convenient shorthand, we write $m_{it}(\boldsymbol{\beta}) = m_t(\mathbf{x}_{it}, \boldsymbol{\beta})$ and $p_{it}(\boldsymbol{\beta}) = p_t(\mathbf{x}_i, \boldsymbol{\beta})$. We can stack the $p_{it}(\boldsymbol{\beta})$ into the $T \times 1$ vector $\mathbf{p}(\mathbf{x}_i, \boldsymbol{\beta})$ and write

$$\mathbf{u}_i(\boldsymbol{\beta}) = \mathbf{y}_i - \mathbf{p}(\mathbf{x}_i, \boldsymbol{\beta}) n_i = \mathbf{y}_i - \mathbf{p}(\mathbf{x}_i, \boldsymbol{\beta}) \mathbf{1}_T' \mathbf{y}_i = \left[\mathbf{I}_T - \mathbf{p}(\mathbf{x}_i, \boldsymbol{\beta}) \mathbf{1}_T'\right] \mathbf{y}_i \tag{1.2.5}$$

where $\mathbf{u}_i(\boldsymbol{\beta})$ is the $T \times 1$ vector with t^{th} element $u_{it}(\boldsymbol{\beta})$ and $\mathbf{1}_T$ is the $T \times 1$ vector with all elements unity. As shown in Wooldridge (1999) under Assumption CM.1,

$$\mathrm{E}\left[\mathbf{u}_i(\boldsymbol{\beta}_0) | \mathbf{x}_i\right] = \mathbf{0} \tag{1.2.6}$$

Further, the score of the quasi-log-likelihood function for random draw i can be written as

$$\mathbf{s}_i(\boldsymbol{\beta}) = \nabla_{\boldsymbol{\beta}} \mathbf{p}(\mathbf{x}_i, \boldsymbol{\beta})' \mathbf{W}(\mathbf{x}_i, \boldsymbol{\beta}) \mathbf{u}_i(\beta) \tag{1.2.7}$$

where

$$\mathbf{W}(\mathbf{x}_i, \boldsymbol{\beta}) = \mathrm{diag}\left\{[p_{i1}(\boldsymbol{\beta})]^{-1}, [p_{i2}(\boldsymbol{\beta})]^{-1}, ..., [p_{iT}(\boldsymbol{\beta})]^{-1}\right\} \tag{1.2.8}$$

is $T \times T$.

It follows immediately that

$$\mathrm{E}\left[\mathbf{s}_i\left(\boldsymbol{\beta}_0\right)|\mathbf{x}_i\right] = \mathbf{0} \tag{1.2.9}$$

and this translates, under standard regularity conditions, into the consistency and $\sqrt{N}$-asymptotic normality of the FEP estimator. For emphasis, only Assumption CM is needed for consistency and asymptotic normality, and fully robust inference using a sandwich estimator is essentially trivial.

Wooldridge (1999) also notes that the conditional moment restrictions in (1.2.6) leads to uncountably many unconditional moment restrictions beyond those used by the FEP estimator, which are given by

$$\mathrm{E}\left[\mathbf{s}_i\left(\boldsymbol{\beta}_0\right)\right] = \mathbf{0}.$$

Wooldridge (1999) suggests some extra moment conditions but makes no attempt to find the optimal estimator based on (1.2.6). In the next section we derive the optimal instruments under a set of second moment assumptions.

1.3 Optimal Instruments under Second Moment Assumptions

Given the moment conditions in (1.2.6), we can apply Chamberlain's (1992) semiparametric efficiency bound to obtain an asymptotically efficient estimator. Define

$$\mathbf{D}_o\left(\mathbf{x}_i\right) \equiv \mathrm{E}\left[\nabla_{\boldsymbol{\beta}}\mathbf{u}_i\left(\boldsymbol{\beta}_0\right)|\mathbf{x}_i\right] \tag{1.3.1}$$

and

$$\mathbf{V}_o\left(\mathbf{x}_i\right) \equiv \mathrm{Var}\left[\mathbf{u}_i\left(\boldsymbol{\beta}_0\right)|\mathbf{x}_i\right] \tag{1.3.2}$$

Under regularity conditions of the kind found in Newey and McFadden (1994), Newey (2001) extended Chamberlain (1992) by allowing $\mathbf{V}_o\left(\mathbf{x}_i\right)$ to be singular and showed that the efficient estimator that uses only (1.2.6) has asymptotic variance

$$\left\{\mathrm{E}\left[\mathbf{D}_o\left(\mathbf{x}_i\right)'\mathbf{V}_o\left(\mathbf{x}_i\right)^{-}\mathbf{D}_o\left(\mathbf{x}_i\right)\right]\right\}^{-1} \tag{1.3.3}$$

where $\mathbf{V}_o(\mathbf{x}_i)^-$ denotes any generalized inverse (g-inverse), which means $\mathbf{V}_o(\mathbf{x}_i)\mathbf{V}_o(\mathbf{x}_i)^-\mathbf{V}_o(\mathbf{x}_i) = \mathbf{V}_o(\mathbf{x}_i)$. Because $\mathbf{V}_o(\mathbf{x}_i)$ is symmetric, a symmetric g-inverse always exists, and it simplifies notation to take $\mathbf{V}_o(\mathbf{x}_i)^-$ to be symmetric. Below we will obtain an explicit formula for a symmetric g-inverse. Given a random sample of size N and knowledge of $\mathbf{D}_o(\mathbf{x}_i)$ and $\mathbf{V}_o(\mathbf{x}_i)$, an estimator $\hat{\beta}_{OPT}$ that achieves this lower bound solves the exactly identified moment equations

$$\sum_{i=1}^{N}\mathbf{D}_o(\mathbf{x}_i)'\mathbf{V}_o(\mathbf{x}_i)^-\mathbf{u}_i\left(\widehat{\boldsymbol{\beta}}_{OPT}\right) = \mathbf{0} \tag{1.3.4}$$

Of course, this estimator is infeasible because $\mathbf{D}_o(\mathbf{x}_i)$ and $\mathbf{V}_o(\mathbf{x}_i)$ are generally unknown. In principle, both can be nonparametrically estimated. However, especially given the often large dimension of $\mathbf{x}_i$, nonparametric estimation of many conditional means, variances, and covariances hardly seems worth it just to improve asymptotic efficiency over the FEP estimator. Plus, the finite-sample properties of the the resulting estimator could be poor. Our goal here is to obtain simple formulas for the optimal IVs $\mathbf{Z}^*(\mathbf{x}_i) \equiv \mathbf{V}_o(\mathbf{x}_i)^-\mathbf{D}_o(\mathbf{x}_i)$ under reasonably flexible parametric second moment assumptions that have antecedents in the count data literature.

To find $\mathbf{D}_o(\mathbf{x}_i)$, note that

$$\nabla_{\boldsymbol{\beta}}\mathbf{u}_i(\boldsymbol{\beta}) = -\nabla_{\boldsymbol{\beta}}\mathbf{p}(\mathbf{x}_i,\boldsymbol{\beta})\,n_i \tag{1.3.5}$$

where, for each t, we can write

$$\nabla_{\boldsymbol{\beta}}p_{it}(\boldsymbol{\beta}) = \left[\sum_{r=1}^{T}m_{ir}(\boldsymbol{\beta})\right]^{-1}\left\{\nabla_{\boldsymbol{\beta}}m_{it}(\boldsymbol{\beta}) - \left[\sum_{r=1}^{T}\nabla_{\boldsymbol{\beta}}m_{ir}(\boldsymbol{\beta})\right]p_{it}(\boldsymbol{\beta})\right\}$$

Therefore,

$$\begin{aligned}\nabla_{\boldsymbol{\beta}}\mathbf{p}_i(\boldsymbol{\beta}) &= \left[\sum_{r=1}^{T}m_{ir}(\boldsymbol{\beta})\right]^{-1}\left\{\nabla_{\boldsymbol{\beta}}\mathbf{m}_i(\boldsymbol{\beta}) - \mathbf{p}_i(\boldsymbol{\beta})\left[\mathbf{1}_T'\nabla_{\boldsymbol{\beta}}\mathbf{m}_i(\boldsymbol{\beta})\right]\right\}\\ &= \left[\sum_{r=1}^{T}m_{ir}(\boldsymbol{\beta})\right]^{-1}\left[\mathbf{I}_T - \mathbf{p}_i(\boldsymbol{\beta})\mathbf{1}_T'\right]\nabla_{\boldsymbol{\beta}}\mathbf{m}_i(\boldsymbol{\beta}) \qquad (1.3.6)\end{aligned}$$

which gives us the necessary gradient.

Further, because

$$\mathrm{E}\left(n_i|\mathbf{x}_i, c_i\right) = c_i\left[\sum_{r=1}^{T} m_{ir}\left(\boldsymbol{\beta}_0\right)\right]$$

we have

$$\mathrm{E}\left[\nabla_{\boldsymbol{\beta}}\mathbf{u}_i\left(\boldsymbol{\beta}_0\right)|\mathbf{x}_i, c_i\right] = -c_i\left[\mathbf{I}_T - \mathbf{p}_i\left(\boldsymbol{\beta}_0\right)\mathbf{1}_T'\right]\nabla_{\boldsymbol{\beta}}\mathbf{m}_i\left(\boldsymbol{\beta}_0\right)$$

Now, let

$$\mu_c\left(\mathbf{x}_i\right) \equiv \mathrm{E}\left(c_i|\mathbf{x}_i\right)$$

Then we have shown

$$\mathbf{D}_o\left(\mathbf{x}_i\right) = -\mu_c\left(\mathbf{x}_i\right)\left[\mathbf{I}_T - \mathbf{p}_i\left(\boldsymbol{\beta}_0\right)\mathbf{1}_T'\right]\nabla_{\boldsymbol{\beta}}\mathbf{m}_i\left(\boldsymbol{\beta}_0\right) \tag{1.3.7}$$

which is the first piece needed to derive the optimal instruments. The unknown function in $\mathbf{D}_o\left(\mathbf{x}_i\right)$, $\mu_c\left(\mathbf{x}_i\right)$, is the conditional mean in the heterogeneity distribution.

Next, consider $\mathbf{V}_o\left(\mathbf{x}_i\right)^{-}$. First, we can write

$$\begin{aligned}\mathbf{V}_o\left(\mathbf{x}_i\right) &\equiv \mathrm{Var}\left[\mathbf{u}_i\left(\boldsymbol{\beta}_0\right)|\mathbf{x}_i\right] = \mathrm{Var}\left\{\left[\mathbf{I}_T - \mathbf{p}_i\left(\boldsymbol{\beta}_0\right)\mathbf{1}_T'\right]\mathbf{y}_i|\mathbf{x}_i\right\} \\ &\equiv \left(\mathbf{I}_T - \mathbf{P}_i\right)\boldsymbol{\Omega}_i\left(\mathbf{I}_T - \mathbf{P}_i'\right) \qquad (1.3.8)\end{aligned}$$

where

$$\boldsymbol{\Omega}_i \equiv \mathrm{Var}\left(\mathbf{y}_i|\mathbf{x}_i\right) \tag{1.3.9}$$

is assumed to be nonsingular (with probability one) and $\mathbf{P}_i \equiv \mathbf{p}_i\left(\boldsymbol{\beta}_0\right)\mathbf{1}_T'$ is $T \times T$. Because the $p_{it}\left(\boldsymbol{\beta}_0\right)$ sum to unity across t, it is easy to show that $\mathbf{P}_i$ is an idempotent (but not symmetric) matrix with $\mathrm{rank}(\mathbf{P}_i) = 1$.

In establishing that the FEP estimator is asymptotically efficient under the Poisson first and second moment assumptions, Verdier (2018) finds a particular symmetric matrix which is inherent to the FEP solution.

The matrix

$$\begin{aligned}\mathbf{V}_o(\mathbf{x}_i)^- &= \boldsymbol{\Omega}_i^{-1} - \boldsymbol{\Omega}_i^{-1}\mathbf{p}_i(\boldsymbol{\beta}_0)\left[\mathbf{p}_i(\boldsymbol{\beta}_0)'\boldsymbol{\Omega}_i^{-1}\mathbf{p}_i(\boldsymbol{\beta}_0)\right]^{-1}\mathbf{p}_i(\boldsymbol{\beta}_0)'\boldsymbol{\Omega}_i^{-1} \\ &= \boldsymbol{\Omega}_i^{-1} - \boldsymbol{\Omega}_i^{-1}\mathbf{m}_i(\boldsymbol{\beta}_0)\left[\mathbf{m}_i(\boldsymbol{\beta}_0)'\boldsymbol{\Omega}_i^{-1}\mathbf{m}_i(\boldsymbol{\beta}_0)\right]^{-1}\mathbf{m}_i(\boldsymbol{\beta}_0)'\boldsymbol{\Omega}_i^{-1} \qquad (1.3.10)\end{aligned}$$

is a generalized inverse of $\mathbf{V}_o(\mathbf{x}_i)$. The second equality in (1.3.10) follows by the definition of $\mathbf{p}_i(\boldsymbol{\beta}_0)$ and by cancelling terms. By simple multiplication it is easily seen that

$$\mathbf{p}_i(\boldsymbol{\beta}_0)'\mathbf{V}_o(\mathbf{x}_i)^- = \mathbf{0}$$

and so

$$\mathbf{D}_o(\mathbf{x}_i)'\mathbf{V}_o(\mathbf{x}_i)^- = -\mu_c(\mathbf{x}_i)\nabla_{\boldsymbol{\beta}}\mathbf{m}_i(\boldsymbol{\beta}_0)'\mathbf{V}_o(\mathbf{x}_i)^- \qquad (1.3.11)$$

The expression for the optimal instruments in (1.3.11) is not directly applicable because $\mu_c(\cdot)$ and $\mathbf{V}_o(\cdot)$ are unknown, with the latter depending on the unknown $\boldsymbol{\Omega}_i$. We now impose assumptions on the structural variance-covariance matrix, $\mathrm{Var}(\mathbf{y}_i|\mathbf{x}_i, c_i)$, that lead to useful simplifications. The first restriction is on the diagonal elements.

Assumption Working Variance 1 (WV.1): For $t = 1, ..., T$, there exists $\alpha > 0$ such that

$$\mathrm{Var}(y_{it}|\mathbf{x}_i, c_i) = \mathrm{Var}(y_{it}|\mathbf{x}_{it}, c_i) = \alpha\mathrm{E}(y_{it}|\mathbf{x}_{it}, c_i) = \alpha c_i m_{it}(\boldsymbol{\beta}_0) \blacksquare \qquad (1.3.12)$$

Assumption WV.1 is motivated by the count data literature, where the assumption that the variance is proportional to the mean is commonly used in generalized linear models (GLM) and GEE settings; see, for example, McCullagh and Nelder (1989), Liang and Zeger (1986), Hardin and Hilbe (2012), and Wooldridge (2010, Section 13.11). Again, one important difference between our setting and the standard GEE setting is that we state the first and second moments conditional on the unobserved heterogeneity, c_i, in addition to the observable variables, $\mathbf{x}_i$. Once the population is effectively partitioned on the basis of $(\mathbf{x}_i, c_i)$, the so-called "GLM variance assumption" is more appealing. We do not restrict the value of $\alpha = \mathrm{Var}(y_{it}|\mathbf{x}_{it}, c_i)/\mathrm{E}(y_{it}|\mathbf{x}_{it}, c_i)$, and so the y_{it}

can exhibit underdispersion or overdispersion relative to the Poisson distribution. This variance-mean relationship also holds for one popular parameterization of the negative binomial distribution (which implies overdispersion), and can hold for continuous outcomes as well, such as a common parameterization of the gamma distribution.

The second working assumption is on the conditional correlation matrix.

Assumption Working Variance 2 (WV.2): For a $T \times T$ symmetric, positive definite matrix $\mathbf{R}$ (with unity down the diagonal),

$$\text{Corr}\left(\mathbf{y}_i|\mathbf{x}_i, c_i\right) = \mathbf{R} \blacksquare \tag{1.3.13}$$

Assumption WV.2 is motivated by the GEE literature, where a constant conditional correlation matrix is the leading example of a working correlation assumption. We do not put restrictions on the elements of $\mathbf{R}$, $\rho_{ts} = \text{Corr}\left(y_{it}, y_{is}|\mathbf{x}_i, c_i\right)$, other than those that ensure $\mathbf{R}$ is a valid correlation matrix. The special case of no serial correlation conditional on $(\mathbf{x}_i, c_i)$ is $\mathbf{R} = \mathbf{I}_T$. One could impose an exchangeability restriction on $\mathbf{R}$, as is common in the GEE literature, but that is less attractive here because we are conditioning on c_i (which would often be assumed to be an explanation for an exchangeable structure without conditioning on c_i). With large N and small T, there is little reason to impose restrictions on $\mathbf{R}$. Again, an important difference with the GEE literature is we condition the correlation matrix on c_i as well as $\mathbf{x}_i$ – which makes $\mathbf{R} = \mathbf{I}_T$ more tenable (but still unnecessary).

We can combine Assumptions WV.1 and WV.2 into a working variance-covariance matrix conditional on $(\mathbf{x}_i, c_i)$:

$$\text{Var}\left(\mathbf{y}_i|\mathbf{x}_i, c_i\right) = \alpha c_i \mathbf{M}_i^{1/2} \mathbf{R} \mathbf{M}_i^{1/2} \tag{1.3.14}$$

where $\mathbf{M}_i \equiv \text{diag}\left\{m_{i1}\left(\boldsymbol{\beta}_0\right), m_{i2}\left(\boldsymbol{\beta}_0\right), ..., m_{iT}\left(\boldsymbol{\beta}_0\right)\right\}$ and $\mathbf{M}_i^{1/2}$ is the obvious matrix square root. If not for conditioning on the unobserved heterogeneity c_i, (1.3.14) has a structure very familiar from the GEE literature on estimating conditional means of count variables with longitudinal data.

In stating Assumptions WV.1 and WV.2, we have opted not to include a “0” subscript on α or $\mathbf{R}$. This decision requires a brief explanation. For deriving the optimal instruments, we are assuming the existence of “true values.” However, when we discuss implementation of our new estimator in Section 1.4, we do not assume Assumptions WV.1 or WV.2 are in force. To ensure that the focus is on estimating $\boldsymbol{\beta}_0$, and to simplify the notation, we omit the “0” subscripts on the parameters in the working assumptions.

Before deriving the optimal instruments, we first obtain $\boldsymbol{\Omega}_i = \text{Var}\left(\mathbf{y}_i|\mathbf{x}_i\right)$ and provide a useful expression for its inverse. As shorthand, let $\mathbf{m}_i$ be the $T \times 1$ vector of $m_{it}\left(\boldsymbol{\beta}_0\right)$, and define $\mathbf{M}_i^{1/2}$ as above. We use $\sqrt{\mathbf{m}_i}$ to denote the $T \times 1$ vector containing the square roots of the $m_{it}\left(\boldsymbol{\beta}_0\right)$. In stating the next lemma, let

$$\sigma_c^2\left(\mathbf{x}_i\right) = \text{Var}\left(c_i|\mathbf{x}_i\right)$$

Lemma 1.3.1. *Under Assumptions CM, WV.1, and WV.2,*

$$\text{Var}\left(\mathbf{y}_i|\mathbf{x}_i\right) = \boldsymbol{\Omega}_i = \alpha\mu_c\left(\mathbf{x}_i\right)\mathbf{M}_i^{1/2}\mathbf{R}\mathbf{M}_i^{1/2} + \sigma_c^2\left(\mathbf{x}_i\right)\mathbf{m}_i\mathbf{m}_i' \tag{1.3.15}$$

which is positive definite. Further,

$$\boldsymbol{\Omega}_i^{-1} = \frac{1}{\left[\alpha\mu_c\left(\mathbf{x}_i\right)\right]}\mathbf{M}_i^{-1/2}\left\{\mathbf{R}^{-1} - \frac{\sigma_c^2\left(\mathbf{x}_i\right)}{\left[\alpha\mu_c\left(\mathbf{x}_i\right) + \sigma_c^2\left(\mathbf{x}_i\right)\sqrt{\mathbf{m}_i}'\mathbf{R}^{-1}\sqrt{\mathbf{m}_i}\right]}\mathbf{R}^{-1}\sqrt{\mathbf{m}_i}\sqrt{\mathbf{m}_i}'\mathbf{R}^{-1}\right\}\mathbf{M}_i^{-1/2}$$

Proof. See Appendix for proof. □

Establishing the formula for $\boldsymbol{\Omega}_i$ uses the law of total variance (for matrices). Positive definiteness of $\boldsymbol{\Omega}_i$ follows because the first term in (1.3.15) is positive definite under WV.1 and WV.2 and the second is always positive semi-definite. As shown in the Appendix, the formula for $\boldsymbol{\Omega}_i^{-1}$ applies a result due to Sherman and Morrison (1950).

Now we can state the main optimal instrument result.

Theorem 1.3.1. *Under Assumptions CM, WV.1, and WV.2, a symmetric generalized inverse of* $\mathbf{V}_o(\mathbf{x}_i)$ *is*

$$\mathbf{V}_o(\mathbf{x}_i)^- = \frac{1}{[\alpha\mu_c(\mathbf{x}_i)]}\mathbf{M}_i^{-1/2}\left[\mathbf{R}^{-1} - \frac{1}{\sqrt{\mathbf{m}_i}'\mathbf{R}^{-1}\sqrt{\mathbf{m}_i}}\mathbf{R}^{-1}\sqrt{\mathbf{m}_i}\sqrt{\mathbf{m}_i}'\mathbf{R}^{-1}\right]\mathbf{M}_i^{-1/2} \qquad (1.3.16)$$

Further, the optimal $T \times K$ *matrix of instruments,* $\mathbf{Z}^*(\mathbf{x}_i)$, *is*

$$\mathbf{Z}^*(\mathbf{x}_i)' \equiv \nabla_{\boldsymbol{\beta}}\mathbf{m}_i(\boldsymbol{\beta}_0)'\,\mathbf{M}_i^{-1/2}\left[\mathbf{R}^{-1} - \frac{1}{\sqrt{\mathbf{m}_i}'\mathbf{R}^{-1}\sqrt{\mathbf{m}_i}}\mathbf{R}^{-1}\sqrt{\mathbf{m}_i}\sqrt{\mathbf{m}_i}'\mathbf{R}^{-1}\right]\mathbf{M}_i^{-1/2} \qquad (1.3.17)$$

where, again, $\mathbf{m}_i$ *and* $\mathbf{M}_i$ *are evaluated at* $\boldsymbol{\beta}_0$. *We have dropped the minus sign in* $\mathbf{D}_o(\mathbf{x}_i)$ *as that does not affect the optimal choice.*

Proof. See Appendix for proof. □

The optimal instrument matrix in (1.3.17) has a rather remarkable feature: it does not depend on the constant α nor on the conditional first two moments of the heterogeneity distribution, $\mu_c(\mathbf{x}_i)$ and $\sigma_c(\mathbf{x}_i)$ – even though $\boldsymbol{\Omega}_i^{-1}$ depends on all of these quantities and $\mathbf{D}_o(\mathbf{x}_i)$ depends on $\mu_c(\mathbf{x}_i)$. Under the working variance matrix assumptions, the optimal instruments depend only on $\boldsymbol{\beta}_0$ and $\mathbf{R}$. We have a natural preliminary estimator of $\boldsymbol{\beta}_0$, namely, the FEP estimator. Estimating $\mathbf{R}$ is much more challenging, and for that we will introduce additional working assumptions – something we take up in the next section.

An interesting special case of Theorem 1.3.1 is when the $\{y_{it} : t = 1, 2, ..., T\}$ are conditionally uncorrelated, an assumption with a long history in linear and nonlinear unobserved effects models. Traditional treatments of linear unobserved effects models – often called "random effects" models – include the assumption that idiosyncratic shocks are serially uncorrelated, which implies that, conditional on $(\mathbf{x}_i, c_i)$, the $\{y_{it} : t = 1, 2, ..., T\}$ are uncorrelated. In using joint maximum likelihood to estimate nonlinear models with unobserved heterogeneity – random effects probit and ordered probit, random effects multinomial logit, random effects Tobit, random effects version of Poisson and negative binomial models, among others – it is almost always assumed that the

$\{y_{it} : t = 1, 2, ..., T\}$ are independent conditional on $(\mathbf{x}_i, c_i)$; see Sections 13.9, 15.8, 17.8, and 18.7 in Wooldridge (2010).

Corollary 1.3.1. *Under Assumptions CM, WV.1, and WV.2 with* $\mathbf{R} = \mathbf{I}_T$*, the FEP estimator is efficient among estimators that use only Assumption CM for consistency.*

Proof. See Appendix for proof. □

Corollary 1.3.1 is a new result that shows the FEP estimator is asymptotically efficient for any $\alpha > 0$ in Assumption WV.1 provided there is no serial correlation. Conditional on $\mathbf{x}_i$ and c_i, any amount of constant underdispersion or overdispersion is allowed. Therefore, Corollary 1.3.1 improves on Verdier (2018), who imposed $\alpha = 1$, the value that holds for the Poisson distribution. That FEP is asymptotically efficient for any α while allowing for any dependence between c_i and $\mathbf{x}_i$ allows us to make an interesting connection with the cross-sectional GLM literature. As pointed out in Wooldridge (2010, Section 13.11.3), the cross-sectional version of Assumption WV.1 implies that the Poisson QMLE is asymptotically efficient among estimators that use only correct specification of the conditional mean function for consistency.

1.4 Operationalizing Optimal IV Estimation

From Theorem 1.3.1, in order to obtain a feasible optimal IV estimator under Assumptions CM, WV.1, and WV.2, we need a preliminary consistent estimator of $\boldsymbol{\beta}_0$ and we either need to know $\mathbf{R}$ or have a consistent estimator of it. If we want to impose a specific structure on $\mathbf{R}$ – say, an AR(1) model with a known AR(1) parameter – then (1.3.17) can be used after replacing $\boldsymbol{\beta}_0$ with $\widehat{\boldsymbol{\beta}}_{FEP}$ (the clear choice for a first-stage estimator of $\boldsymbol{\beta}_0$). Remember, imposing such a restriction when it is incorrect would not affect consistency of the method of moments estimator; but the estimator would not be asymptotically efficient. Generally, we want to estimate $\mathbf{R}$ without imposing any restrictions.

In order to ignore the first-stage estimation when obtaining the asymptotic variance of $\sqrt{N}\left(\hat{\beta}_{OPT} - \boldsymbol{\beta}_0\right)$, the first-stage estimators of should be $\sqrt{N}$-consistent – a weak requirement because we are assuming random sampling and smooth moment and objective functions. See Wooldridge (2010, Chapter 14) for discussion. As mentioned earlier, it is very natural to use the FEP estimator as the initial estimator of $\boldsymbol{\beta}_0$. Estimation of $\mathbf{R}$ is more difficult because it is the (working) correlation matrix conditional on the unobserved heterogeneity, c_i, in addition to $\mathbf{x}_i$.

The key to estimating $\mathbf{R}$ is the relationship in (1.3.15). To see how (1.3.15) can be used, define a $T \times 1$ vector of errors

$$\mathbf{v}_i \equiv \mathbf{y}_i - \mathrm{E}\left(\mathbf{y}_i|\mathbf{x}_i\right) = \mathbf{y}_i - \mu_c\left(\mathbf{x}_i\right)\mathbf{m}_i \tag{1.4.1}$$

Then

$$\mathrm{E}\left(\mathbf{v}_i\mathbf{v}_i'|\mathbf{x}_i\right) = \alpha\mu_c\left(\mathbf{x}_i\right)\mathbf{M}_i^{1/2}\mathbf{R}\mathbf{M}_i^{1/2} + \sigma_c^2\left(\mathbf{x}_i\right)\mathbf{m}_i\mathbf{m}_i' \tag{1.4.2}$$

which we can write in matrix error form as

$$\mathbf{v}_i\mathbf{v}_i' = \alpha\mu_c\left(\mathbf{x}_i\right)\mathbf{M}_i^{1/2}\mathbf{R}\mathbf{M}_i^{1/2} + \sigma_c^2\left(\mathbf{x}_i\right)\mathbf{m}_i\mathbf{m}_i' + \mathbf{S}_i$$

with

$$\mathrm{E}\left(\mathbf{S}_i|\mathbf{x}_i\right) = \mathbf{0} \tag{1.4.3}$$

Next, define

$$\mathbf{k}_i \equiv \mathrm{E}\left(\mathbf{y}_i|\mathbf{x}_i\right) = \mu_c\left(\mathbf{x}_i\right)\mathbf{m}_i \tag{1.4.4}$$

and let $\mathbf{K}_i$ be the diagonalized version of $\mathbf{k}_i$. Then

$$\mathbf{v}_i\mathbf{v}_i' - \sigma_c^2\left(\mathbf{x}_i\right)\mathbf{m}_i\mathbf{m}_i' = \alpha\sqrt{\mathbf{K}_i}\mathbf{R}\sqrt{\mathbf{K}_i} + \mathbf{S}_i \tag{1.4.5}$$

and so

$$\mathbf{K}_i^{-1/2}\left[\mathbf{v}_i\mathbf{v}_i' - \sigma_c^2\left(\mathbf{x}_i\right)\mathbf{m}_i\mathbf{m}_i'\right]\mathbf{K}_i^{-1/2}/\alpha = \mathbf{R} + \mathbf{K}_i^{-1/2}\mathbf{S}_i\mathbf{K}_i^{-1/2}/\alpha \tag{1.4.6}$$

By (1.4.3) and iterated expectations, the second term in (1.4.6), $\mathbf{K}_i^{-1/2}\mathbf{S}_i\mathbf{K}_i^{-1/2}/\alpha$, has a mean of zero.

Therefore, we have shown

$$\mathbf{R} = \mathrm{E}\left\{\alpha^{-1}\mathbf{K}_i^{-1/2}\left[\mathbf{v}_i\mathbf{v}_i' - \sigma_c^2\left(\mathbf{x}_i\right)\mathbf{m}_i\mathbf{m}_i'\right]\mathbf{K}_i^{-1/2}\right\} \tag{1.4.7}$$

Combining (1.4.7) with (1.3.17) shows that α appears as a multiplicative factor in $\mathbf{Z}^*\left(\mathbf{x}_i\right)$, and therefore does not affect the optimal choice of instruments.

Equation (1.4.7) for $\mathbf{R}$ suggests simply computing the sample analog of the matrix inside the expected value. However, we must deal with the fact that the matrix depends on three unknown quantities: the parameter α, the conditional mean function $\mu_c\left(\cdot\right)$ (which appears in the definition of $\mathbf{v}_i$), and the conditional variance function $\sigma_c^2\left(\cdot\right)$.

There are different ways to approach estimation of $\mu_c\left(\cdot\right)$. For example, under Assumption CM,

$$\mathrm{E}\left(n_i|\mathbf{x}_i, c_i\right) = c_i\left[\sum_{r=1}^{T} m_{ir}\left(\boldsymbol{\beta}_0\right)\right] \tag{1.4.8}$$

and so

$$\mathrm{E}\left[\left.\frac{n_i}{\sum_{r=1}^{T} m_{ir}\left(\boldsymbol{\beta}_0\right)}\right|\mathbf{x}_i\right] = \mu_c\left(\mathbf{x}_i\right) \tag{1.4.9}$$

Alternatively, we can write

$$\mathrm{E}\left[\left.T^{-1}\sum_{t=1}^{T}\frac{y_{it}}{m_{it}\left(\boldsymbol{\beta}_0\right)}\right|\mathbf{x}_i\right] = \mu_c\left(\mathbf{x}_i\right) \tag{1.4.10}$$

Because we have available $\sqrt{N}$-consistent estimators of $\boldsymbol{\beta}_0$, expressions (1.4.9) and (1.4.10) show that $\mu_c\left(\cdot\right)$ is nonparametrically identified. In fact, we can use these expressions to motivate a nonparametric estimator. Almost certainly the initial estimator of $\boldsymbol{\beta}_0$ is $\widehat{\boldsymbol{\beta}}_{FEP}$, in which case we construct a dependent variable, $n_i/\left[\sum_{r=1}^{T}\hat{m}_{ir}\right]$, where $\hat{m}_{ir} = m_{ir}\left(\widehat{\boldsymbol{\beta}}_{FEP}\right)$, and use it in a cross-sectional nonparametric regression to obtain $\hat{\mu}_c\left(\cdot\right)$. For $\sigma_c^2\left(\cdot\right)$, the law of total variance gives the conditional form given $\boldsymbol{x}_i$.

We have

$$\begin{aligned} \mathrm{E}\left(v_{it}^{2}|\mathbf{x}_{i}\right) &= \mathrm{Var}\left(y_{it}|\mathbf{x}_{i}\right) = \mathrm{E}\left[\mathrm{Var}\left(y_{it}|\mathbf{x}_{i}, c_{i}\right)|\mathbf{x}_{i}\right] + \mathrm{Var}\left[\mathrm{E}\left(y_{it}|\mathbf{x}_{i}, c_{i}\right)|\mathbf{x}_{i}\right] \\ &= \mathrm{E}\left[\alpha c_{i} m_{it}\left(\boldsymbol{\beta}_{0}\right)|\mathbf{x}_{i}\right] + \mathrm{Var}\left[c_{i} m_{it}\left(\boldsymbol{\beta}_{0}\right)|\mathbf{x}_{i}\right] \\ &= \alpha \mu_{c}\left(\mathbf{x}_{i}\right) m_{it}\left(\boldsymbol{\beta}_{0}\right) + \sigma_{c}^{2}\left(\mathbf{x}_{i}\right)\left[m_{it}\left(\boldsymbol{\beta}_{0}\right)\right]^{2} \qquad (1.4.11) \end{aligned}$$

where we impose the working variance Assumption WV.1. Given that $\mu_c(\mathbf{x}_i)$ is identified from the previous argument, this expression identifies α and $\sigma_c^2(\cdot)$. In fact, after obtaining (semiparametric) residuals $\hat{v}_{it} = y_{it} - \hat{\mu}_c(\mathbf{x}_i) m_{it}\left(\widehat{\boldsymbol{\beta}}_{FEP}\right)$, we can use the squared residuals, $\hat{v}_{it}^2$, as the dependent variable in nonparametric estimation of $\sigma_c^2(\cdot)$. Therefore, a semiparametric approach to estimating the optimal IVs is available under Assumptions CM, WV.1, and WV.2.

For practical reasons, our suggestion is to avoid estimating either $\mu_c(\cdot)$ and $\sigma^2(\cdot)$ nonparametrically. Remember, we only need to estimate these conditional moments to obtain IVs more efficient than those used by the FEP estimator. The dimension of $\mathbf{x}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{iT})$ is often large. We can reduce the dimension by using a nonparametric Mundlak (1978) device, which would have $\mu_c(\cdot)$ and $\sigma^2(\cdot)$ depending only on time averages $\bar{\mathbf{x}}_i \equiv T^{-1}\sum_{r=1}^{T}\mathbf{x}_{ir}$. Nevertheless, estimating a conditional variance along with a conditional mean when K is even moderately large is still challenging, both theoretically and practically. It would involve choosing at least two tuning parameters. From a robustness perspective, we cannot improve over the FEP estimator because it is consistent under Assumption CM. High-dimensional nonparametric estimation seems unnecessary to improve over the usual FEP estimator in the presence of serial correlation and under- or overdispersion, especially if one factors in finite-sample considerations. Instead, we draw on the literature on models for nonnegative responses to suggest working assumptions for the conditional mean and variance of the heterogeneity – as summarized, for example, in Wooldridge (2010, Section 18.7.3).

For concreteness, and because it is by far the leading case, we now assume that $m_{it}(\boldsymbol{\beta}_0) = \exp(\mathbf{x}_{it}\boldsymbol{\beta}_0)$. Other forms of $m_{it}(\boldsymbol{\beta}_0)$ are easily handled, but the formulas and connections with other literatures is not as straightforward. In fact, we do not even need a generalized linear model form in our current setting, though such a mean function tends to lead to easier interpretation.

Assumption WH.1: For known $1 \times Q$ functions $\mathbf{h}(\mathbf{x}_i)$, a scalar η, and λ a $Q \times 1$ vector,

$$\mu_c(\mathbf{x}_i) \equiv \mathrm{E}(c_i|\mathbf{x}_i) = \exp[\eta + \mathbf{h}(\mathbf{x}_i)\lambda] \ \blacksquare \tag{1.4.12}$$

The leading case is to use the (nonredundant) time averages of $\{\mathbf{x}_{it} : t = 1, ..., T\}$, which is an extension of the Mundlak (1978) device to the nonlinear case, so that $\mathbf{h}(\mathbf{x}_i) = \bar{\mathbf{x}}_i$. But we can also use Chamberlain's (1980) less restrictive version, or include other functions of $\{\mathbf{x}_{it} : t = 1, ..., T\}$, such as unit-specific trends or even unit-specific second moments. It seems sensible to use something simple, such as the Mundlak device, as we are only using WH.1 to generate instruments.

When we combine Assumption WH.1 with the exponential conditional mean for $\mathrm{E}(y_{it}|\mathbf{x}_i, c_i)$, we obtain, by iterated expectations,

$$\mathrm{E}(y_{it}|\mathbf{x}_i) = \exp[\eta + \mathbf{h}(\mathbf{x}_i)\lambda]\exp(\mathbf{x}_{it}\boldsymbol{\beta}_0) = \exp[\mathbf{x}_{it}\boldsymbol{\beta}_0 + \eta + \mathbf{h}(\mathbf{x}_i)\lambda] \tag{1.4.13}$$

The parameters in this conditional mean function can be consistently estimated using a variety of methods. A simple approach is to exploit equation (1.4.9) or (1.4.10) using exponential mean functions. After obtaining the FEP estimator $\widehat{\boldsymbol{\beta}}_{FEP}$, estimate η and λ by a cross sectional Poisson regression with mean function $\exp[\eta + \mathbf{h}(\mathbf{x}_i)\lambda]$ and one of the dependent variables

$$\frac{n_i}{\sum_{r=1}^{T}\exp\left(\mathbf{x}_{ir}\widehat{\boldsymbol{\beta}}_{FEP}\right)} \text{ or } T^{-1}\sum_{t=1}^{T}\frac{y_{it}}{\exp\left(\mathbf{x}_{it}\widehat{\boldsymbol{\beta}}_{FEP}\right)} \tag{1.4.14}$$

Even if the original y_{it} are count variables – and there is no presumption that they are – neither of the regressands in (1.4.14) would be a count variable. Of course, this is of no consequence because of the robustness of the Poisson QMLE for estimating the parameters of the conditional mean regardless of the nature of the dependent variable (provided it is nonnegative).

Alternatively, $\boldsymbol{\beta}_0$, η, and λ can be estimated jointly using the pooled Poisson QMLE. The pooled Poisson QMLE is completely robust to distributional misspecification and serial correlation. Of course, to preserve consistency of the resulting method of moments estimator we do not need Assumption WH.1 to hold; we are using it to estimate the optimal instruments derived earlier.

The second working assumption on the heterogeneity distribution imposes a restriction on the variance-mean relationship.

Assumption WH.2: For $\delta > 0$,

$$\sigma_c^2(\mathbf{x}_i) \equiv \mathrm{Var}(c_i|\mathbf{x}_i) = \delta\left[\mu_c(\mathbf{x}_i)\right]^2 = \delta\left\{\exp\left[\eta + \mathbf{h}(\mathbf{x}_i)\,\lambda\right]\right\}^2 \blacksquare \qquad (1.4.15)$$

Assumption WH.2 is very common in settings with nonnegative, continuous heterogeneity (including so-called random effects Poisson and negative binomial models). The condition that the variance is proportional to the square of the mean holds for the natural parameterizations of the gamma and lognormal distributions, and holds whenever

$$c_i = h_i \mu_c(\mathbf{x}_i) \qquad (1.4.16)$$

for $h_i \geq 0$ and independent of $\mathbf{x}_i$, without any further restrictions on the distribution of h_i. Like Assumption WH.1, Assumption WH.2 is not needed for consistent estimation using the method of moments estimator but only to estimate the optimal instruments under the working Assumptions WV.1 and WV.2.

Using Assumptions CM, WV.1, WH.1, and WH.2 we can obtain estimating equations for α and δ. First, note that

$$\mathrm{E}\left(v_{it}^2|\mathbf{x}_i\right) = \alpha k_{it} + \delta k_{it}^2 \qquad (1.4.17)$$

where

$$k_{it} \equiv \mathrm{E}(y_{it}|\mathbf{x}_i) = \exp\left[\mathbf{x}_{it}\boldsymbol{\beta}_0 + \eta + \mathbf{h}(\mathbf{x}_i)\,\lambda\right]$$

An immediate implication of equation (1.4.17) is

$$\mathrm{E}\left[\left(\frac{v_{it}}{\sqrt{k_{it}}}\right)^2 \middle| \mathbf{x}_i\right] = \alpha + \delta k_{it} \qquad (1.4.18)$$

which is the basis for estimating variance parameters in common cross-sectional models where heterogeneity is assumed independent of the covariates. A simple way to operationalize the conditional mean is

$$\hat{v}_{it} = y_{it} - \hat{k}_{it} = y_{it} - \exp\left[\mathbf{x}_{it}\widehat{\boldsymbol{\beta}}_{FEP} + \hat{\eta} + \mathbf{h}\left(\mathbf{x}_i\right)\widehat{\boldsymbol{\lambda}}\right] \tag{1.4.19}$$

where $\hat{\eta}$ and $\widehat{\boldsymbol{\lambda}}$ are from one of the Poisson regressions described in equation (1.4.13). Then $\hat{\alpha}$ and $\hat{\delta}$ are, respectively, the intercept and slope in the pooled simple regression

$$\frac{\hat{v}_{it}^2}{\hat{k}_{it}} \text{ on } 1, \hat{k}_{it}, t = 1, ..., T;\ i = 1, ..., N \tag{1.4.20}$$

It is clear from equation (1.3.17) that $\hat{\alpha}$ does not appear in the optimal instruments, but we need to estimate α in order to obtain $\hat{\delta}$. In order to conclude the working assumptions are a reasonable approximation to reality, both $\hat{\alpha}$ and $\hat{\delta}$ should be nonnegative. If one of them is negative (most likely $\hat{\delta}$) then $\hat{\delta}$ should be set to zero. Because $\hat{\alpha}$ drops out of the optimal IVs, we need not estimate it when we set $\hat{\delta} = 0$. Nevertheless, one may be curious about the estimated amount of overdispersion when δ is set to zero. With $\delta = 0$, the estimate of α is simply

$$\hat{\alpha} = (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\hat{v}_{it}^2/\hat{k}_{it}\right) \tag{1.4.21}$$

and this is guaranteed to be nonnegative. However, as mentioned above, $\hat{\alpha}$ does not affect estimation of the optimal IVs when $\delta = 0$.

When we add Assumptions WH.1 and WH.2 to the previous assumptions, we obtain a simple form for $\mathbf{R}$:

$$\mathbf{R} = \mathrm{E}\left\{\mathbf{K}_i^{-1/2}\left[\mathbf{v}_i\mathbf{v}_i' - \delta\mathbf{k}_i\mathbf{k}_i'\right]\mathbf{K}_i^{-1/2}/\alpha\right\}$$

which leads immediately to the method-of-moments/plug-in estimator

$$\hat{\mathbf{R}} = \left(\frac{1}{\hat{\alpha}}\right)N^{-1}\sum_{i=1}^{N}\hat{\mathbf{K}}_i^{-1/2}\left(\hat{\mathbf{v}}_i\hat{\mathbf{v}}_i' - \hat{\delta}\hat{\mathbf{k}}_i\hat{\mathbf{k}}_i'\right)\hat{\mathbf{K}}_i^{-1/2} \tag{1.4.22}$$

By a standard application of the uniform weak law of large numbers [Wooldridge (2010, Lemma 12.1)], $\hat{\mathbf{R}} \xrightarrow{p} \mathbf{R}$. For each $t \neq s$, the correlations are estimated as

$$\hat{\rho}_{st} = \left(\frac{1}{\hat{\alpha}}\right) N^{-1} \sum_{i=1}^{N} \frac{\left(\hat{v}_{is}\hat{v}_{it} - \hat{\delta}\hat{k}_{is}\hat{k}_{it}\right)}{\sqrt{\hat{k}_{is}\hat{k}_{it}}} \tag{1.4.23}$$

From the definition of $\hat{\alpha}$ and $\hat{\delta}$ obtained from (1.4.18), it is easily seen that $\hat{\rho}_{tt} = 1$ for $t = 1, ..., T$, and so this estimator imposes the logical requirement that a correlation matrix must have unity down its diagonal.

If we set $\delta = 0$, $\hat{\mathbf{R}}$ reduces to

$$\hat{\mathbf{R}} = \left(\frac{1}{\hat{\alpha}}\right) N^{-1} \sum_{i=1}^{N} \hat{\mathbf{K}}_i^{-1/2} \left(\hat{\mathbf{v}}_i \hat{\mathbf{v}}_i'\right) \hat{\mathbf{K}}_i^{-1/2} \tag{1.4.24}$$

With this choice of $\hat{\mathbf{R}}$, we can make a direct connection with the GEE literature by ignoring the presence of c_i and working off the first two conditional moments of $\mathbf{y}_i$ given $\mathbf{x}_i$ – see, for example, Liang and Zeger (1986) and Wooldridge (2010, Sections 13.11.4 and 18.7.3). Namely, under the full set of working assumptions with $\delta = 0$,

$$\begin{aligned} \mathrm{E}\left(y_{it}|\mathbf{x}_i\right) &= \exp\left[\mathbf{x}_{it}\boldsymbol{\beta}_0 + \eta + \mathbf{h}\left(\mathbf{x}_i\right)\boldsymbol{\lambda}\right] = k_{it},\ t = 1, ..., T & (1.4.25) \\ \mathrm{Var}\left(y_{it}|\mathbf{x}_i\right) &= \alpha \mathrm{E}\left(y_{it}|\mathbf{x}_i\right),\ t = 1, ..., T & (1.4.26) \\ \mathrm{Corr}\left(\mathbf{y}_i|\mathbf{x}_i\right) &= \alpha \mathbf{K}_i^{1/2}\mathbf{R}\mathbf{K}_i^{1/2} & (1.4.27) \end{aligned}$$

This collection of moment assumptions is precisely what is used in GEE applications of Poisson regression (whether or not y_{it} is a count variable), with the addition of the vector of functions $\mathbf{h}\left(\mathbf{x}_i\right)$. We emphasize that these are *all* working assumptions in the current context. Not even the conditional mean function in (1.4.25) is assumed to hold for consistency because (1.4.25) is obtained from Assumptions CM and WH.1, whereas we are only require Assumption CM for consistency. We impose Assumptions WH.1 and WH.2 in order to estimate $\mathbf{R}$ and then to estimate $\mathbf{\Omega}_i$. Provided it leads to a positive definite estimate, we prefer (1.4.20) because it is the correct expression under all of the working assumptions.

Under Assumption CM and the full set of working assumptions, we can estimate the optimal IVs, for each i, as

$$\nabla_{\beta}\hat{\mathbf{m}}_i'\hat{\mathbf{M}}_i^{-1/2}\left[\hat{\mathbf{R}}^{-1}-\frac{1}{\sqrt{\hat{\mathbf{m}}_i}'\hat{\mathbf{R}}^{-1}\sqrt{\hat{\mathbf{m}}_i}}\hat{\mathbf{R}}^{-1}\sqrt{\hat{\mathbf{m}}_i}\sqrt{\hat{\mathbf{m}}_i}'\hat{\mathbf{R}}^{-1}\right]\hat{\mathbf{M}}_i^{-1/2} \tag{1.4.28}$$

where "^" means the quantity is evaluated at a first-round estimator, most likely $\widehat{\boldsymbol{\beta}}_{FEP}$, and $\hat{\mathbf{R}}$ is from (1.4.22) or, if necessary, (1.4.24). [In either case, $\hat{\alpha}$ drops out of (1.4.28).] However, without the full set of working assumptions, this choice of IVs is *not* guaranteed to improve over the FEP estimator because of its dependence on $\hat{\mathbf{R}}$. A somewhat subtle point is that (1.4.28) is not even optimal under Assumptions CM, WV.1, and WV.2 because consistency of $\hat{\mathbf{R}}$ for $\mathbf{R}$ generally requires correct specification of the heterogeneity mean and variance – that is, Assumptions WH.1 and WH.2. As mentioned previously, if we did not have to estimate $\mathbf{R}$, we could use (1.4.28) with $\hat{\mathbf{R}}$ replaced by $\mathbf{R}$, and then we would have just identification as with the FEP estimator. Naturally, we want to use the data to provide an estimator of $\mathbf{R}$ better than just guessing. Incidentally, expression (1.4.28) shows that the estimator $\hat{\alpha}$ has no direct effect on the optimal IVs because it factors out as a constant.

In order to ensure improvements over FEP, our recommendation is to stack the FEP and the new "optimal" IVs to form an expanded IV matrix and use GMM. The resulting estimator, which we simply call the "GMM estimator," is guaranteed to be asymptotically at least as efficient as the FEP and GFEP estimators; usually it is strictly more efficient than both. In other words, the $T \times 2K$ matrix of IVs is $\hat{\mathbf{Z}}_i$, written in transposed form as

$$\hat{\mathbf{Z}}_i' = \begin{pmatrix} \nabla_{\beta}\hat{\mathbf{m}}_i'\hat{\mathbf{M}}_i^{-1/2}\left[\mathbf{I}_T-\sqrt{\hat{\mathbf{p}}_i}\sqrt{\hat{\mathbf{p}}_i}'\right]\hat{\mathbf{M}}_i^{-1/2} \\ \nabla_{\beta}\hat{\mathbf{m}}_i'\hat{\mathbf{M}}_i^{-1/2}\left[\hat{\mathbf{R}}^{-1}-\frac{1}{\sqrt{\hat{\mathbf{m}}_i}'\hat{\mathbf{R}}^{-1}\sqrt{\hat{\mathbf{m}}_i}}\hat{\mathbf{R}}^{-1}\sqrt{\hat{\mathbf{m}}_i}\sqrt{\hat{\mathbf{m}}_i}'\hat{\mathbf{R}}^{-1}\right]\hat{\mathbf{M}}_i^{-1/2} \end{pmatrix} \tag{1.4.29}$$

Given this choice of $\hat{\mathbf{Z}}_i$, the mechanics of GMM are straightforward. After obtaining $\widehat{\boldsymbol{\beta}}_{FEP}$, obtain the $T \times 1$ residual vectors

$$\tilde{\mathbf{u}}_i = \mathbf{y}_i - \mathbf{p}\left(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}_{FEP}\right)n_i \tag{1.4.30}$$

Then, given the estimators of η, $\boldsymbol{\lambda}$, α, δ, and $\mathbf{R}$ described above, obtain the $2K \times 2K$ matrix,

$$\hat{\boldsymbol{\Psi}} = N^{-1} \sum_{i=1}^{N} \hat{\mathbf{Z}}_i' \tilde{\mathbf{u}}_i \tilde{\mathbf{u}}_i' \hat{\mathbf{Z}}_i \tag{1.4.31}$$

Assuming $\hat{\boldsymbol{\Psi}}$ is positive definite (which generally holds with probability approaching one), the optimal GMM estimator, $\hat{\beta}_{GMM}$, solves

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^K} \left(\sum_{i=1}^{N} \mathbf{u}_i (\boldsymbol{\beta})' \hat{\mathbf{Z}}_i \right) \hat{\boldsymbol{\Psi}}^{-1} \left(\sum_{i=1}^{N} \hat{\mathbf{Z}}_i' \mathbf{u}_i (\boldsymbol{\beta}) \right) \tag{1.4.32}$$

Because we have chosen very smooth mean, variance, and correlation functions, the consistency and $\sqrt{N}$-asymptotic normality are standard; see, for example, Wooldridge (2010, Chapter 14). Remember, $\hat{\boldsymbol{\Psi}}^{-1}$ is an (estimated) optimal weighting matrix given the choice of instruments; the standard GMM inference does not require that $\hat{\mathbf{Z}}_i$ is optimal.

Regardless of the size of T, the GMM estimator generates K overidentification restrictions that can be used to test Assumption CM.

1.5 A Small Simulation Study

We now present the results of a small Monte Carlo simulation to demonstrate the efficacy of the improved GMM estimator. The conditional mean model, which has an exponential form, includes three time-varying explanatory variables and multiplicative heterogeneity. We consider two conditional distributions for the outcome variable, y_{it}. In the first case, y_{it} is a count variable generated as

$$y_{it} | \mathbf{x}_i, c_i, \mathbf{e}_i \sim \text{Poisson}\left[c_i \exp\left(\mathbf{x}_{it}\boldsymbol{\beta} + e_{it}\right)\right] \tag{1.5.1}$$

where $\mathbf{e}_i = (e_{i1}, e_{i2}, ..., e_{iT})'$ is distributed as multivariate normal with unit variances. In order to generate serial dependence in $\{y_{it} : t = 1, ..., T\}$ conditional on $(\mathbf{x}_i, c_i)$, $\{e_{it} : t = 1, 2, ..., T\}$ follows an AR(1) process with first-order correlation $\phi \in \{0, 0.25, 0.75\}$. This autoregressive process generates no conditional dependence when $\phi = 0$ and fairly strong time series dependence when $\phi = 0.75$. Because of the inclusion of e_{it}, the conditional distribution $\mathrm{D}\left(y_{it} | \mathbf{x}_i, c_i\right)$ is not Poisson; in fact, it exhibits overdispersion because $\exp(e_{it})$ is integrated out in obtaining $\mathrm{D}(y_{it}|\mathbf{x}_i, c_i)$. However, consistency of the estimators requires only that that $\mathrm{E}(y_{it}|\mathbf{x}_i, c_i)$ has the exponential form with multiplicative c_i.

The strictly exogenous explanatory variables, $\mathbf{x}_{it}$, are generated as a trivariate, stationary vector autoregression, where the stochastic term is an independent multivariate standard normal distribution with autocorrelation parameter 0.125. The processes $\mathbf{x}_i = (\mathbf{x}_{i1}, ..., \mathbf{x}_{iT})$ and $\mathbf{e}_i$ are independent. The vector $\boldsymbol{\beta}$ is set to $\boldsymbol{\beta}' = (0.15, 0.25, 0.35)$ (where we drop the o subscript to make the tables easier to read).

To generate correlation between c_i and $\mathbf{x}_i$, we first use an exponential version of the Mundlak (1978) device and an exponential distribution:

$$c_i|\mathbf{x}_i \sim \text{Exponential}\left[\exp\left(\eta + \bar{\mathbf{x}}_i\boldsymbol{\lambda}\right)\right] \tag{1.5.2}$$

Under this specification, the working assumptions WH.1 and WH.2 are both satisfied with $\mathbf{h}(\mathbf{x}_i) = \bar{\mathbf{x}}_i$ and, in the case of WH.2, $\delta = 1$.

We estimate the parameters in the heterogeneity moments using a two-step pooled Poisson QMLE with the FEP estimator as the first-stage estimator of $\boldsymbol{\beta}$. The estimates $\hat{\alpha}$ and $\hat{\delta}$ are estimated via the pooled OLS regression in equation (1.4.20) and $\hat{\mathbf{R}}$ is estimated as in (1.4.22). When $\hat{\mathbf{R}}$ is not positive definite for a particular draw, we set $\hat{\delta} = 0$ and estimate $\hat{\mathbf{R}}$ as in (1.4.24) (in which case the value of $\hat{\alpha}$ plays no role in the estimation of $\boldsymbol{\beta}$). This situation occurs between 60% and 80% of the simulations.

We use $N = 300$, $T \in \{4, 8\}$, and $1,000$ replications in the simulations. The findings are reported in Table 1.1.

Some general patterns emerge from Table 1.1. First, the FEP estimator shows very little bias, and its bias is almost always smaller than the GFEP and GMM estimators. The GFEP estimator generally shows the most bias – as high as nine percent in some cases. Still, we only have $N = 300$, which is not especially large. Interestingly, the bias in the GMM estimator – which combines both sets of moment conditions – is well below that of the GFEP estimator. The bias in both the GFEP

Table 1.1: Conditional Poisson distribution

			Bias			**SD**			**RMSE**	
		FEP	**GFEP**	**GMM**	**FEP**	**GFEP**	**GMM**	**FEP**	**GFEP**	**GMM**
$\boldsymbol{\phi} = \mathbf{0}$	**T = 4**	0.002	-0.004	0.000	0.082	0.075	0.072	0.082	0.075	0.072
		0.001	-0.011	-0.003	0.083	0.078	0.072	0.083	0.079	0.072
		-0.001	-0.016	-0.005	0.083	0.079	0.075	0.083	0.081	0.075
	T = 8	0.011	-0.010	-0.005	0.052	0.044	0.041	0.052	0.045	0.041
		0.000	-0.020	-0.011	0.053	0.044	0.042	0.053	0.049	0.044
		0.001	-0.027	-0.014	0.051	0.045	0.042	0.052	0.052	0.045
$\boldsymbol{\phi} = \mathbf{0}.25$	**T = 4**	-0.007	-0.016	0.008	0.081	0.074	0.072	0.081	0.076	0.073
		-0.003	-0.014	0.004	0.082	0.075	0.070	0.079	0.077	0.070
		0.002	-0.015	0.003	0.079	0.075	0.070	0.079	0.077	0.070
	T = 8	-0.001	-0.014	-0.007	0.051	0.045	0.042	0.051	0.047	0.043
		0.000	-0.021	-0.010	0.048	0.044	0.040	0.048	0.049	0.042
		-0.001	-0.029	-0.015	0.051	0.046	0.043	0.051	0.054	0.046
$\boldsymbol{\phi} = \mathbf{0}.75$	**T = 4**	-0.001	-0.007	-0.003	0.057	0.054	0.051	0.057	0.055	0.051
		0.005	-0.008	0.001	0.060	0.058	0.052	0.061	0.059	0.052
		0.001	-0.014	-0.002	0.060	0.059	0.053	0.060	0.060	0.053
	T = 8	0.001	-0.012	-0.004	0.043	0.035	0.034	0.043	0.037	0.034
		-0.001	-0.023	-0.011	0.044	0.036	0.034	0.044	0.043	0.036
		-0.002	-0.032	-0.015	0.047	0.038	0.036	0.047	0.050	0.039

and GMM estimators appears to increase with T. Overall, the bias in the GMM estimator seems acceptable, especially given the small N.

The GMM estimator always has the smallest sampling standard deviation, sometimes being about 80% of the FEP standard error. The SD of the GFEP estimator falls in between that of the FEP and GMM estimators. In a few cases the FEP estimator has smaller root mean squared error (RMSE) than the GFEP estimator. The asymptotic theory of GMM estimation implies that the GMM estimator is asymptotically more efficient than FEP or GFEP because, in the setting of the simulation, the entire set of working assumptions does not hold, and so GFEP does not use the optimal IVs. The ranking of the estimators in terms of the root mean squared error favors the GMM estimator in every case.

To see how the estimators perform when y_{it} is a continuous outcome, we generated y_{it} as

$$y_{it}|\mathbf{x}_i, c_i, \mathbf{e}_i \sim \text{Gamma}\left[\exp\left(\mathbf{x}_{it}\boldsymbol{\beta} + e_{it}\right), c_i\right] \tag{1.5.3}$$

where the gamma distribution is parameterized so that $\mathrm{E}\left(y_{it}|\mathbf{x}_{it}, c_i, \mathbf{e}_i\right) = c_i \exp\left(\mathbf{x}_{it}\boldsymbol{\beta} + e_{it}\right)$, as before. The conditional variance is $\mathrm{Var}\left(y_{it}|\mathbf{x}_{it}, c_i, \mathbf{e}_i\right) = c_i^2 \exp\left(\mathbf{x}_{it}\boldsymbol{\beta} + e_{it}\right)$. We use the same process in (1.5.2) to generate c_i. The simulation findings are reported in Table 1.2.

Table 1.2: Conditional Gamma distribution

			Bias			**SD**			**RMSE**	
		FEP	**GFEP**	**GMM**	**FEP**	**GFEP**	**GMM**	**FEP**	**GFEP**	**GMM**
$\boldsymbol{\phi} = \mathbf{0}$	**T = 4**	0.000	-0.006	-0.002	0.090	0.087	0.081	0.090	0.087	0.081
		0.003	-0.008	0.003	0.089	0.085	0.080	0.089	0.085	0.080
		0.001	-0.014	0.000	0.090	0.088	0.083	0.090	0.089	0.083
	T = 8	0.000	-0.012	-0.006	0.056	0.049	0.048	0.056	0.051	0.048
		-0.001	-0.019	-0.009	0.052	0.050	0.047	0.052	0.054	0.048
		-0.001	-0.027	-0.014	0.054	0.051	0.048	0.054	0.058	0.050
$\boldsymbol{\phi} = \mathbf{0}.25$	**T = 4**	0.002	-0.007	0.002	0.086	0.082	0.078	0.086	0.082	0.078
		-0.003	-0.016	-0.004	0.085	0.082	0.077	0.085	0.084	0.078
		0.002	-0.014	-0.001	0.086	0.084	0.081	0.086	0.085	0.081
	T = 8	0.000	-0.013	-0.006	0.057	0.050	0.048	0.057	0.052	0.048
		0.000	-0.019	-0.009	0.055	0.050	0.048	0.055	0.053	0.049
		-0.001	-0.033	-0.017	0.058	0.053	0.051	0.058	0.062	0.053
$\boldsymbol{\phi} = \mathbf{0}.75$	**T = 4**	0.001	-0.006	0.000	0.069	0.067	0.063	0.069	0.067	0.063
		0.000	-0.012	-0.001	0.074	0.072	0.067	0.074	0.073	0.067
		0.000	-0.016	-0.001	0.070	0.072	0.064	0.070	0.074	0.064
	T = 8	0.001	-0.014	-0.005	0.049	0.041	0.040	0.049	0.044	0.040
		0.000	-0.023	-0.008	0.048	0.042	0.039	0.048	0.048	0.040
		-0.001	-0.034	-0.013	0.050	0.046	0.043	0.050	0.057	0.045

The general pattern found in Table 1.1 continues to hold in Table 1.2. The FEP estimator generally has the lowest bias, although the GMM estimator also does well with bias. The GFEP estimator, which uses only the "optimal" IVs, shows more bias – again, sometimes on the order of more than nine percent. In terms of precision and RMSE, the GMM estimator outperforms FEP and GFEP in all scenarios, although the gains are modest in some cases.

We tried several additional scenarios, including cases where Assumption WH.2 is violated – by drawing c_i from a Poisson distribution – and cases where, conditional on $(\mathbf{x}_i, c_i)$ – y_{it} is an underdispersed gamma random variable. In the former case, we found only minor differences among the estimators, although sometimes the FEP estimator outperformed the other two in terms of RMSE. In the latter case, where we did not allow serial correlation, the estimators perform very similarly. As a final set of simulations, we misspecified the conditional mean $\mathrm{E}(c_i|\mathbf{x}_i)$ in (1.5.2) by letting the mean depend on the average of the first and last time periods rather than $\bar{\mathbf{x}}_i$. In other words, Assumption WH.1 is violated. The GMM estimator uniformly performed the best based on RMSE and exhibited biases on the order of those reported in Tables 1.1 and 1.2. These simulations are available upon request from the authors.

1.6 Summary and Conclusion

We have characterized the optimal instruments in a multiplicative panel model under a general set of working assumptions. The variance-mean relationship, conditional on unobserved heterogeneity as well as covariates, is allowed to be any positive number. The conditional correlation matrix is assumed to be constant but is otherwise unrestricted. Under these assumptions, the optimal IVs depends only on the unknown correlation matrix, $\mathbf{R}$ (and the value of the conditional mean parameters, $\boldsymbol{\beta}_0$). In the special case that $\mathbf{R} = \mathbf{I}_T$, we show that the FEP estimator achieves the asymptotic efficiency bound for any amount of overdispersion or underdispersion, thereby relaxing the assumptions under which the FEP estimator is known to be asymptotically efficient. When $\mathbf{R}$ is not the identity matrix, it is possible to improve on the FEP estimator.

To operationalize the optimal IVs in order to exploit serial correlation, we add working first and second moment assumptions on the conditional heterogeneity distribution. These assumption are common in literatures that allows nonnegative heterogeneity in cross-sectional and panel data models. We show that estimating the optimal IVs is straightforward, and suggest a GMM approach that is guaranteed to improve asymptotic efficiency whether or not serial correlation is present. Our simulations show that the GMM estimator that combines the FEP moment conditions and the new "optimal" moment conditions has very good bias properties and provides nontrivial efficiency gains – even when the cross-sectional sample size is only $N = 300$.

Our results and new estimator are appealing for cases where N is substantially larger than T, as we have used the standard microeconometric setting where T is fixed in the asymptotic analysis. Naturally, this is not the only possibility. For example, Fernández-Val and Weidner

(2018) and Chen, Fernández-Val, and Weidner (2020) have proposed quasi-MLEs that allow more heterogeneity. However, consistency requires $T \to \infty$ along with $N \to \infty$, and necessarily restricts the amount of time series heterogeneity and dependence.

CHAPTER 2

INFORMATION EQUIVALENCE AMONG TRANSFORMATIONS OF SEMIPARAMETRIC NONLINEAR PANEL DATA MODELS

2.1 Introduction

In the standard linear panel data model with additive unobserved heterogeneity, it is well known that numerous transformations can be used to eliminate the heterogeneity prior to estimation. The most common methods are the within and first-differencing transformations[1]. Similarly, when the heterogeneity appears as a multiplicative term in the conditional mean like in certain Generalized Linear Model settings, modified within and differencing transformations can control for the heterogeneity and provide moment conditions for estimation. There exist other transformations which control for heterogeneity but are clearly absurd. For example, multiplying all the data by zero eliminates the heterogeneity along with all information for estimation. For a less trivial example, suppose the population model is linear with a single additive effect and the first-differenced errors are homoskedastic and uncorrelated. Then second-differencing is still consistent but less efficient than first-differencing. These examples raise the question of how to evaluate methods for eliminating heterogeneity while preserving information for estimation.

This paper considers conditional mean models with unobserved heterogeneity. The general framework derived within encompasses a large class of both linear and strictly nonlinear models, examples of which are given in Section 2.2.1. The models are referred to as "semiparametric" in the sense that nothing is assumed about the relationship between the heterogeneity and observables other than regularity conditions needed for asymptotic analysis. In place of assumptions on the conditional distribution of the heterogeneity, these models often require a transformation to eliminate or control for the term. I provide a unified framework for comparing such transformations in terms of the information they preserve. Those that yield the same moment conditions, given

[1] For a comprehensive review of linear panel models with additive heterogeneity, see Chapters 10 and 11 of Wooldridge (2010).

certain regularity assumptions, will provide the same $\sqrt{N}$-asymptotic efficiency bound if they have equal rank.

As mentioned above, the within and first-differencing transformations are the most common in the linear panel case for eliminating additive heterogeneity. When the covariates are strictly exogenous with respect to the idiosyncratic errors, these transformations provide conditional moment restrictions which can be exploited for estimation of the population parameters. For a given conditional variance matrix, Arellano and Bover (1995) suggest that Generalized Least Squares (GLS) on the demeaned equations is equivalent to the efficient 3SLS estimator. This claim was later proven in Im et al. (1999) along with a proof that the GLS estimators on the demeaned and first-differenced estimators are equivalent. Their result shows that two commonly used methods of estimation preserve the same information in the linear case. However, they limit their investigation to a small number of estimators and only allow for a single time-invariant individual effect. My approach can derive the same result as Im et al. (1999) as well as general factor-augmented panels with an arbitrary number of individual effects with time-varying coefficients.

For nonlinear models with a multiplicative heterogeneity term, one approach to estimation is the fixed effects Poisson (FEP) estimator. Hausman et al. (1984) derive the FEP as the maximum likelihood estimator of a multinomial distribution[2]. Wooldridge (1999) shows that the FEP is in fact consistent under a much weaker strict exogeneity assumption. One proof of this result shows that the score of the likelihood function has a mean of zero at the true parameter value due the likelihood's implicit transformation of the data. This transformation subtracts the weighted time averages from each outcome and so I refer to it as the generalized within transformation. Another approach is the generalized next-differencing transformation first studied by Chamberlain (1992) and Wooldridge (1997), which subtracts from a time period the next period outcome, weighted by the quotient of the mean functions. While generalized next-differencing was originally proposed for a sequential exogeneity setting, I study it here in the context of strict exogeneity. I also consider the residual maker matrix from regressing on the outcome variable's mean function. To the best of

[2] Similar to the linear fixed effects estimator, the FEP estimator is a true fixed effects procedure as it can be derived by estimating via pooled Poisson regression and treating the multiplicative terms as parameters to estimate.

my knowledge, this paper is the first to show information equivalence of these transformations.

In Section 2.2, I define information equivalence in a first order asymptotic sense. The efficiency bounds studied will apply to "small-T" settings where asymptotics are derived with T fixed as $N \to \infty$. I then derive sufficient conditions under which transformations of the data which yield moment restrictions for estimation preserve the same information. This result is general and can apply to a number of finite and asymptotic settings. In Section 2.3, I apply the main result from Section 2.2 to a nonlinear multiplicative model, a linear model with an unknown factor structure, and a linear random trend model. Section 2.4 discusses further practical suggestions like implementation and extensions. Section 2.5 provides concluding remarks along with potential directions for future research.

2.2 Information equivalence

As mentioned in the Introduction, the results of this section apply to population moments. In what follows, $(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{c}_i)$ is assumed to be a random draw from an infinite population. The matrix $(\boldsymbol{y}_i, \boldsymbol{x}_i)$ is $T \times (1 + K)$ and observable whereas the random $p \times 1$ vector $\boldsymbol{c}_i$ is not. All statements involving expressions of random variables hold almost surely. For example, conditional means and rank conditions for random matrices hold with probability one. Finally, I assume regularity conditions suitable for asymptotic analysis such as bounds on the higher-order moments of the data.

2.2.1 Model

The following conditional mean assumption specifies the empirical setting:

Assumption CM: For $t = 1, ..., T$,

$$E(y_{it}|\boldsymbol{x}_i, \boldsymbol{c}_i) = m_t(\boldsymbol{x}_{it}, \boldsymbol{\beta}_0, \boldsymbol{c}_i) \tag{2.2.1}$$

where $m_t(\boldsymbol{x}, ., \boldsymbol{c}) : \mathbb{R}^K \to \mathbb{R}$ is a known Borel twice-differentiable function for every $\boldsymbol{x} \in \mathcal{X}_t$ and $\boldsymbol{c} \in C$, where $\mathcal{X}_t$ and C are the respective supports of $\boldsymbol{x}_{it}$ and $\boldsymbol{c}_i$. ■

Equation (2.2.1) specifies a nonlinear semiparametric conditional mean function with strictly exogenous covariates where $\boldsymbol{\beta}_0$ is a $K \times 1$ vector of parameters[3]. The mean function itself is allowed to vary over time periods. The heterogeneity is also allowed to enter the mean function in any arbitrary way. In the linear panel case, the simplest and most common specification is an individual-specific intercept. In nonlinear cases, the heterogeneity is often included as a multiplicative term.

I do not place any identifying assumptions directly on m_t. These implicit identification conditions will come later in the form of rank assumptions. Essentially, the results contained in this paper apply to nontrivial empirical situations. For example, consider a model $y_{i1} = c_i + \beta y_{i2}$ where c_i is an individual-specific intercept and y_{i2} is an indicator variable associated with a treatment or policy intervention. If c_i has a mass point at zero, it must be the case that there is variation, so that $y_{i1} \neq 0$ for all i.

The following examples illustrate some common empirical settings for which Assumption CM applies:

Example 1 (Linear model with additive effects): Consider the following specification:

$$y_{it} = c_i + \boldsymbol{x}_{it}\boldsymbol{\beta}_0 + u_{it}$$

This model is common among applied microeconometric researchers. Im et al. (1999) shows that the 3SLS estimator of $\boldsymbol{\beta}_0$ using the differenced covariates as instruments is algebraically equivalent to GLS estimators based off of both the within and differenced transformed residuals[4]. This example will be discussed in Sections 2.2.2 and 2.3.1.

We can include multiple individual effects loaded onto macro shocks in the form

$$y_{it} = \boldsymbol{c}_i'\boldsymbol{f}_t + \boldsymbol{x}_{it}\boldsymbol{\beta}_0 + u_{it}$$

where $\boldsymbol{c}_i'\boldsymbol{f}_t = \sum_{r=1}^{p} c_{ir} f_{rt}$ and $\boldsymbol{f}_t$ is observable. An example of the general setting is the random trend linear model.

$$y_{it} = c_i + a_i t + \boldsymbol{x}_{it}\boldsymbol{\beta}_0 + u_{it}$$

[3] In this context, nonlinear does not mean 'strictly nonlinear', but can also include linear models.

[4] The setting studied by Im et al. is motivated by considering covariates which satisfy $E(\boldsymbol{x}_i \otimes \boldsymbol{u}_i) = \boldsymbol{0}$. The equivalence result provided in their paper, however, is purely algebraic in nature and holds regardless of the covariance between the covariates and idiosyncratic errors.

The standard approach to estimation is to first-difference the outcomes to yield another linear model with only an additive individual effect. If strict exogeneity is assumed with respect to $\boldsymbol{x}_i$, we have the same empirical setting as above, and so the same analysis will apply. I discuss the general model in Section 2.3.2. ■

Example 2 (Exponential mean): Consider the following mean function:

$$E(y_{it}|\boldsymbol{x}_i, c_i) = \exp(c_i + \boldsymbol{x}_{it}\boldsymbol{\beta}_0)$$

The exponential mean function is most popularly employed to study count data. The most common estimator of the parameters in this model is the FEP estimator. Wooldridge (1999) shows that Assumption CM is sufficient for identification using the following transformation:

$$y_{it} - \left(\sum_{s=1}^{T} y_{is}\right)\left(\frac{\exp(\boldsymbol{x}_{it}\boldsymbol{\beta}_0)}{\sum_{s=1}^{T}\exp(\boldsymbol{x}_{is}\boldsymbol{\beta}_0)}\right)$$

This transformation will be referred to as the generalized within transformation and provides the basis of the FEP estimator since it shows up in the score function of the Poisson QMLE and has an expectation of zero conditional on $\boldsymbol{x}_i$. Another possible transformation is

$$y_{it} - y_{i,t+1}\frac{\exp(\boldsymbol{x}_{it}\boldsymbol{\beta}_0)}{\exp(\boldsymbol{x}_{i,t+1}\boldsymbol{\beta}_0)}$$

which I refer to as the generalized next-differencing transformation. Both of this transformations are studied in generality in Section 2.3.

In an analogy to the linear setting, we can discuss an exponential random trend model with multiplicative specification

$$E(y_{it}|\boldsymbol{x}_i, \boldsymbol{c}_i) = c_i a_i^t \exp(\boldsymbol{x}_{it}\boldsymbol{\beta}_0)$$

which can be motivated by the form $E(y_{it}|\boldsymbol{x}_i, \boldsymbol{c}_i) = \exp(\gamma_i + \alpha_i t + \boldsymbol{x}_{it}\boldsymbol{\beta}_0)$. This model has received no attention in the econometric literature to the best of my knowledge. I discuss how the results of this paper could apply to such a model in Section 2.3.1. ■

Example 3 (Production functions): Suppose the dependent variable is firm output which follows the given production technology:

$$Q_{it} = \exp(\epsilon_{it} - c_i)L_{it}^{\beta_1}K_{it}^{\beta_2}$$

where (L, K) are labor and capital stock respectively. The heterogeneity can be written $\exp(-c_i)$. If $E(\epsilon_{it}|c_i, L_{it}, K_{it})$ is assumed constant[5], then the transformations studied in Section 2.3 can be used for estimation of the parameters and average partial effects under weak assumptions on the heterogeneity term. This example serves as an interesting bridge between the linear and nonlinear specifications as production theory can be stated in the above nonlinear fashion, but production function estimation is often carried out after log-linearization for which the results of Im et al. (1999) would apply. The specific form of the error is reminiscent of a stochastic frontier model with a time-invariant inefficiency term. See Section V of Amsler, Lee, and Schmidt (2009). ■

For the general treatment of the paper, I consider transformations of the mean function which provide moment conditions for estimating $\boldsymbol{\beta}_0$. Assumption MAT characterizes such matrix transformations:

Assumption MAT: Let $L \leq T$, and let $\boldsymbol{A}(\boldsymbol{x}, \boldsymbol{\beta})$ be an $L \times T$ matrix which satisfies

$$\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)E(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{c}_i) = \boldsymbol{0} \tag{2.2.2}$$

and is differentiable in $\boldsymbol{\beta}$ over int$(\boldsymbol{\Theta})$ for every $\boldsymbol{x} \in \mathcal{X}$. ■

$\boldsymbol{A}$ is a residual maker matrix which is zero at the true parameter value $\boldsymbol{\beta}_0$. I assume $L \leq T$ which corresponds to the examples studied in Section 2.3. While $L > T$ is theoretically possible and would rely on the same theory of g-inverses employed in this paper, I do not consider such a case. In fact, cases of the examples in Section 2.3 where $L > T$ often correspond to linearly dependent and hence redundant sets of moment conditions.

Under the previous assumptions,

$$E(\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)\boldsymbol{y}_i|\boldsymbol{x}_i) = \boldsymbol{0} \tag{2.2.3}$$

by iterated expectations. We can thus use equation (2.2.3) as the basis of a GMM estimator of $\boldsymbol{\beta}_0$, where any function of $\boldsymbol{x}_i$ can be used as an instruments for $\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)\boldsymbol{y}_i$ to improve efficiency. Note

[5] The value of $E(\epsilon_{it}|L_{it}, K_{it})$ is allowed to differ over time as long as it is not a function of observables. The researcher can then just specify time dummies in the mean function to capture the temporal change.

that $\boldsymbol{A}$ could contain external instrumental variables which do not appear in the mean function. This more general case is considered in Section 2.2.2.

The following Lemma demonstrates a useful fact for characterizing information equivalent transformations and has clear parallels in the linear model case. First define $\boldsymbol{m}_i(\boldsymbol{\beta}) = (m_t(\boldsymbol{x}_{i1}, \boldsymbol{\beta}, \boldsymbol{c}_i), ..., m_t(\boldsymbol{x}_{iT}, \boldsymbol{\beta}, \boldsymbol{c}_i))'$.

Lemma 2.2.1. *Suppose $\boldsymbol{A}(\boldsymbol{x}, \boldsymbol{\beta})$ is an $L \times T$ matrix satisfying Assumption MAT. Then for any $(\boldsymbol{x}^0, \boldsymbol{c}^0) \in \mathcal{X} \times C$ such that $|m_t(\boldsymbol{x}^0_t, \boldsymbol{\beta}_0, \boldsymbol{c}^0)| > 0$ for some t, Rank$(\boldsymbol{A}(\boldsymbol{x}^0, \boldsymbol{\beta}_0)) < T$.*

Proof. $\boldsymbol{A}(\boldsymbol{x}^0, \boldsymbol{\beta}_0)\boldsymbol{m}_i(\boldsymbol{\beta}_0) = \boldsymbol{0}$ over the supports of $\boldsymbol{x}_i$ and $\boldsymbol{c}_i$ by Assumption MAT. As $|m_t(\boldsymbol{x}^0_t, \boldsymbol{\beta}_0, \boldsymbol{c}^0)| > 0$, $\boldsymbol{A}(\boldsymbol{x}^0, \boldsymbol{\beta}_0)$ has a nontrivial null space, and hence its rank is less than T. □

The theory for choosing optimal instruments is well-known: when the conditional variance is nonsingular, the optimal GMM estimator uses instruments $(Var(\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)\boldsymbol{y}_i|\boldsymbol{x}_i)^{-1}E(\nabla_{\boldsymbol{\beta}}\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)\boldsymbol{y}_i|\boldsymbol{x}_i))'$. However, in most nontrivial cases when $\boldsymbol{A}$ is $T \times T$, the conditional variance matrix of $\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)\boldsymbol{y}_i$ is singular even when $Var(\boldsymbol{y}_i|\boldsymbol{x}_i)$ is nonsingular. I make one additional assumption on the transformation studied which allows for such a generality. Assumption SYS specifies consistency of a particular linear system which is necessary for the definition of the asymptotic efficiency bound. It will allow us to use a certain class of generalized inverses when the conditional variance is singular.

Assumption SYS: The system $Var(\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)\boldsymbol{y}_i|\boldsymbol{x}_i)\boldsymbol{F}(\boldsymbol{x}_i) = E(\nabla_{\boldsymbol{\beta}}\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)\boldsymbol{y}_i|\boldsymbol{x}_i)$ is consistent in $\boldsymbol{F}(\boldsymbol{x}_i)$ and $E(\boldsymbol{F}(\boldsymbol{x}_i)'Var(\boldsymbol{A}(\boldsymbol{\beta}_0)\boldsymbol{y}_i|\boldsymbol{x}_i)\boldsymbol{F}(\boldsymbol{x}_i))$ is nonsingular for a given solution. ■

Consistency of a linear system only requires the existence of a solution and not necessarily uniqueness. In fact, Section 2.3 considers relevant cases for which uniqueness does not hold. Assumption SYS is posed in Newey (2001) for studying censored and truncated regression. It holds trivially when the conditional variance is nonsingular, in which case the unique solution is $Var(\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)\boldsymbol{y}_i|\boldsymbol{x}_i)^{-1}E(\nabla_{\boldsymbol{\beta}}\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)\boldsymbol{y}_i|\boldsymbol{x}_i)$. The results in Chamberlain (1987) and Newey (2001) show that the semiparametric efficiency bound for estimating $\boldsymbol{\beta}_0$ using equation (2.2.3) and Assumptions CM, MAT, and SYS is

$$E\left(E(\nabla_{\boldsymbol{\beta}}\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)\boldsymbol{y}_i|\boldsymbol{x}_i)'Var(\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)\boldsymbol{y}_i|\boldsymbol{x}_i)^{-}E(\nabla_{\boldsymbol{\beta}}\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)\boldsymbol{y}_i|\boldsymbol{x}_i)\right)^{-1} \tag{2.2.4}$$

where "−" denotes a symmetric g-inverse[6]. That is, no $\sqrt{N}$-consistent estimator of $\boldsymbol{\beta}_0$ based off of equation (2.2.3) has a smaller asymptotic variance than (2.2.4).

Theorem 5.2 in Newey (2001) shows that the efficiency bound in (2.2.4) is invariant to choice of symmetric g-inverse under Assumption SYS. If the conditional variance is nonsingular, then the g-inverse can be replaced by a proper inverse as in Chamberlain (1987). Otherwise any g-inverse will work as long as the consistency assumption holds. The matrix in (2.2.4) is also equivalent to the asymptotic variance of the GMM estimator based off of the moment conditions in (2.2.3) which uses the optimal instruments $(Var(\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)\boldsymbol{y}_i|\boldsymbol{x}_i)^{-}E(\nabla_{\boldsymbol{\beta}}\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)\boldsymbol{y}_i|\boldsymbol{x}_i))'$. The system is just identified and so no weight matrix is required for the asymptotic bound. Realizing this efficiency bound is the subject of Section 2.4.

The rest of the paper is concerned with studying transformations of the observed data which provide the same semiparametric efficiency bound as defined in (2.2.4). The following definition characterizes the types of transformations I consider:

Definition: Let Assumption CM hold, and let $\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta})$ and $\boldsymbol{B}(\boldsymbol{x}_i, \boldsymbol{\beta})$ be $L \times T$ and $M \times T$, respectively. Given $\boldsymbol{A}$ and $\boldsymbol{B}$ satisfy Assumptions MAT and SYS, the matrices are **information equivalent transformations** if their semiparametric efficiency bounds given by (2.2.4) are equal. ■

Information equivalence defined above is an equivalence relation on the set of $K \times K$ real-valued matrices since it is defined via matrix equivalences. This fact will be used in Section 2.3 to show information equivalence between general forms of applied transformations since it is a transitive property and it is easiest to evaluate the information bound in relation to the generalized within transformation. Information equivalence is similar to the definition of redundancy of moment conditions as given by Breusch et al. (1999). However, the results in this paper are not direct consequences of their redundancy results as I allow the moment conditions to have singular covariance matrices which directly applies to the examples in Section 2.3.

[6] A g-inverse for matrix $\boldsymbol{\Omega}$ is a matrix $\boldsymbol{\Omega}^{-}$ such that $\boldsymbol{\Omega}\boldsymbol{\Omega}^{-}\boldsymbol{\Omega} = \boldsymbol{\Omega}$. This condition is weaker than the Moore-Penrose inverse which requires three other non-redundant properties. It is worth noting that the Moore-Penrose inverse is unique, but a g-inverse is not necessarily; this fact will be used to prove the main results in Section 2.2.2. For a general treatment of g-inverses, see Rao and Mitra (1978).

2.2.2 General equivalence result

I now prove the general unifying theory of information equivalence. Consider the empirical setting proposed in Section 2.2.1 where Assumption CM holds. I suppose there is a $T \times T$ matrix $\boldsymbol{M}(\boldsymbol{z}_i, \boldsymbol{\beta})$ satisfying Assumptions MAT and SYS where $\boldsymbol{z}_i$ is allowed to include any element of $\boldsymbol{x}_i$ as well as outside instruments. Dropping the arguments and writing $\boldsymbol{M}_i = \boldsymbol{M}(\boldsymbol{z}_i, \boldsymbol{\beta}_0)$ for simplicity, we have the following moment conditions:

$$E(\boldsymbol{M}_i \boldsymbol{y}_i | \boldsymbol{z}_i) = \boldsymbol{0} \tag{2.2.5}$$

Equation (2.2.5) includes the case of unconditional moment restrictions.

I denote $\boldsymbol{V}_i = E(\boldsymbol{y}_i \boldsymbol{y}_i' | \boldsymbol{z}_i)$. I now consider transformations which still yield valid moment conditions. Let $\boldsymbol{B}_i = \boldsymbol{B}(\boldsymbol{z}_i, \boldsymbol{\beta}_0)$ be a $J \times T$ matrix such that $E(\boldsymbol{B}_i \boldsymbol{y}_i | \boldsymbol{z}_i) = \boldsymbol{0}$. Now I make the following assumptions which are pivotal for the general result of this section, and thus refer to them as Assumptions GR.1 and GR.2.

Assumption GR.1: $\boldsymbol{B}_i \boldsymbol{M}_i = \boldsymbol{B}_i$ and $Rank(\boldsymbol{M}_i \boldsymbol{V}_i \boldsymbol{M}_i') = Rank(\boldsymbol{M}_i) = J < T$. ■

Assumption GR.2: $Rank(\boldsymbol{B}_i \boldsymbol{V}_i \boldsymbol{B}_i') = Rank(\boldsymbol{B}_i) = J$. ■

The notation for $\boldsymbol{M}_i$ in Assumption GR.1 is motivated by the standard notation for a residual maker matrix. In fact, one possible sufficient condition for Assumption GR.1 is that $Rank(\boldsymbol{V}_i) = J$ and that $\boldsymbol{V}_i$ shares a null space with $\boldsymbol{B}_i$. This assumption would also suffice for Assumption GR.2 since $\boldsymbol{B}_i'$ spans the column space of $\boldsymbol{V}_i$, and is relevant in linear panel models with additive heterogeneity. We can then let $\boldsymbol{M}_i$ be a residual maker matrix from regressing on a basis vector for the null space of $\boldsymbol{B}_i$. Another relevant setting to this paper is when $\boldsymbol{M}_i = \boldsymbol{I}_T - \boldsymbol{P}_i$ where $\boldsymbol{P}_i$ has rank $T - J$ and $\boldsymbol{B}_i \boldsymbol{P}_i = \boldsymbol{0}$. This setting characterizes the nonlinear models studied in Section 2.3 and is also sufficient for Assumptions GR.1 and GR.2.

Given the discussion above, I now prove a lemma which is essential to the proof of the general equivalence result.

Lemma 2.2.2. *$\boldsymbol{B}_i'(\boldsymbol{B}_i \boldsymbol{V}_i \boldsymbol{B}_i')^{-1} \boldsymbol{B}_i$ is a g-inverse of $\boldsymbol{M}_i \boldsymbol{V}_i \boldsymbol{M}_i'$.*

Proof. $\boldsymbol{B}_i'(\boldsymbol{B}_i\boldsymbol{V}_i\boldsymbol{B}_i')^{-1}\boldsymbol{B}_i\boldsymbol{M}_i\boldsymbol{V}_i\boldsymbol{M}_i'\boldsymbol{B}_i'(\boldsymbol{B}_i\boldsymbol{V}_i\boldsymbol{B}_i')^{-1}\boldsymbol{B}_i = \boldsymbol{B}_i'(\boldsymbol{B}_i\boldsymbol{V}_i\boldsymbol{B}_i')^{-1}\boldsymbol{B}_i\boldsymbol{V}_i\boldsymbol{B}_i'(\boldsymbol{B}_i\boldsymbol{V}_i\boldsymbol{B}_i')^{-1}\boldsymbol{B}_i = \boldsymbol{B}_i'(\boldsymbol{B}_i\boldsymbol{V}_i\boldsymbol{B}_i')^{-1}\boldsymbol{B}_i$. Since $Rank(\boldsymbol{B}_i'(\boldsymbol{B}_i\boldsymbol{V}_i\boldsymbol{B}_i')^{-1}\boldsymbol{B}_i) = J$ by Assumption GR.2 and $Rank(\boldsymbol{M}_i\boldsymbol{V}_i\boldsymbol{M}_i') = J$ by Assumption GR.1, $\boldsymbol{B}_i'(\boldsymbol{B}_i\boldsymbol{V}_i\boldsymbol{B}_i')^{-1}\boldsymbol{B}_i$ is a g-inverse of $\boldsymbol{M}_i\boldsymbol{V}_i\boldsymbol{M}_i'$ by Theorem 2.6 of Rao and Mitra (1971). □

Theorem 2.2.1. *The equality*

$$\boldsymbol{B}_i'(\boldsymbol{B}_i\boldsymbol{V}_i\boldsymbol{B}_i')^{-1}\boldsymbol{B}_i = \boldsymbol{M}_i'(\boldsymbol{M}_i\boldsymbol{V}_i\boldsymbol{M}_i')^{-}\boldsymbol{M}_i \tag{2.2.6}$$

holds for any choice of matrix $\boldsymbol{B}_i$ satisfying Assumptions GR.1 and GR.2 for the same $\boldsymbol{M}_i$ and for any g-inverse of $\boldsymbol{M}_i\boldsymbol{V}_i\boldsymbol{M}_i'$.

Proof. By Rao and Mitra (1971, p. 603), the expression

$$\boldsymbol{M}_i'(\boldsymbol{M}_i\boldsymbol{V}_i\boldsymbol{M}_i')^{-}\boldsymbol{M}_i \tag{2.2.7}$$

is invariant to the choice of g-inverse as $Rank(\boldsymbol{M}_i\boldsymbol{V}_i\boldsymbol{M}_i') = Rank(\boldsymbol{M}_i)$ by Assumption GR.1. Since $\boldsymbol{B}_i'(\boldsymbol{B}_i\boldsymbol{V}_i\boldsymbol{B}_i')^{-1}\boldsymbol{B}_i$ is such a g-inverse by Lemma 2.2.2 and $\boldsymbol{B}_i\boldsymbol{M}_i = \boldsymbol{B}_i$ we have

$$\begin{aligned}\boldsymbol{B}_i'(\boldsymbol{B}_i\boldsymbol{V}_i\boldsymbol{B}_i')^{-1}\boldsymbol{B}_i &= \boldsymbol{M}_i'\boldsymbol{B}_i'(\boldsymbol{B}_i\boldsymbol{V}_i\boldsymbol{B}_i')^{-1}\boldsymbol{B}_i\boldsymbol{M}_i \\ &= \boldsymbol{M}_i'(\boldsymbol{M}_i\boldsymbol{V}_i\boldsymbol{M}_i')^{-}\boldsymbol{M}_i\end{aligned}$$

which is independent of $\boldsymbol{B}_i$. □

Equation (2.2.6) of Theorem 1 provides the framework for evaluating information equivalence. To see how, I include an additional orthogonality assumption which simplifies the efficiency bound in (2.2.4).

Assumption ORTH: $\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta})$ is an $L \times T$ matrix, $L \leq T$, such that $\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta})\boldsymbol{m}_i(\boldsymbol{\beta}) = \boldsymbol{0}$ for all $\boldsymbol{\beta}$ is some open ball about $\boldsymbol{\beta}_0$. ■

Assumption ORTH is clearly sufficient for Assumption MAT. The transformations studied in the next section satisfy Assumption ORTH for all values of $\boldsymbol{\beta} \in \mathbb{R}^K$ for which the mean function is well-defined. However it only needs to be defined on a relatively small open and connected set so that it applies with respect to differentiation. Note that ORTH does not say anything about point identification of $\boldsymbol{\beta}_0$. Assumption CM guarantees $E(\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)\boldsymbol{y}_i|\boldsymbol{x}_i) = \boldsymbol{0}$ only at $\boldsymbol{\beta}_0$ because $E(\boldsymbol{y}_{it}|\boldsymbol{x}_i, \boldsymbol{c}_i) = m_t(\boldsymbol{x}_{it}, \boldsymbol{\beta}_0, \boldsymbol{c}_i)$. I also note that every transformation considered in Section 2.3 satisfies Assumption ORTH.

The following lemma is a consequence of Assumption ORTH which greatly simplifies the bound in (2.2.4).

Lemma 2.2.3. *Let $\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta})$ satisfy Assumption ORTH. Then under regularity conditions which allow us to pass the gradient operator through the conditional expectation,*

$$E(\nabla_{\boldsymbol{\beta}}\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)\boldsymbol{y}_i|\boldsymbol{x}_i) = \boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)\nabla_{\boldsymbol{\beta}}\boldsymbol{m}_i(\boldsymbol{\beta}_0)$$

Proof. See Appendix for proof. □

Lemma 2.2.3 greatly simplifies the efficiency bound in (2.2.4). It also allows us to say something about finite sample equivalence among certain types of transformations. I summarize these results here:

Corollary 2.2.1. *Let $\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta})$ be a $L \times T$ matrix satisfying Assumptions MAT, SYS, and ORTH. Then $\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)$ has the following efficiency bound:*

$$E\left(\nabla_{\boldsymbol{\beta}}\boldsymbol{m}_i(\boldsymbol{\beta}_0)'\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)'(\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)E(\boldsymbol{y}_i\boldsymbol{y}_i'|\boldsymbol{x}_i)\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)')^{-}\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)\nabla_{\boldsymbol{\beta}}\boldsymbol{m}_i(\boldsymbol{\beta}_0)\right)^{-1} \tag{2.2.8}$$

Corollary 2.2.2. *Suppose $\boldsymbol{A}_i$ and $\boldsymbol{B}_i$ are $J \times T$ matrices and $\boldsymbol{M}_i$ is a $T \times T$ matrix such that Assumptions GR.1 and GR.2 hold for $\boldsymbol{A}$ and $\boldsymbol{B}$. Further suppose $\boldsymbol{A}_i$, $\boldsymbol{B}_i$, $\boldsymbol{M}_i$, and the conditional gradient $\nabla_{\boldsymbol{\beta}}E(\boldsymbol{y}_i|\boldsymbol{z}_i)$ are independent of $\boldsymbol{\beta}$. Then*

$$\nabla_{\boldsymbol{\beta}}\boldsymbol{m}_i'\boldsymbol{A}_i'(\boldsymbol{A}_i\boldsymbol{V}_i\boldsymbol{A}_i')^{-1}\boldsymbol{A}_i\boldsymbol{m}_i(\boldsymbol{\beta}) = \nabla_{\boldsymbol{\beta}}\boldsymbol{m}_i'\boldsymbol{B}_i'(\boldsymbol{B}_i\boldsymbol{V}_i\boldsymbol{B}_i')^{-1}\boldsymbol{B}_i\boldsymbol{m}_i(\boldsymbol{\beta}) \tag{2.2.9}$$

for any value of $\boldsymbol{\beta}$ in $\boldsymbol{m}_i(\boldsymbol{\beta})$.

Corollary 2.2.1 allows us to directly apply the result from Theorem 2.2.1 to the relevant cases in Section 2.3. For information equivalence, it will suffice to show that the relevant transformations satisfying Assumptions MAT, SYS, and ORTH only need to satisfy a rank assumption to be information equivalent. The choice of $\boldsymbol{M}$ will become apparent based on the empirical setting.

Corollary 2.2.2 gives an even more powerful result than equivalence of efficiency bounds. For example, if the moment conditions in (2.2.5) are conditional on $\boldsymbol{x}_i$, the efficient GMM estimator of $\boldsymbol{\beta}_0$, say $\widehat{\boldsymbol{\beta}}$, solves

$$\sum_{i=1}^{N} \nabla_{\boldsymbol{\beta}} \boldsymbol{m}_i' \boldsymbol{M}_i' (\boldsymbol{M}_i \boldsymbol{V}_i \boldsymbol{M}_i')^{-} \boldsymbol{M}_i \boldsymbol{m}_i(\widehat{\boldsymbol{\beta}}) = \boldsymbol{0} \tag{2.2.10}$$

Corollary 2.2.2 tells us that the efficient estimator based off of $E(\boldsymbol{A}_i \boldsymbol{y}_i | \boldsymbol{x}_i)$ and $E(\boldsymbol{B}_i \boldsymbol{y}_i | \boldsymbol{x}_i)$ are algebraically equivalent. When the transformations are themselves functions of the parameters, implementation of the efficient instruments depends on first-stage estimators whereas the transformation $\boldsymbol{A}_i \boldsymbol{m}_i$ depends on the FOC solution, so the results only hold asymptotically. The proof of Theorem 4.2 in Im et al. (1999) uses a specific form of the argument in the proof above. This fact suggests further applications to panel data transformations with strictly exogenous covariates which I explore in the next section.

2.3 Examples of information equivalence

This section considers the application of Theorem 2.2.1 to a variety of interesting empirical settings.

2.3.1 Multiplicative heterogeneity

I now consider the case of a single multiplicative heterogeneous effect:

$$E(y_{it} | \boldsymbol{x}_i, c_i) = c_i m_t(\boldsymbol{x}_{it}, \boldsymbol{\beta}_0) \tag{2.3.1}$$

This specification has grown in popularity in recent years. For example, see Krapf, Ursprung, and Zimmermann (2017), Fischer, Royer, and White (2018), Castillo, Mejía, and Restrepo (2020), Schlenker and Walker (2016), McCabe and Snyder (2014, 2015), and Williams et al. (2020). The most common specification of equation (2.3.1) in practice is the exponential mean function as demonstrated in Example 2. Often the data generating process is a count variable with a mass point at zero, but the model can apply to any nonnegative outcome. This typically implies $m_t(\boldsymbol{x}, \boldsymbol{\beta}_0) > 0$ for all $\boldsymbol{x} \in \mathcal{X}$ which the rank assumptions made in this section will imply.

I consider the following generalized residual functions first introduced in Example 2:

$$u_{it}(\boldsymbol{\beta}) = y_{it} - \left(\sum_{s=1}^{T} y_{is}\right) p_{it}(\boldsymbol{\beta}) \tag{2.3.2}$$

$$r_{i,t,s}(\boldsymbol{\beta}) = y_{it} - y_{is}\frac{m_t(\boldsymbol{x}_{it}, \boldsymbol{\beta})}{m_s(\boldsymbol{x}_{is}, \boldsymbol{\beta})} \tag{2.3.3}$$

where $p_{it}(\boldsymbol{\beta}) = m_t(\boldsymbol{x}_{it}, \boldsymbol{\beta}) \left(\sum_{s=1}^{T} m_s(\boldsymbol{x}_{is}, \boldsymbol{\beta})\right)^{-1}$. Equation (2.3.2) is reminiscent of the linear within transformation. However, the transformation in the linear case demeans using the time averages, whereas the generalized within transformation weights by the pseudo-probability $p_{it}(\boldsymbol{\beta})$. The generalized differencing residual in equation (2.3.3) allows a large number of differencing procedures, including next- and first-differencing as well as differencing one time period from the others in which t is fixed and s is allowed to vary. Any other number of arbitrary generalized differencing is allowed so long as it produces a full rank transformation.

In contrast to the linear model with an additive effect, the transformations in equations (2.3.2) and (2.3.3) will not eliminate the heterogeneity but still creates valid moment conditions. For example, taking the mean of equation (2.3.3) conditional on $(\boldsymbol{x}_i, c_i)$ gives

$$\begin{aligned} E(r_{i,t,s}(\boldsymbol{\beta}_0)|\boldsymbol{x}_i, c_i) &= c_i m_t(\boldsymbol{x}_{it}, \boldsymbol{\beta}_0) - c_i m_s(\boldsymbol{x}_{is}, \boldsymbol{\beta}_0)\frac{m_t(\boldsymbol{x}_{it}, \boldsymbol{\beta}_0)}{m_s(\boldsymbol{x}_{is}, \boldsymbol{\beta}_0)} \\ &= c_i(m_t(\boldsymbol{x}_{it}, \boldsymbol{\beta}_0) - m_t(\boldsymbol{x}_{it}, \boldsymbol{\beta}_0)) \\ &= 0 \end{aligned}$$

which still yields conditional moment restrictions.

Define the respective $T \times 1$ and $(T-1) \times 1$ residual vectors

$$\boldsymbol{u}_i(\boldsymbol{\beta}) = (\boldsymbol{I}_T - \boldsymbol{p}_i(\boldsymbol{\beta})\boldsymbol{1}')\boldsymbol{y}_i \tag{2.3.4}$$

$$\boldsymbol{r}_i(\boldsymbol{\beta}) = \boldsymbol{D}_i(\boldsymbol{\beta})\boldsymbol{y}_i \tag{2.3.5}$$

where $\boldsymbol{1}$ is a $T \times 1$ vector of ones and $\boldsymbol{D}_i(\boldsymbol{\beta})$ is the $T-1 \times T$ weighted generalized differencing matrix which yields the desired residuals as in (2.3.3). I refer to transformations in Equations (2.3.4) and

(2.3.5) as the **generalized within** and **generalized differencing** transformations respectively. Then an iterated expectations argument shows $E(\boldsymbol{u}_i(\boldsymbol{\beta}_0)|\boldsymbol{x}_i) = \mathbf{0}$ and $E(\boldsymbol{r}_i(\boldsymbol{\beta}_0)|\boldsymbol{x}_i) = \mathbf{0}$. Thus equations (2.3.4) and (2.3.5) satisfy Assumption MAT and suggest moment conditions for efficient GMM estimation which could reach their respective efficiency bounds in (2.2.4).

As discussed in the Introduction, equation (2.3.4) is the foundation of the FEP estimator. The FEP is defined in Hausman et al. (1984) as the MLE of a conditional Multinomial distribution with probability and count parameters $\boldsymbol{p}_i(\boldsymbol{\beta}_0) = (p_{i1}(\boldsymbol{\beta}_0), ..., p_{iT}(\boldsymbol{\beta}_0))'$ and n_i. Wooldridge (1999) shows that the FEP is consistent under Assumption CM using the fact that equation (2.3.4) has a zero conditional mean at $\boldsymbol{\beta}_0$ regardless of the true distribution of $\boldsymbol{y}_i|\boldsymbol{x}_i$. This robustness result helped lead to its proliferation in empirical research. As for efficiency, Hahn (1997) shows that the FEP is asymptotically efficient under the full set of Multinomial distributional assumptions. Verdier (2018) strengthens this result substantially by showing efficiency under just zero conditional correlation and conditional mean-variance equality. Brown and Wooldridge (2021) extends this result to allow arbitrary constant conditional mean-variance dispersion.

Equation (2.3.5) was first studied by Chamberlain (1992) and Wooldridge (1997) in the context of next-differencing for nonlinear models. It can also allow for estimation of $\boldsymbol{\beta}_0$ under weaker forms of exogeneity, like sequential exogeneity in the next-differencing case of $s = t + 1$, rather than the strict exogeneity implied by Assumption CM. Sequential exogeneity allows the researcher to specify lag dynamics in the mean function which violates strict exogeneity. However, remarkably less is known about efficient estimation based off of equation (2.3.5) when compared to equation (2.3.4) in the context of strict exogeneity as studied here.

The transformations defined in (2.3.4) and (2.3.5) are clearly not the only transformations which satisfy Assumption MAT. Consider the residual maker matrix from regressing on the mean function defined by equation (2.3.1): $(\boldsymbol{I}_T - \boldsymbol{m}_i(\boldsymbol{\beta})(\boldsymbol{m}_i(\boldsymbol{\beta})'\boldsymbol{m}_i(\boldsymbol{\beta}))^{-1}\boldsymbol{m}_i(\boldsymbol{\beta})')$. This matrix satisfies Assumption ORTH and thus Assumption MAT since it is algebraically orthogonal to the mean function by construction. It is also well-known that the matrix is symmetric, idempotent, and has rank $T - 1$. I will refer to this matrix as the **residual maker** transformation.

The main theorem of this section proves the information equivalence between the generalized within, generalized differencing, and residual maker transformations. This result is similar to Theorem 4.2 of Im et al. (1999) who proves algebraic equivalences of GLS estimators based off of strictly exogenous covariates in linear panel data models with additive effects. There are two primary differences between Theorem 2.3.1 in this paper and Theorem 4.2 in Im et al. First, the heterogeneity is multiplicative rather than additive. This difference is not made without loss of generality as rewriting the terms causes the heterogeneity to have time variation[7]. Second, Im et al. shows an algebraic equivalence between the estimators studied, while I show an asymptotic equivalence. As mentioned after Theorem 2.2.1, finite sample equality will not necessarily follow when the transformations are functions of the parameter $\boldsymbol{\beta}_0$ and require a first-step estimator to implement.

By Lemma 2.2.1, the conditional variance of the generalized within transformation is necessarily singular, so I will need to show that its efficiency bound is well-defined and invariant to the choice of symmetric g-inverse. Lemma 1 of Verdier (2018) shows that it has rank $T-1$ at the true parameter value. This fact suggests that deleting a row to remove the rank degeneracy leads to a transformation with a nonsingular variance matrix. Im et al. (1999) takes this approach when showing equivalence between the within and differenced linear estimators. Let $\boldsymbol{Q}$ be a $T-1 \times T$ matrix which removes any arbitrary row from a given $T \times T$ matrix. Then the transformation $\boldsymbol{Q}(\boldsymbol{I}_T - \boldsymbol{p}_i(\boldsymbol{\beta}_0)\mathbf{1}')$ is the generalized within transformation with an arbitrary row deleted. A similar procedure can be used to make the residual maker transformation full rank. The main result will show that information equivalence is invariant to the row deleted.

Lemma 2.3.1 will show that the efficiency bound of the within and residual maker transformations are well-defined. First I will assume that $E(\boldsymbol{y}_i\boldsymbol{y}_i'|\boldsymbol{x}_i)$ is strictly positive definite, a weaker assumption than the conditional variance of $\boldsymbol{y}_i$ itself being positive definite. Under this assumption, the conditional variance of the generalized differencing transformation is nonsingular under a rank condition provided below. Before I can verify Assumption SYS, I will need an additional rank

[7] If $y_{it} = m_t + u_i$, then rewriting this as $y_{it} = c_i m_t$ implies $c_i = \frac{m_t+u_i}{m_t}$ which depends on the time period specified.

assumption for each respective transformation.

Assumption RK.1: $Rank(\boldsymbol{D}_i(\boldsymbol{\beta}_0)) = T - 1$. ■

Assumption RK.1 states that the differencing matrix has full row rank. It requires that none of the differences used for estimation are redundant in the sense that some row or rows are linear combinations of the others. Necessarily the researcher cannot reuse rows, and if y_{it} is differenced from y_{is}, then y_{is} cannot be differenced from y_{it}. Further, we must have $s \neq t$ for each row so that $\boldsymbol{D}$ does not have any zero rows. For example, including all pairwise differences leads to linear dependence which causes RK.1 to fail.

Assumption RK.2: Let $\boldsymbol{\Sigma}_i = E(\boldsymbol{y}_i\boldsymbol{y}_i'|\boldsymbol{x}_i)$ be positive definite. Define $\boldsymbol{V}_i^- = (\boldsymbol{\Sigma}_i^{-1} - \frac{1}{a_i}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{m}_i(\boldsymbol{\beta}_0)\boldsymbol{m}_i(\boldsymbol{\beta}_0)'\boldsymbol{\Sigma}_i^{-1})$ where $a_i = \boldsymbol{m}_i(\boldsymbol{\beta}_0)'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{m}_i(\boldsymbol{\beta}_0)$. Then the square matrix $E(\nabla_{\boldsymbol{\beta}}\boldsymbol{m}_i(\boldsymbol{\beta}_0)'\boldsymbol{V}_i^-\nabla_{\boldsymbol{\beta}}\boldsymbol{m}_i(\boldsymbol{\beta}_0))$ has full rank. ■

$\boldsymbol{V}_i^-$ is a symmetric g-inverse of $Var((\boldsymbol{I}_T - \boldsymbol{p}_i(\boldsymbol{\beta}_0)\boldsymbol{1}')\boldsymbol{y}_i|\boldsymbol{x}_i)$. In fact, it also satisfies the property

$$\boldsymbol{V}_i^- \left[Var((\boldsymbol{I}_T - \boldsymbol{p}_i(\boldsymbol{\beta}_0)\boldsymbol{1}')\boldsymbol{y}_i|\boldsymbol{x}_i)\right] \boldsymbol{V}_i^- = \boldsymbol{V}_i^- \tag{2.3.6}$$

as shown in Lemma 2 of Verdier (2018) so that it is a reflexive inverse and is also clearly symmetric. Assumption RK.2 suffices for the bound in (2.2.4) existing, as I show in the next lemma that $\boldsymbol{V}_i^-\nabla_{\boldsymbol{\beta}}\boldsymbol{m}_i(\boldsymbol{\beta}_0)$ is a solution to the system in Assumption SYS. This fact, along with the fact that $\boldsymbol{V}_i^-\boldsymbol{m}_i(\boldsymbol{\beta}_0) = \boldsymbol{0}$ and Lemma 2.2.3, gives the bound in (2.2.4) as the expectation above. The following lemma shows that all transformations studied satisfy Assumption SYS and so any symmetric g-inverse will suffice.

Lemma 2.3.1. *Suppose Assumptions CM, RK.1, and RK.2 hold and that $E(\boldsymbol{y}_i\boldsymbol{y}_i'|\boldsymbol{x}_i)$ is positive definite. Then the generalized differencing, generalized within, and residual maker transformations satisfy Assumption SYS. Further, either of the $T \times T$ transformations with any arbitrary row deleted also satisfy Assumption SYS.*

Proof. See Appendix for proof. □

The main consequence of Lemma 2.3.1 is that the asymptotic efficiency bound is well-defined and invariant to symmetric g-inverse for all of the transformations studied in this section. Now I can formally state the application of the main equivalence theorem to the transformations studied in this section. First note that Assumptions CM, RK.1, RK.2, and the positive definiteness of $E(\boldsymbol{y}_i\boldsymbol{y}_i'|\boldsymbol{x}_i)$ are sufficient for each of the transformations studied to satisfy Assumptions SYS and ORTH (and thus MAT) so that their asymptotic efficiency bounds are well-defined and given by (2.2.8).

Theorem 2.3.1. *Suppose Assumptions CM, RK.1, and RK.2 hold and that $E(\boldsymbol{y}_i\boldsymbol{y}_i'|\boldsymbol{x}_i)$ is positive definite. $(\boldsymbol{I}_T - \boldsymbol{p}_i(\boldsymbol{\beta}_0)\mathbf{1}')$, $\boldsymbol{D}_i(\boldsymbol{\beta}_0)$, $(\boldsymbol{I}_T - \boldsymbol{m}_i(\boldsymbol{\beta})(\boldsymbol{m}_i(\boldsymbol{\beta})'\boldsymbol{m}_i(\boldsymbol{\beta}))^{-1}\boldsymbol{m}_i(\boldsymbol{\beta})')$, $\boldsymbol{Q}(\boldsymbol{I}_T - \boldsymbol{p}_i(\boldsymbol{\beta}_0)\mathbf{1}')$, and $\boldsymbol{Q}((\boldsymbol{I}_T - \boldsymbol{m}_i(\boldsymbol{\beta})(\boldsymbol{m}_i(\boldsymbol{\beta})'\boldsymbol{m}_i(\boldsymbol{\beta}))^{-1}\boldsymbol{m}_i(\boldsymbol{\beta})'))$ are information equivalent and invariant to the row deleted by $\boldsymbol{Q}$.*

Proof. See Appendix for proof. □

The proof of Theorem 2.3.1 is independent of which row is deleted in choosing $\boldsymbol{Q}$ and the type of differencing chosen in $\boldsymbol{D}$ satisfying Assumption RK.1, reinforcing the importance of the rank assumptions. As in Theorem 2.2.1, transformations with rank $L < T$ can be shown to be information equivalent via a similar argument, but this fact is not directly relevant to the current results. It's also important to note that the list of information equivalent transformations is not necessarily exhaustive, as any $T \times T$ or $T - 1 \times T$ matrix with rank $T - 1$ and respective orthogonality condition will be information equivalent to the transformations in Theorem 2.3.1 by Theorem 2.2.1.

Similar to the discussion after Theorem 2.2.1, the results in Theorem 2.3.1 could also apply to mean function which have already been transformed. For example, consider the multiplicative random trend from Example 2, $y_{it} = c_i a_i^t m_t(\boldsymbol{x}_{it}, \boldsymbol{\beta}_0)u_{it}$ where u_{it} is an idiosyncratic error. If we assume the outcomes are bounded away from zero, we could first divide each outcome by the previous period. We now have the multiplicative model $y_{it}^* = a_i \frac{m_t(\boldsymbol{x}_{it}, \boldsymbol{\beta}_0)}{m_{t-1}(\boldsymbol{x}_{i,t-1}, \boldsymbol{\beta}_0)} \frac{u_{it}}{u_{i,t-1}}$. If $\frac{u_{it}}{u_{i,t-1}}$ is independent of $\boldsymbol{x}_i$ and a_i with mean 1, we have the model from equation (2.3.1). Then all of the transformations studied here are information equivalent on the transformed outcomes $\boldsymbol{y}_i^*$.

2.3.2 Linear factor model

This section considers linear panels with a factor-augmented error:

$$y_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta}_0 + \boldsymbol{f}_t'\boldsymbol{\gamma}_i + u_{it} \tag{2.3.7}$$

where $\boldsymbol{f}_t$ is a $p \times 1$ vector of common factors. Stacking the factors into the $T \times p$ matrix $\boldsymbol{F} = (\boldsymbol{f}_1, ..., \boldsymbol{f}_T)'$, Pesaran (2006) adds the additional reduced form equation

$$\boldsymbol{x}_i = \boldsymbol{F}\boldsymbol{\Gamma}_i + \boldsymbol{v}_i \tag{2.3.8}$$

where $\boldsymbol{\Gamma}_i$ is a $p \times K$ matrix of “factor loadings" and $\boldsymbol{v}_i$ is a $T \times K$ matrix of mean zero idiosyncratic errors. Write $\boldsymbol{z}_i = (\boldsymbol{y}_i, \boldsymbol{x}_i)$. Under the assumptions in Pesaran (2006), equations (2.3.7) and (2.3.8) imply

$$E(\boldsymbol{z}_i) = \boldsymbol{FCQ} \tag{2.3.9}$$

where $\boldsymbol{CQ}$ is a $p \times K + 1$. Assuming $p \leq K + 1$, $\boldsymbol{CQ}$ is full rank which suggests that $E(\boldsymbol{z}_i)$ can control for the space spanned by $\boldsymbol{F}$. The pooled common correlated effects estimator (CCEP) is defined as

$$\widehat{\boldsymbol{\beta}}_{CCEP} = \left(\sum_{i=1}^{N} \boldsymbol{x}_i'\boldsymbol{M}_{\widehat{\boldsymbol{F}}}\boldsymbol{x}_i\right)^{-1} \sum_{i=1}^{N} \boldsymbol{x}_i'\boldsymbol{M}_{\widehat{\boldsymbol{F}}}\boldsymbol{y}_i \tag{2.3.10}$$

where $\widehat{\boldsymbol{F}} = \overline{\boldsymbol{Z}} = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{y}_i, \boldsymbol{x}_i)$. Westerlund et al. (2019) shows that when T is fixed and $N \to \infty$, $\boldsymbol{M}_{\widehat{\boldsymbol{F}}} \xrightarrow{p} \boldsymbol{M}_{\boldsymbol{F}} - \boldsymbol{P}_{-p}$ where $\boldsymbol{P}_{-p}$ is a nonlinear function of the model's errors. When $p = K + 1$ and the number of cross-sectional averages equals the number of factors, $\boldsymbol{P}_{-p} = \boldsymbol{0}$ and so the CCEP removes the factors and nothing else.

Another fixed-T approach comes from Ahn et al. (2013). They do not make the reduced form assumption in equation (2.3.8). Instead, they introduce new parameters which eliminate $\boldsymbol{F}$. As both $\boldsymbol{F}$ and $\boldsymbol{\gamma}_i$ are unobserved, they impose the p^2 normalizations on the factor matrix

$$\boldsymbol{F} = (\boldsymbol{\Theta}', -\boldsymbol{I}_p)' \tag{2.3.11}$$

where Θ is a $(T-p) \times p$ matrix of unrestricted parameters. Let $\boldsymbol{\theta} = \text{vec}(\boldsymbol{\Theta})$. They then define the quasi-long-differencing (QLD) matrix

$$\boldsymbol{H}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{I}_{T-p} \\ \boldsymbol{\Theta}' \end{pmatrix} \tag{2.3.12}$$

so that $\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{F} = \boldsymbol{0}$.

The Ahn et al. (2013) technique involves jointly estimating $(\boldsymbol{\beta}_0', \boldsymbol{\theta}')'$ with the use of many instruments. Instead, I focus on the QLD transformation and compare it to the asymptotic CCE transformation. Suppose $\boldsymbol{\Omega}_i = E(\boldsymbol{u}_i\boldsymbol{u}_i'|\boldsymbol{x}_i)$ is known and has full rank. Define the CCE GLS and QLD GLS estimators as

$$\widehat{\boldsymbol{\beta}}_{CCEGLS} = \left(\sum_{i=1}^{N} \boldsymbol{x}_i'\boldsymbol{M_F}(\boldsymbol{M_F}\boldsymbol{\Omega}_i\boldsymbol{M_F})^{-}\boldsymbol{M_F}\boldsymbol{x}_i\right)^{-1} \sum_{i=1}^{N} \boldsymbol{x}_i'\boldsymbol{M_F}(\boldsymbol{M_F}\boldsymbol{\Omega}_i\boldsymbol{M_F})^{-}\boldsymbol{M_F}\boldsymbol{y}_i \tag{2.3.13}$$

$$\widehat{\boldsymbol{\beta}}_{QLDGLS} = \left(\sum_{i=1}^{N} \boldsymbol{x}_i'\boldsymbol{H}(\boldsymbol{\theta})(\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{\Omega}_i\boldsymbol{H}(\boldsymbol{\theta}))^{-1}\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{x}_i\right)^{-1} \sum_{i=1}^{N} \boldsymbol{x}_i'\boldsymbol{H}(\boldsymbol{\theta})(\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{\Omega}_i\boldsymbol{H}(\boldsymbol{\theta}))^{-1}\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{y}_i \tag{2.3.14}$$

Theorem 2.3.2. *Suppose Assumption CM holds, $E(\boldsymbol{y}_i\boldsymbol{y}_i'|\boldsymbol{x}_i)$ is positive definite, and $Rank(\boldsymbol{F}) = p < T$. Then $\widehat{\boldsymbol{\beta}}_{CCEGLS} = \widehat{\boldsymbol{\beta}}_{QLDGLS}$.*

Proof. $Rank(\boldsymbol{H}(\boldsymbol{\theta})) = Rank(\boldsymbol{M_F}) = T-p$ so $\boldsymbol{M_F}(\boldsymbol{M_F}\boldsymbol{\Omega}_i\boldsymbol{M_F})^{-}\boldsymbol{M_F} = \boldsymbol{H}(\boldsymbol{\theta})(\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{\Omega}_i\boldsymbol{H}(\boldsymbol{\theta}))^{-1}\boldsymbol{H}(\boldsymbol{\theta})'$ by Theorem 1. □

Because $\boldsymbol{H}(\boldsymbol{\theta})$ and $\boldsymbol{M_F}$ are only available asymptotically, the best we can hope to achieve is an asymptotic equivalence result. Further, as discussed earlier, the CCE transformation $\boldsymbol{M}_{\widehat{\boldsymbol{F}}}$ only converges in probability to $\boldsymbol{M_F}$ when $p = K + 1$. Other fixed-T approaches in the literature include Robertson and Sarafidis (2015) who parameterize the correlation between the exogenous instruments and the factor loadings. They show that one of their estimators is asymptotically equivalent to the full QLD GMM estimator of Ahn et al. (2013) which suggests a similar efficiency result as Theorem 3. Westerlund (2020) studies the principal components (PC) estimator using the Pesaran (2006) CCE model. PC estimation is essentially fixed effects OLS which estimates the factors and loadings as additional parameters. If the estimator of $\boldsymbol{M_F}$ is consistent for $\boldsymbol{M_F}$, it can be made asymptotically efficient in the sense of Theorem 2.3.2 and thus a possible efficient alternative to CCE estimation when T is fixed.

2.3.3 Random trend

I now consider a particular factor specification which is common in applied settings. This linear model with additive effects as described in Example 1 of Section 2.2.1. takes the form

$$y_{it} = c_i + a_i t + \boldsymbol{x}_{it}\boldsymbol{\beta}_0 + u_{it} \tag{2.3.15}$$

Such a model is often called a random trend model because the outcome variable has an unobserved heterogeneous response to the observable time trend[8]. A standard technique in dealing with the heterogeneous trend is to first-difference. Define $\Delta y_{it} = y_{it} - y_{i,t-1}$ with similar definitions for $\Delta \boldsymbol{x}_{it}$ and Δu_{it}. Then

$$\Delta y_{it} = a_i + \Delta \boldsymbol{x}_{it}\boldsymbol{\beta}_0 + \Delta u_{it} \tag{2.3.16}$$

Under the strict exogeneity assumption of Assumption CM, we have $E(\Delta u_{it}|\boldsymbol{x}_i) = \boldsymbol{0}$ for each $t \geq 2$. Thus we have strictly exogenous covariates with an additive heterogeneity term. The most popular technique for estimating $\boldsymbol{\beta}_0$ in a linear model with additive heterogeneity is fixed effects estimation which applies the within transformation, $\boldsymbol{I}_{T-1} - \frac{1}{T-1}\mathbf{1}_{T-1}\mathbf{1}_{T-1}'$ where here $\mathbf{1}_{T-1}$ is a $T-1 \times 1$ vector of ones, to the first differenced residuals $\Delta y_{it} - \Delta \boldsymbol{x}_{it}\boldsymbol{\beta}_0$.

Another way to eliminate the heterogeneity in equation (2.3.15) is to apply the first-differencing transformation again on equation (2.3.16). This technique is often referred to as second-differencing. The regression is then run for $\Delta y_{it} - \Delta y_{i,t-1}$ on $\Delta \boldsymbol{x}_{it} - \Delta \boldsymbol{x}_{i,t-1}$. Since the heterogeneous terms correspond to a known intercept and time trend, we can also run a full fixed regression on equation (2.3.15) which treats $(c_1, ..., c_N, a_1, ..., a_N)$ as parameters.

One final transformation to consider is the forward orthogonal deviations (FOD) operator in Arellano and Bover (1995). This matrix applies the following transformation to the errors u_{it} in

[8] See Section 11.7.1 of Wooldridge (2010).

equation (2.2.16):

$$\frac{(T-t)}{(T-t+1)}\left(u_{it}-\frac{1}{(T-t)}(u_{i,t+1}+...+u_{iT})\right) \tag{2.3.17}$$

The transformation can be written in matrix form as

$$\text{diag}(\frac{T-1}{T},...,\frac{1}{2})^{1/2}\times\begin{pmatrix} 1 & -(T-1)^{-1} & -(T-1)^{-1} & \dots & -(T-1)^{-1} & -(T-1)^{-1} & -(T-1)^{-1} \\ 0 & 1 & -(T-2)^{-1} & \dots & -(T-2)^{-1} & -(T-2)^{-1} & -(T-2)^{-1} \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & 0 & 0 & \dots & 0 & 1 & -1 \end{pmatrix} \tag{2.3.18}$$

I denote this FOD transformation as the matrix $\boldsymbol{F}$. For each of the first $T-1$ observations, $\boldsymbol{F}$ subtracts off a weighted mean of the rest of the independent variables. While initially studied in the context of sequential exogeneity and predetermined systems like first-differencing, I study it here in the context of strict exogeneity to determine information equivalence. Since I am also assuming the structure in (2.3.16) where first-differencing has already occurred, I consider the $T-2\times T-1$ matrix $\boldsymbol{F}$ which corresponds to the definition in equation (2.3.18) but only assumes $T-1$ dependent variables instead of T. Regardless of the number of time periods considered, $\boldsymbol{F}$ has full row rank which is $T-2$ in this case.

To show information equivalence of the techniques described, let $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ be the respective $T-1\times T$ and $T-2\times T-1$ full rank first-differencing matrices, $\boldsymbol{W}=\boldsymbol{I}_{T-1}-\frac{1}{T-1}\boldsymbol{1}_{T-1}\boldsymbol{1}'_{T-1}$ be the $T-1\times T-1$ within transformation which has rank $T-2$, $\boldsymbol{F}$ be the $T-2\times T-1$ full rank matrix defined similarly to equation (2.3.18), and $\boldsymbol{M}$ be the $T\times T$ residual maker matrix from regressing on $(1,t)$. Then

$$\boldsymbol{D}_2\boldsymbol{D}_1E((\boldsymbol{y}_i-\boldsymbol{x}_i\boldsymbol{\beta}_0)|\boldsymbol{x}_i)=E(\boldsymbol{D}_2\boldsymbol{D}_1(\boldsymbol{y}_i-\boldsymbol{x}_i\boldsymbol{\beta}_0)|\boldsymbol{x}_i)=\boldsymbol{0} \tag{2.3.19}$$

$$\boldsymbol{W}\boldsymbol{D}_1E((\boldsymbol{y}_i-\boldsymbol{x}_i\boldsymbol{\beta}_0)|\boldsymbol{x}_i)=E(\boldsymbol{W}\boldsymbol{D}_1(\boldsymbol{y}_i-\boldsymbol{x}_i\boldsymbol{\beta}_0)|\boldsymbol{x}_i)=\boldsymbol{0} \tag{2.3.20}$$

$$\boldsymbol{M}E((\boldsymbol{y}_i-\boldsymbol{x}_i\boldsymbol{\beta}_0)|\boldsymbol{x}_i)=E(\boldsymbol{M}(\boldsymbol{y}_i-\boldsymbol{x}_i\boldsymbol{\beta}_0)|\boldsymbol{x}_i)=\boldsymbol{0} \tag{2.3.21}$$

$$\boldsymbol{F}\boldsymbol{D}_1E((\boldsymbol{y}_i-\boldsymbol{x}_i\boldsymbol{\beta}_0)|\boldsymbol{x}_i)=E(\boldsymbol{F}\boldsymbol{D}_1(\boldsymbol{y}_i-\boldsymbol{x}_i\boldsymbol{\beta}_0)|\boldsymbol{x}_i)=\boldsymbol{0} \tag{2.3.22}$$

where equations (2.3.19)-(2.3.22) correspond to the residuals from the second-differencing, first-differencing then within, first-differencing then forward orthogonal deviations, and full fixed effects transformations respectively. Thus each of the transformations satisfy Assumption MAT and so we can apply the general theory from Section 2.2.2.

Theorem 2.3.3. *Suppose Assumption CM holds and $E(\boldsymbol{y}_i\boldsymbol{y}_i'|\boldsymbol{x}_i)$ is positive definite. Then $\boldsymbol{D}_2\boldsymbol{D}_1$, $\boldsymbol{W}\boldsymbol{D}_1$, $\boldsymbol{F}\boldsymbol{D}_1$ and $\boldsymbol{M}$ are information equivalent.*

Proof. As $\boldsymbol{D}_1$ is full rank, $Rank(\boldsymbol{D}_2\boldsymbol{D}_1) = Rank(\boldsymbol{W}\boldsymbol{D}_1) = Rank(\boldsymbol{F}\boldsymbol{D}_1) = T-2$. Since $Rank(\boldsymbol{M}) = T-2$ by definition, the result holds by Theorem 1. □

The simplicity of the proof follows from the general nature of the unified theory proved in Section 2.2 and thus demonstrates its usefulness. In the language of Im et al. (1999), the GLS estimators based off of the residuals in equations (2.3.19)-(2.3.22) for a given $E(\boldsymbol{u}_i\boldsymbol{u}_i'|\boldsymbol{X}_i)$ are algebraically equivalent for a given covariance matrix. Finally, Theorem 2.3.3 can be seen as a generalization of Theorem 4.3 of Im et al. (1999).

2.4 Practical considerations

The final section of the paper provides useful applications of the results in the previous two sections. I first consider implementation of the efficiency bounds discussed in the paper. Given a transformation $\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta})$ satisfying Assumptions SYS and ORTH (and thus MAT), I describe the efficient estimator. The estimator $\widehat{\boldsymbol{\beta}}_A$ which solves

$$\sum_{i=1}^{N} \nabla_{\boldsymbol{\beta}}\boldsymbol{m}_i(\boldsymbol{\beta}_0)'\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)'(\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)E(\boldsymbol{y}_i\boldsymbol{y}_i'|\boldsymbol{x}_i)\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)')^{-}\boldsymbol{A}(\boldsymbol{x}_i, \widehat{\boldsymbol{\beta}}_A)\boldsymbol{y}_i = \boldsymbol{0} \tag{2.4.1}$$

is $\sqrt{N}$-asymptotically normal with asymptotic variance equal to the efficiency bound given by equation (2.2.4).

First-stage estimation of $\boldsymbol{\beta}_0$ comes from a GMM estimator with an arbitrary weight matrix. Second, one needs to consistently estimate $E(\boldsymbol{y}_i\boldsymbol{y}_i'|\boldsymbol{x}_i)$. A nonparametric regression estimator can be used in principle, but in practice this estimator may give highly imprecise estimates when T

and K are relatively large. In the multiplicative heterogeneity setting, Brown and Wooldridge (2021) provides a simple and attractive parametric framework for the FEP setting. They assume $Var(y_{it}|\boldsymbol{x}_i, c_i) = \alpha E(y_{it}|\boldsymbol{x}_i, c_i)$ where $\alpha > 0$ is an identified coefficient along with a constant conditional correlation matrix.

Asymptotically justified standard errors can be derived using the familiar sample analog to the efficiency bound in (2.2.4). The researcher can then test the validity of parts of Assumption CM. For strict exogeneity, Wooldridge (2010, Chapter 18) suggests including functions of lead values of independent variables and running a joint test of significance. This method's most attractive feature is the weakness of its alternative hypothesis. The null maintains strict exogeneity while the alternative is merely that strict exogeneity fails. It is also easy to implement and can be tested in most standard statistical packages. However, there is no guidance on how to choose which regressors to include or their functional forms.

Another possible way to examine strict exogeneity is via a Hausman test. The researcher could choose a competing estimator based on the desired alternative hypothesis. In the nonlinear multiplicative example of Section 2.3.1, suppose the researcher believes that sequential exogeneity holds, or that $E(y_{it}|\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{it}, c_i) = m_t(\boldsymbol{x}_{it}, \boldsymbol{\beta}_0)$. Then the generalized next-differencing transformation $\boldsymbol{D}_i(\boldsymbol{\beta}) = (\boldsymbol{r}_{i,1,2}(\boldsymbol{\beta}), ..., \boldsymbol{r}_{i,T-1,T}(\boldsymbol{\beta}))'$ still provides valid moment conditions. However, the instruments are designed to reach the efficiency bound in (2.2.4) will not be valid under sequential exogeneity alone. Chamberlain (1992b) derives the asymptotic efficiency bound for moment conditions under sequential exogeneity and provides an implementable estimator which reaches said bound. Under the null hypothesis, both estimators are consistent with the generalized next-differencing estimator as in (2.3.5) being asymptotically efficient. Under the alternative, only Chamberlain's instruments are valid (and in fact asymptotically efficient among $\sqrt{N}$-asymptotically normal estimators). Thus we can use a Hausman statistic to test the assumption of strict exogeneity.

The Chamberlain estimator described in the Hausman statistic procedure is difficult to implement as the instruments may be comprised of multiple sums of conditional moments. The researcher will need to either greatly strengthen the assumptions of the model to allow for parametric forms of these moments or utilize a large number of nonparametric regressions. Either way, this computational burden makes the Chamberlain estimator difficult to implement.

Another possible application of the results involve finite-sample and computational concerns. Phillips (2020) demonstrates that matrix inversion for estimators based on first-differencing can involve significantly more computational resources than those based on forward orthogonal deviations. He demonstrates with simulation evidence that computational time increases quickly with T even for relatively small values of N. While instruments need to satisfy two conditions given in Phillips (2020) which are not necessarily assumed here, I reiterate that the results in Section 2.2 are purely algebraic and can be applied in a large number of settings.

2.5 Conclusion

This paper considers linear transformations of nonlinear panel models with unobserved heterogeneity. When covariates are strictly exogenous in the zero conditional mean sense, such transformations provide uncountable moment conditions exploitable for estimation. I consider specifically the asymptotic efficiency bound for estimating the model's parameter which is reached by the optimal choice of instruments. This matrix specifies a lower bound on how efficient any $\sqrt{N}$-asymptotically normal estimator of $\boldsymbol{\beta}_0$ can possibly be.

Transformations of the data are said to be information equivalent if they yield the same asymptotic efficiency bound. The main result of Section 2.2 is a unified framework for evaluating the efficiency bounds of transformations that provide moment conditions for estimation. It shows that besides regularity conditions, matrix transformations which yield conditional moment restrictions and have the same rank yield the same information bound. I also simplify the form of the efficiency bound under a general and easily verifiable algebraic orthogonality property which could potentially help in determining other interesting relationships between instrumental variable estimators.

The general framework is applied to show that the generalized within transformation, which provides the basis of the FEP estimator, is in fact information equivalent to a number of other transformations. These transformations, which include generalizations of varying differencing techniques used in the linear panel data context such as next-, first-, and long-differencing, as well as the residual maker matrix from regression on the outcome variable's mean function, are only required to satisfy a rank condition for the main theorem to hold. It is also shown that any $T-1 \times T$ matrix which is algebraically orthogonal to the mean function of the outcome and of full rank is information equivalent, which includes deleting any arbitrary row from the generalized within transformation and removing the linear redundancy does not lose any information.

I also generalize a result of Im et al. (1999) on linear panels with an additive heterogeneity term to a general factor-augmented error structure as studied in Pesaran (2006), Ahn et al. (2013), and Westerlund (2020). I show that any transformation of the data which is full rank and eliminates the factors is information equivalent. I use this result to show that in the case of a random heterogeneous trend model, first-differencing twice, first-differencing and then using a within transformation, and the true fixed effects estimator are information equivalent. For arbitrary factor structures, the QLD transformation of Ahn et al. (2013) is information equivalent to the infeasible fixed effects GLS estimator which takes the unobserved effects as known.

The work in this paper provides a basic framework for comparison of parametric estimators for a broad class of nonlinear models. I primarily consider strictly exogenous covariates so I could compare estimators using theoretically efficient instruments. However, the finite sample algebraic results hold regardless of validity of the instruments. As such, the main theorem in Section 2.2 can apply to any comparison of efficiency for instrumental variable estimators.

CHAPTER 3

MOMENT-BASED ESTIMATION OF LINEAR PANEL DATA MODELS WITH FACTOR-AUGMENTED ERRORS

3.1 Introduction

The prevalence of panel data in modern economics has led theorists and practitioners to pay more attention to unobserved and interactive heterogeneity. A popular representation of unobserved effects is the linear factor structure $\sum_{j=1}^{p} f_{tj}\gamma_{ji}$ where f_{tj} is a time-varying macro effect or "common factor" and γ_{ji} is an individually heterogeneous response or "factor loading". In studying the statistical properties of estimators of factor models, most theoretical treatments have relied on asymptotic expansions where the number of time periods T grows large with the number of cross-sectional units N. As the vast majority of microeconometric data sets have only a few time periods, the recent literature assumes T is fixed while N goes to infinity.

One of the most popular approaches is the common correlated effects (CCE) estimator of Pesaran (2006). He assumes that the covariates are a linear function of the common factors plus a matrix of independent idiosyncratic errors. The pooled CCE estimator comes from the OLS regression which estimates unit-specific slopes on the cross-sectional averages of the dependent and independent variables. CCE is similar to a fixed effects treatment which seeks to eliminate the factors and remove a source of both endogeneity and cross-sectional dependence. Consistency and asymptotic normality was originally proved for sequences of N and T going to infinity.

Recent work extends the CCE framework to a fixed-T setting. De Vos and Everaert (2021) derive a fixed-T consistency correction for the dynamic CCE estimator but requires $T \to \infty$ for asymptotic normality. Westerlund et al. (2019) provide the first asymptotic normality derivation of pooled CCE when T is fixed and $N \to \infty$. However, they still maintain stringent assumptions on the model's DGP. For example, they assume that the factor loadings are independent of the idiosyncratic errors. My estimators do not require this assumption for consistency, though making it simplifies the standard errors. Further, the CCE estimator generally uses more factor proxies than necessary which can lead to inefficiency. Finally, the CCE estimator requires $T > K + 1$ which is highly restrictive in microeconometric settings. For example, an intervention analysis with only pre-treatment, treatment, and post-treatment observations, classical CCE would require the treatment indicator to be the only regressor.

Aside from CCE, most existing fixed-T techniques create moment conditions by including additional parameters to estimate or by eliminating the factors with observed proxies. A few examples include Hayakawa (2012), Ahn et al. (2001, 2013), Robertson and Sarafidis (2015), and Juodis and Sarafidis (2018, 2020)[1]. Of these approaches, I focus on Ahn et al. (2013), who define a parameterized quasi-long-differencing (QLD) transformation that eliminates the factor structure. The QLD residuals then form the basis for a GMM estimator which uses all available exogenous variables to generate moment conditions. I focus on the QLD technique for the sake of comparison to CCE as both approaches eliminate the factor structure and allow for “fixed effects" assumptions. For example, Robertson and Sarafidis (2015) parameterize the correlation between the exogenous variables and the factor loadings. Ahn (2015) points out that if the factor loadings’ distributions change over the cross-sectional units, identification in Robertson and Sarafidis (2015) does not hold.

Ahn et al. (2013) do not assume a pure factor structure in the covariates like Pesaran (2006) and leaves the distribution of the covariates unspecified. However, the generality of Ahn et al. (2013) comes at the cost of identifying assumptions, which may explain its lack of use in the empirical literature. The QLD GMM estimator requires many moments to identify all the model’s parameters. If either T or the number of factors is large, their GMM estimator may require outside instruments. Their estimator also requires nonlinear optimization with a large number of moments and parameters. Hayakawa (2016) provides a simple example where the global identifying assumptions fail and there exist local stationary points.

[1] Juodis and Sarafidis (2021) allows for a linear estimator which requires no additional parameters, However, the fixed-T analysis requires strong assumptions on the loadings which this paper avoids. See Assumption S.1.1(d) in their Appendix.

I synthesize both approaches and weaken both the Pesaran (2006) and Ahn et al. (2013) assumptions. I use a weakened CCE model without any independence assumptions to provide a first-stage estimator of the additional QLD parameters. Using the QLD transformation, I then derive pooled and mean group linear estimators and provide standard errors which are valid even when the heterogeneity is correlated with the model's errors. These novel estimators have desirable rank conditions and do not require outside instruments like in Ahn et al. (2013). They also do not restrict the number of covariates to be less than the number of time periods minus one, an improvement over fixed-T CCE. Simulations suggest that the linear QLD estimators outperform the CCE and QLD GMM estimators in finite samples.

Another potential source of heterogeneity in linear models comes from the slope coefficients on the observed variables of interest. Pesaran (2006) proves fixed-T consistency of the mean group CCE estimator under random slopes but assumes they are independent of everything else in the model. Asymptotic normality requires $T \to \infty$ and pooled CCE is studied under constant slopes. I prove fixed-T consistency and asymptotic normality of the new pooled and mean group QLD estimators. I show that the first-stage estimation of the QLD parameters does not affect consistency, which mirrors the pooled OLS result of Wooldridge (2005), who assumes known factors. To the best of my knowledge, this paper is the first to consider arbitrary random slopes in the context of fixed-T panels with factor-driven endogeneity.

The rest of the paper is structured as follows: Section 3.2 describes the main model of interest which is weaker than that in Westerlund et al. (2019). Section 3.3 provides the assumptions which underlie the model and discusses implementation of the QLD-based estimators. Section 3.4 introduces random slopes. Section 3.5 provides simulation evidence for the finite sample properties of the QLD estimators. Section 3.6 compares the pooled QLD estimator to two-way fixed effects (TWFE) and CCE in estimating the effect of education expenditure on standardized test performance using a school district-level data set from the state of Michigan. Section 3.7 concludes with a brief summary and suggestions for future research.

3.2 Model

This section lays out the models considered in Westerlund et al. (2019) and Ahn et al. (2013), the fixed-T CCE and QLD approaches respectively. Throughout the paper, the equation of interest is

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta_0} + \boldsymbol{F}_0\boldsymbol{\gamma}_i + \boldsymbol{u}_i \tag{3.2.1}$$

where $\boldsymbol{y}_i$ is a $T \times 1$ vector of outcomes, $\boldsymbol{X}_i$ is $T \times K$ matrix of covariates, $\boldsymbol{F}_0$ is a $T \times p_0$ matrix of factors common to all units in the population, $\boldsymbol{\gamma}_i$ is a $p_0 \times 1$ vector of factor loadings, $\boldsymbol{u}_i$ is a $T \times 1$ vector of idiosyncratic shocks. A '0' subscript denotes the true or realized value of an unobserved parameter. p_0 is then unobserved because $\boldsymbol{F}_0$ and $\boldsymbol{\gamma}_i$ are unobserved. Later, p denotes the number of factors specified by the econometrician. $\boldsymbol{\beta}_0$ is the object of interest and the factor structure $\boldsymbol{F}_0\boldsymbol{\gamma}_i$ is treated as a collection of nuisance parameters.

This paper defines p_0 as the number of factors whose loadings correlate with $\boldsymbol{X}_i$. This interpretation is similar to Ahn et al. (2013) and implicit to the CCE model as discussed in the following section. One justification of this interpretation is to write the full error as $\boldsymbol{D}_0\boldsymbol{\rho}_i + \boldsymbol{\epsilon}_i$ where $\boldsymbol{D}_0$ is a possibly infinite dimensional matrix of common factors and $\boldsymbol{\epsilon}_i$ is a vector of idiosyncratic errors. Then $\boldsymbol{F}_0\boldsymbol{\gamma}_i$ is the set of variables from $\boldsymbol{D}_0\boldsymbol{\rho}_i$ which are correlated with $\boldsymbol{X}_i$ and the rest are absorbed into the error. However, it is entirely likely that $\boldsymbol{\gamma}_i$ is correlated with the other loadings which are uncorrelated with $\boldsymbol{X}_i$. This correlation can cause problems for inference and is addressed in Section 3.3.

Finally, I assume the factors in $\boldsymbol{F}_0$ are constant for the purpose of asymptotic analysis. The alternative setting is to assume the factors are stochastic and independent of the other terms, or make the modeling assumptions conditional on the sigma-algebra generated by the factors like in Ahn et al. (2013). When T is fixed, the stochastic nature of the factors is less relevant for the asymptotic arguments. Standard errors do not change as properly studentized test statistics converge to their usual distributions[2]. As such, I consider the standard microeconometric assumption of random

[2] See Section 6 of Andrews (2005).

sampling in the cross-section. Hsiao (2018) provides examples of papers which make either the fixed or random assumption on the factors.

3.2.1 Common Correlated Effects

The CCE model in Pesaran (2006) and Westerlund et al. (2019) adds an additional reduced form equation which represents the relationship between the covariates and the factor structure:

$$\boldsymbol{X}_i = \boldsymbol{F}_0 \boldsymbol{\Gamma}_i + \boldsymbol{V}_i \tag{3.2.2}$$

where $\boldsymbol{\Gamma}_i$ is a $p_0 \times K$ matrix of factor loadings and $\boldsymbol{V}_i$ is a $T \times K$ matrix of idiosyncratic errors. Westerlund et al. (2019) follows Pesaran (2006) in assuming $\boldsymbol{V}_i$, $\boldsymbol{\Gamma}_i$, $\boldsymbol{\gamma}_i$, and $\boldsymbol{u}_i$ are mutually independent[3]. Assuming that the idiosyncratic errors have mean zero, CCE estimates the factors with the matrix $\widehat{\boldsymbol{F}} = (\overline{\boldsymbol{y}}, \overline{\boldsymbol{X}})$ where $(\overline{\boldsymbol{y}}, \overline{\boldsymbol{X}}) = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{y}_i, \boldsymbol{X}_i)$ are the cross-sectional averages of $\boldsymbol{y}_i$ and $\boldsymbol{X}_i$.

The **pooled common correlated effects (CCEP)** estimator treats the cross-sectional averages as fixed effects and can be represented as

$$\widehat{\boldsymbol{\beta}}_{CCEP} = \left(\sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{M}_{\widehat{\boldsymbol{F}}} \boldsymbol{X}_i\right)^{-1} \sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{M}_{\widehat{\boldsymbol{F}}} \boldsymbol{y}_i \tag{3.2.3}$$

where $\boldsymbol{M}_{\widehat{\boldsymbol{F}}} = \boldsymbol{I}_T - \widehat{\boldsymbol{F}}(\widehat{\boldsymbol{F}}'\widehat{\boldsymbol{F}})^{+}\widehat{\boldsymbol{F}}'$. Here $'+'$ denotes a Moore-Penrose inverse which can be replaced by a proper inverse in samples where $\widehat{\boldsymbol{F}}'\widehat{\boldsymbol{F}}$ has full rank. Pesaran (2006) derives the CCE estimator under the following intuition: first, write $\boldsymbol{Z}_i = (\boldsymbol{y}_i, \boldsymbol{X}_i)$. The two models in equations (3.2.1) and (3.2.2) imply

$$E(\boldsymbol{Z}_i) = \boldsymbol{F}_0 E(\boldsymbol{C}_i) \boldsymbol{Q} \tag{3.2.4}$$

where $\boldsymbol{C}_i = (\boldsymbol{\gamma}_i, \boldsymbol{\Gamma}_i)$ and

$$\boldsymbol{Q} = \begin{pmatrix} 1 & \boldsymbol{0}_{1\times K} \\ \boldsymbol{\beta}_0 & \boldsymbol{I}_K \end{pmatrix}$$

[3] Westerlund et al. (2019) assume the loadings come from a fixed series of constant matrices which is more general than the Pesaran (2006) assumption that the loadings are iid.

Thus, $\boldsymbol{M}_{\widehat{\boldsymbol{F}}}$ asymptotically eliminates the space spanned by $\boldsymbol{F}_0$ which includes $\boldsymbol{F}_0\boldsymbol{\gamma}_i$.

Westerlund et al. (2019) show that $\boldsymbol{M}_{\widehat{\boldsymbol{F}}}$ generally converges to the space orthogonal to both $\boldsymbol{F}$ and a random term which is a function of the model's idiosyncratic errors. For the sake of simplicity, suppose that $\boldsymbol{M}_{\widehat{\boldsymbol{F}}} \xrightarrow{p} \boldsymbol{M}_{\boldsymbol{F}_0}$ which is the case when $p_0 = K + 1$. Then the pooled CCE estimator is based off of the moment conditions

$$E(\boldsymbol{X}_i' \boldsymbol{M}_{\boldsymbol{F}_0}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})) = \boldsymbol{0}$$

Assuming $E(\boldsymbol{V}_i) = \boldsymbol{0}$ as in Pesaran (2006) and Westerlund et al. (2019), the reduced form portion of the CCE model also implies $E(\boldsymbol{M}_{\boldsymbol{F}_0}\boldsymbol{X}_i) = \boldsymbol{0}$. Since the CCE approach estimates no parameters in this set of moments, the additional moments are unused by the CCE residual above. I show how these reduced form moments can be exploited for additional information in Section 3.3.

A particularly harsh restriction of the pooled CCE estimator is the rank condition required for the denominator. $\boldsymbol{M}_{\widehat{\boldsymbol{F}}}$ is a residual-maker matrix and so it has rank $T - (K + 1)$. For the estimator to be well-defined, we require $T > K + 1$. This constraint is trivially nonbinding when $T \to \infty$ like in the prior literature. However, when T is fixed like in this paper, we need $K < T - 1$. For example, if we only observe three time periods, we can only incorporate one regressor. Also, when $K + 1 > p_0$, the CCE estimator unnecessarily removes variation from the data which could improve precision of the estimator. I address both of these problems in Section 3.2.2.

3.2.2 Quasi-long-differencing

Ahn et al. (2013) do not assume the pure factor structure in $\boldsymbol{X}_i$. They start with equation (3.2.1) then parameterize the factors for the purpose of eliminating them. Before discussing how this process works, I introduce the 'rotation problem', a well-known issue in the factor literature. Since both $\boldsymbol{F}$ and $\boldsymbol{\gamma}_i$ are unobservable, they cannot be separately identified. To see why, consider any nonsingular $p \times p$ matrix $\boldsymbol{A}$. Then $\boldsymbol{F}_0\boldsymbol{\Gamma}_i = \boldsymbol{F}^*\boldsymbol{\Gamma}_i^*$ where $\boldsymbol{F}^* = \boldsymbol{F}_0\boldsymbol{A}$ and $\boldsymbol{\Gamma}_i^* = \boldsymbol{A}^{-1}\boldsymbol{\Gamma}_i$. We can only hope to identify the factors up to an arbitrary rotation of their linear subspace. Ahn et al. (2013)

suggest the following p_0^2 normalizations based off of a row-reduction rotation:

$$\boldsymbol{F}_0 = (\boldsymbol{\Theta}_0', -\boldsymbol{I}_{p_0})' \tag{3.2.5}$$

where $\boldsymbol{\Theta}_0$ is a $(T - p_0) \times p_0$ matrix of unrestricted parameters. The given normalization is irrelevant because I am not interested in estimating $\boldsymbol{F}_0$. In this case, I only assume that the factors are full rank; the normalization chosen merely reflects this fact.

Given the normalization of the general factor matrix $\boldsymbol{F}_0$ in equation (3.2.5), Ahn et al. (2013) define the **quasi-long-differencing (QLD)** matrix

$$\boldsymbol{H}(\boldsymbol{\theta}_0) = \begin{pmatrix} \boldsymbol{I}_{T-p_0} \\ \boldsymbol{\Theta}_0' \end{pmatrix} \tag{3.2.6}$$

where $\boldsymbol{\theta}_0 = \text{vec}(\boldsymbol{\Theta}_0)$. The QLD transformation eliminates the factors for any given $\boldsymbol{\theta}_0$: $\boldsymbol{H}(\boldsymbol{\theta}_0)'\boldsymbol{F}(\boldsymbol{\theta}_0) = \boldsymbol{0}$. This differencing technique allows for the construction of the QLD residual studied in Ahn et al. (2013):

$$E(\boldsymbol{w}_i \otimes \boldsymbol{H}(\boldsymbol{\theta}_0)'(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_0)) = \boldsymbol{0} \tag{3.2.7}$$

where $\boldsymbol{w}_i$ is a vector of instruments which may contain $\text{vec}(\boldsymbol{X}_i)$. The normalization in (3.2.5) and implicit in (3.2.6) is only one particular choice of rotation. The Ahn et al. (2013) estimator depends on the choice of normalization which is unaddressed in the original paper. I discuss this issue in the Appendix and provide potential solutions for the estimators derived in Section 3.2.

While Ahn et al. (2013) provide a general framework for estimating $\boldsymbol{\beta}_0$ without strong restrictions on the distribution of $\boldsymbol{X}_i$, it requires at least $p_0 + K/(T - p_0)$ instruments in $\boldsymbol{w}_i$ to identify all of the model's parameters. If some of the variables are not exogenous in each time period like with weakly exogenous or predetermined variables, or if p_0 is large, we may require outside instruments. Hayakawa (2016) demonstrates an example where the objective function based off of equation (3.2.7) suffers from non-global stationary points due to the nonlinear nature of estimation with a large number of moments and parameters.

The pure factor structure in equation (3.2.4) can thus be used for estimating the parameters in equation (3.2.6). If we assume $\boldsymbol{X}_i = \boldsymbol{F}_0\boldsymbol{\Gamma}_i + \boldsymbol{V}_i$ where $E(\boldsymbol{V}_i) = \boldsymbol{0}$, then

$$E(\boldsymbol{H}(\boldsymbol{\theta}_0)'\boldsymbol{Z}_i) = \boldsymbol{0} \tag{3.2.8}$$

and $\boldsymbol{\theta}_0$ is identified by equation (3.2.8) which substantially reduces the number of moments needed to identify $\boldsymbol{\beta}_0$. I also show explicitly in the following section how and when these additional moments are useful for the purpose of identification and efficiency.

3.3 Estimation

I now state this paper's primary assumptions. The first assumption is similar to the 'Basic Assumptions' of Ahn et al. (2013) and is made for the sake of comparison to their approach. The second set specifies the pure factor structure in $\boldsymbol{X}_i$ similar to Westerlund et al. (2019). I specify the models in the assumptions as the main results of the paper depend on which model is being assumed. Conditional moments hold almost surely.

Assumption 1 (Linear population model):

1. $\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta}_0 + \boldsymbol{F}_0\boldsymbol{\gamma}_i + \boldsymbol{u}_i$.

■

Assumption 2 (CCE reduced form equations):

1. $\boldsymbol{X}_i = \boldsymbol{F}_0\boldsymbol{\Gamma}_i + \boldsymbol{V}_i$.

2. $(\boldsymbol{\gamma}_i, \boldsymbol{\Gamma}_i, \boldsymbol{V}_i, \boldsymbol{u}_i)$ are independent and identically distributed across i with finite fourth moments.

3. $E(\boldsymbol{V}_i) = \boldsymbol{0}$ and $E(\boldsymbol{u}_i|\boldsymbol{V}_i) = \boldsymbol{0}$.

4. $\text{Rk}(\boldsymbol{F}_0) = p_0$ and $\text{Rk}(E([\boldsymbol{\gamma}_i, \boldsymbol{\Gamma}_i])) = p_0 \leq K + 1$.

■

Assumption 1 simply defines the relevant population model. I will not require the strong rank conditions of Ahn et al. (2013) which can be found in the Appendix, nor will I require outside instruments. Assumption 2 specifies the pure factor assumption similar to Pesaran (2006) and Westerlund et al. (2019). I assume random sampling in the cross-section to simplify the asymptotic analysis, though this restriction is unnecessary.

Westerlund et al. (2019) follow the classical CCE approach in assuming independence between all stochastic components of the model which is unrealistic in microeconometric settings. Further, the asymptotic normality derivation in Westerlund et al. (2019) relies on the assumption that $\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\gamma}_i' \otimes \boldsymbol{V}_i' = O_p(N^{-1/2})$. I demonstrate in Section 3.3.2 that it is unnecessary for consistency and asymptotic normality, and how misspecification causes inconsistency in the standard errors and bootstrapped test statistics provided in Westerlund et al. (2019). The factor structure allows us to weaken the Ahn et al. (2013) assumption from $E(\boldsymbol{u}_i|\boldsymbol{X}_i) = \boldsymbol{0}$ to $E(\boldsymbol{u}_i|\boldsymbol{V}_i) = \boldsymbol{0}$. Finally, I do not assume the reduced form equation is a conditional mean specification like Westerlund et al. (2019). They assume $E(\boldsymbol{V}_i|\boldsymbol{\Gamma}_i) = \boldsymbol{0}$, where I only need $E(\boldsymbol{V}_i) = \boldsymbol{0}$ and place no restrictions on $D(\boldsymbol{V}_i|\boldsymbol{u}_i)$.

Another way in which QLD can help weaken the CCE model is the relevant order conditions. As described earlier, Westerlund et al. (2019) require $T > K + 1$ for CCE estimation but I will directly use the moments $E(\boldsymbol{H}_0'\boldsymbol{Z}_i) = \boldsymbol{0}$ to remove the factors which only requires $K \geq p_0 + 1$, a restriction also made by Pesaran (2006) and Westerlund et al. (2019). Ahn et al. (2013) does not require this condition but assumes the existence of outside instruments which may be infeasible given the application. I also discuss in Section 3.3.2 how to include known factors like a heterogeneous intercept which decreases the number of relevant factors and makes the assumption even less restrictive.

3.3.1 CCE Moment Conditions

I now look at the moment conditions implied by Assumption 2. Equation (3.2.8) of Section 3.2, $E(\boldsymbol{H}_0'\boldsymbol{Z}_i) = \boldsymbol{0}$ where $\boldsymbol{Z}_i = (\boldsymbol{y}_i, \boldsymbol{X}_i)$, implies that Assumption 2 provides information on $\boldsymbol{\theta}_0$ which leads to more efficient estimation of $\boldsymbol{\beta}_0$ and provides a first-stage estimator which negates the need for the full joint estimator of Ahn et al. (2013). I first consider identification of $\boldsymbol{\theta}_0$ from the pure factor structure alone to show that it in fact yields valid moments. As in Ahn et al. (2013), identification hinges on correctly specifying $p = p_0$ where p is the number of factors specified by the econometrician.

Lemma 3.3.1. *Under Assumption 2, $\boldsymbol{\theta}_0$ is identified by $E(\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{Z}_i) = \mathbf{0}$ if and only if $p = p_0$.*

Proof. Assumption 2(3) implies

$$E(\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{Z}_i) = \boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{F}_0 E(\boldsymbol{C}_i)\boldsymbol{Q} \tag{3.3.1}$$

where $\boldsymbol{E}(C_i) = E([\boldsymbol{\gamma}_i, \boldsymbol{\Gamma}_i])$ and $\boldsymbol{Q}$ is given in Section 3.2.1. $\boldsymbol{Q}$ is nonsingular and $E(\boldsymbol{C}_i)$ has full row rank by Assumption 2(4), so equation (3.3.1) is zero if and only if $\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{F}_0 = \mathbf{0}$. When $p = p_0$, $H(\boldsymbol{\theta})'\boldsymbol{F}_0 = \boldsymbol{\Theta}_0 - \boldsymbol{\Theta}$ which is zero if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. See the Appendix for the $p \neq p_0$ cases. □

Remark (Misspecification): A possible reason for the lack of use of CCE estimation among microeconomists is the model in Assumption 2(1). This assumption is in fact not strictly necessary for identifying $\boldsymbol{\theta}_0$. Consider the following linear projection:

$$E(\boldsymbol{Z}_i) = \boldsymbol{F}_0\boldsymbol{G} + \boldsymbol{E}$$

where $\boldsymbol{F}_0'\boldsymbol{E} = \mathbf{0}$. Then $\boldsymbol{\theta}_0$ is still identified by the moments $E(\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{Z}_i)$ if $\boldsymbol{G}$ has full rank. ■

We can use Lemma 3.3.1 to provide an estimator of $\boldsymbol{\theta}_0$ based off of the covariates alone. Let $\widehat{\boldsymbol{H}} = \boldsymbol{H}(\widehat{\boldsymbol{\theta}})$, $\boldsymbol{D}_{\boldsymbol{\theta}} = E(\nabla_{\boldsymbol{\theta}}\text{vec}(\boldsymbol{H}_0'\boldsymbol{X}_i))$, and $\boldsymbol{A}_{\boldsymbol{\theta}} = E(\text{vec}(\boldsymbol{H}_0'\boldsymbol{Z}_i)\text{vec}(\boldsymbol{H}_0'\boldsymbol{Z}_i)')$.

Theorem 3.3.1. *Suppose Assumption 2 holds, and let $\widehat{\boldsymbol{\theta}}$ be the GMM estimator based off of $E(vec(\boldsymbol{H}_0'\boldsymbol{Z}_i)) = \mathbf{0}$ using a consistent estimator of the optimal weight matrix. Then*

1. $\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, (\boldsymbol{D}_{\boldsymbol{\theta}}'\boldsymbol{A}_{\boldsymbol{\theta}}^{-1}\boldsymbol{D}_{\boldsymbol{\theta}})^{-1})$.

Now suppose that $\widehat{\boldsymbol{A}}_{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{A}_{\boldsymbol{\theta}}$ using first-step estimator $\widehat{\boldsymbol{\theta}}$.

1. *If $p_0 = p$ then $N^{-1}\left(\sum_{i=1}^N vec(\widehat{\boldsymbol{H}}'\boldsymbol{Z}_i)\right)'\widehat{\boldsymbol{A}}_{\boldsymbol{\theta}}^{-1}\left(\sum_{i=1}^N vec(\widehat{\boldsymbol{H}}'\boldsymbol{Z}_i)\right) \xrightarrow{d} \chi^2((T - p_0)(K + 1 - p_0))$.*

2. *If $p_0 > p$, then $N^{-1}\left(\sum_{i=1}^N vec(\widehat{\boldsymbol{H}}'\boldsymbol{Z}_i)\right)'\widehat{\boldsymbol{A}}_{\boldsymbol{\theta}}^{-1}\left(\sum_{i=1}^N vec(\widehat{\boldsymbol{H}}'\boldsymbol{Z}_i)\right) \xrightarrow{p} \infty$.*

Proof. The proof comes from standard theory; see Hansen (1982). The estimator of the optimal weight matrix is $\widehat{\boldsymbol{A}}_{\boldsymbol{\theta}} = \frac{1}{N}\sum_{i=1}^{N} \text{vec}(\boldsymbol{H}(\tilde{\boldsymbol{\theta}})'\boldsymbol{Z}_i)\text{vec}(\boldsymbol{H}(\tilde{\boldsymbol{\theta}})'\boldsymbol{Z}_i)'$ where $\tilde{\boldsymbol{\theta}}$ is a consistent first-stage estimator of $\boldsymbol{\theta}_0$. □

It is entirely possible there are variables in the data set which are linear in the factors but not relevant for estimation. In this case, one can simply use them to estimate $\boldsymbol{\theta}_0$ but drop them from the estimating equation. Further, if relevant variables are not linear in $\boldsymbol{F}_0$, they should be dropped from the estimation in Theorem 3.3.1. This can occur if there are polynomial or interactive functions of the covariates in the estimating equation. De Vos and Westerlund (2019) study this case in the context of CCE.

I also note that the just identified case $p_0 = K + 1$ corresponds to a simple M-estimator:

Corollary 3.3.1. *When $p_0 = K + 1$, the estimator $\widehat{\boldsymbol{\theta}}$ solves*

$$\widehat{\boldsymbol{H}}'(\overline{\boldsymbol{y}}, \overline{\boldsymbol{X}}) = \boldsymbol{0}$$

Corollary 3.3.1 provides important robustness properties in Section 3.3. For now, I point out how Theorem 3.3.1 can help test for p_0. There are $(T - p_0)(K + 1)$ moments and $(T - p_0)p_0$ parameters, so the system is underidentified when $K + 1 < p_0$ and just identified like in Corollary 3.3.1 when $K + 1 = p_0$. When $K + 1 > p_0$, we have overidentifying restrictions to test for p_0. Ahn et al. (2013) recommend testing for p_0 by first setting $p = 0$ and setting $\boldsymbol{H} = \boldsymbol{I}_T$. If the hypothesis is rejected using the statistic in part (2) of Theorem 3.3.1, move to $p = 1$. Continue until the null hypothesis cannot be rejected. I refer the reader to Section 3 of Ahn et al. (2013) for additional details and tests. I follow a similar approach to testing based off of the moments in Theorem 3.3.1.

I now demonstrate that the additional moments generally improve efficiency of the Ahn et al. (2013) GMM estimator by demonstrating that the CCE model's reduced form assumption implies additional non-redundant moment conditions. The following theorem completely characterizes when the moments $E(\boldsymbol{H}_0'\boldsymbol{X}_i) = E(\boldsymbol{H}_0'\boldsymbol{V}_i) = \boldsymbol{0}$ are partially redundant for estimating $\boldsymbol{\beta}_0$ using the Ahn et al. (2013) estimator, meaning its asymptotic variance is the same with or without the additional moments. I do not include $E(\boldsymbol{H}_0'\boldsymbol{y}_i) = \boldsymbol{0}$ because the efficiency result would require additional assumptions on $Var(\boldsymbol{u}_i)$. Let $\boldsymbol{g}_{i1}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \text{vec}(\boldsymbol{X}_i) \otimes \boldsymbol{H}(\boldsymbol{\theta})'(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})$ and $\boldsymbol{g}_{i2}(\boldsymbol{\theta}) = \boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{V}_i$ be the residuals associated with the moment conditions from equations (3.2.7) and (3.2.8) respectively. Let $\boldsymbol{D}_{11} = E(\nabla_{\boldsymbol{\beta}}\boldsymbol{g}_{i1}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0))$, $\boldsymbol{D}_{12} = E(\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_{i1}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0))$, and $\boldsymbol{\Omega}_{11} = Var(\boldsymbol{g}_{i1}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0))$.

Theorem 3.3.2. *Given Assumptions 1 and 2, suppose $E(\boldsymbol{u}_i|\boldsymbol{X}_i)$ and the Identifying Assumptions in the Appendix hold. Then the moment conditions $E(\boldsymbol{g}_{i2}(\boldsymbol{\theta}_0)) = \boldsymbol{0}$ are partially redundant for estimating $\boldsymbol{\beta}_0$ if and only if*

$$\boldsymbol{D}_{12}'\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{D}_{11} = \boldsymbol{0} \tag{3.3.2}$$

Proof. See Appendix for proof. The extra assumptions are only needed so that $(\boldsymbol{\beta}_0', \boldsymbol{\theta}_0')'$ are identified by $E(\boldsymbol{g}_{i1}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)) = \boldsymbol{0}$ and are equivalent to the Basic Assumptions of Ahn et al. (2013). I assume $E(\boldsymbol{u}_i|\boldsymbol{X}_i) = \boldsymbol{0}$ whereas Assumption 2 implies the weaker $E(\boldsymbol{u}_i|\boldsymbol{V}_i) = \boldsymbol{0}$. I make the stronger exogeneity assumption for simplicity, though the moment conditions in $\boldsymbol{g}_{i1}$ could be reformulated with $\boldsymbol{H}_0'\boldsymbol{V}_i \subset \boldsymbol{w}_i$. □

There is no reason to believe equation (3.3.2) holds in general, and so the additional moments improve the efficiency of estimating $\boldsymbol{\beta}_0$ using the QLD residual in equation (3.2.7). Trivial cases where equation (3.3.2) holds includes $\boldsymbol{\theta}_0$ being known to the researcher and $p_0 = 0$.

3.3.2 Pooled and Mean Group QLD

The QLD GMM approach of Ahn et al. (2013) can select appropriate instruments for a given time period. However, an abundance of moment conditions can induce finite-sample bias and local stationary points in the GMM objective function. This section introduces the linear pooled and mean group estimators based off of the QLD transformation. They allow for a variety of rank and exogeneity conditions which are especially useful when the researcher includes heterogeneous slopes in the model, like in Section 3.4. I propose first estimating the parameters $\boldsymbol{\theta}_0$ using the pure factor structure assumed in $\boldsymbol{Z}_i$ and then running the relevant regressions using the "defactored" data

$\widehat{\boldsymbol{H}}'\boldsymbol{y}_i$ and $\widehat{\boldsymbol{H}}'\boldsymbol{X}_i$:

$$\widehat{\boldsymbol{\beta}}_{QLDP} = \left(\sum_{i=1}^{N} \boldsymbol{X}_i'\widehat{\boldsymbol{H}}\widehat{\boldsymbol{H}}'\boldsymbol{X}_i\right)^{-1} \sum_{i=1}^{N} \boldsymbol{X}_i'\widehat{\boldsymbol{H}}\widehat{\boldsymbol{H}}'\boldsymbol{y}_i \tag{3.3.3}$$

$$\widehat{\boldsymbol{\beta}}_{QLDMG} = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{X}_i'\widehat{\boldsymbol{H}}\widehat{\boldsymbol{H}}'\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'\widehat{\boldsymbol{H}}\widehat{\boldsymbol{H}}'\boldsymbol{y}_i \tag{3.3.4}$$

The **pooled quasi-long-differencing (QLDP)** estimator defined by equation (3.3.3) is the pooled OLS estimator from regressing $\widehat{\boldsymbol{H}}'\boldsymbol{y}_i$ on $\widehat{\boldsymbol{H}}'\boldsymbol{X}_i$. A similar estimator was mentioned in Breitung and Hansen (2020) but not thoroughly studied. The **mean group quasi-long-differencing (QLDMG)** estimator defined by equation (3.3.4) can be obtained by running the $T - p$ observation time series regression $\widehat{\boldsymbol{H}}'\boldsymbol{y}_i$ on $\widehat{\boldsymbol{H}}'\boldsymbol{X}_i$ for each i, and then averaging each of the N estimates. It should be noted that $\widehat{\boldsymbol{H}}'$ can be used to “defactor" any variables which are linear in $\boldsymbol{F}_0$ and not just those used in the estimator of $\boldsymbol{\theta}_0$. This observation allows for 2SLS estimation using outside instruments.

Intuitively, the mean group estimator should allow for arbitrarily random slopes at the cost of rank assumptions and precision. If the model is thought to have homogeneous slopes, one should generally choose the pooled estimator over the mean group one. I ignore its asymptotic properties until Section 3.4 when I introduce random slopes. However, the pooled QLD allows us to relax the rank conditions used in Ahn et al. (2013) and Westerlund et al. (2019). Instead of $E(\text{vec}(\boldsymbol{X}_i) \otimes \boldsymbol{H}_0'(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_0)) = \boldsymbol{0}$, we can use the moments $E(\boldsymbol{X}_i'\boldsymbol{H}_0\boldsymbol{H}_0'(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_0)) = \boldsymbol{0}$. This residual represents a just-identified system of moments, requires no outside instruments, and allows $E(\boldsymbol{\gamma}_i\boldsymbol{\gamma}_i')$ and $E(\boldsymbol{\gamma}_i)$ to be completely arbitrary. Further, since estimation of $\boldsymbol{\theta}_0$ comes from the reduced form moments, I do not require $T > K + 1$.

Before proving asymptotic normality, I point out that the case of $p = K + 1$ implies a powerful algebraic fact about the pooled QLD estimator: it is the same whether or not the researcher includes common variables in the regression. That is, all variables which do not vary over i are irrelevant to the estimation of $\boldsymbol{\beta}_0$, which includes time dummies. Further, the pooled QLD residuals are the same with or without the inclusion of common variables. Note that I say $p = K + 1$ instead of $p_0 = K + 1$ as the following theorem is purely algebraic and independent of model specification or statistical properties.

Let $\boldsymbol{W}$ be a $(T-p)\times q$ matrix of common variables, and let $(\tilde{\boldsymbol{\alpha}}', \tilde{\boldsymbol{\beta}}')'$ be the estimates from the pooled regression of $\widehat{\boldsymbol{H}}'\boldsymbol{y}_i$ on $\widehat{\boldsymbol{H}}'[\boldsymbol{W}, \boldsymbol{X}_i]$. Finally, let $\widehat{\boldsymbol{\epsilon}_i} = (\boldsymbol{y}_i - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}}_{QLDP})$ and $\tilde{\boldsymbol{\epsilon}}_i = (\boldsymbol{y}_i - \boldsymbol{X}_i\tilde{\boldsymbol{\beta}} - \boldsymbol{W}\tilde{\boldsymbol{\alpha}})$ be the associated residuals.

Theorem 3.3.3. *Suppose $p = K+1$. If $Rk(\widehat{\boldsymbol{H}}'\boldsymbol{W}) = q$, then*

1. $\widehat{\boldsymbol{\beta}}_{QLDP} = \tilde{\boldsymbol{\beta}}$.

2. $\tilde{\boldsymbol{\alpha}} = \boldsymbol{0}$.

3. $\widehat{\boldsymbol{\epsilon}_i} = \tilde{\boldsymbol{\epsilon}}_i$.

Proof. By Corollary 3.3.1, the first-stage estimator $\widehat{\boldsymbol{\theta}}$ solves $\widehat{\boldsymbol{H}}'[\overline{\boldsymbol{y}}, \overline{\boldsymbol{X}}] = \boldsymbol{0}$.

$$\sum_{i=1}^{N} \boldsymbol{X}_i'\widehat{\boldsymbol{H}}\boldsymbol{W} = N\overline{\boldsymbol{X}}'\widehat{\boldsymbol{H}}\boldsymbol{W} = \boldsymbol{0}$$

by Corollary 3.3.1, so $\widehat{\boldsymbol{H}}'\boldsymbol{X}_i$ and $\boldsymbol{W}$ are uncorrelated in the sample. Thus $\tilde{\boldsymbol{\beta}}_{QLDP} = \widehat{\boldsymbol{\beta}}_{QLDP}$. Using the same argument,

$$\begin{aligned}\tilde{\boldsymbol{\alpha}} &= \left(\sum_{i=1}^{N} \boldsymbol{W}'\widehat{\boldsymbol{H}}\widehat{\boldsymbol{H}}'\boldsymbol{W}\right)^{-1} \sum_{i=1}^{N} \boldsymbol{W}'\widehat{\boldsymbol{H}}\widehat{\boldsymbol{H}}'\boldsymbol{y}_i \\ &= N\left(\boldsymbol{W}'\widehat{\boldsymbol{H}}\widehat{\boldsymbol{H}}'\boldsymbol{W}\right)^{-1} \boldsymbol{W}'\widehat{\boldsymbol{H}}\widehat{\boldsymbol{H}}'\overline{\boldsymbol{y}} = \boldsymbol{0}\end{aligned}$$

As $\tilde{\boldsymbol{\alpha}} = \boldsymbol{0}$ and $\tilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{QLDP}$, we have $\tilde{\boldsymbol{\epsilon}}_i = \widehat{\boldsymbol{\epsilon}_i}$. □

The above result suggests that when $p = K+1$, the QLD matrix suffices to remove all unobserved time effects in the population, even those which do not interact with the heterogeneity. The intuition is similar to the ‘zero sum’ class of estimators studied by Westerlund (2019).

It may appear that Theorem 3.3.3 only applies in very special scenarios; however, simulation evidence in the Appendix suggests that overestimating p_0 does not cause inconsistency. These results bolster the simulation evidence from Ahn et al. (2013) which suggests the same thing when using their GMM estimator. Breitung and Hansen (2020) also demonstrate that the Ahn et al. (2013) estimator performs well under the BIC method of estimating p_0 which has a tendency to overestimate the number of factors. Overestimating p_0 includes the case of incorrectly estimating factors when $p_0 = 0$. Under strict exogeneity, CCE and QLD procedures will be consistent because their factor proxies are just functions of the exogenous variables. Reporting the QLDP which takes $p = K + 1$ could then serve as a robustness check if the estimated p_0 is less than $K + 1$. This fact is explored in a brief simulation study in Section 3.5.2.

I now show asymptotic normality for the pooled QLD estimator. I demonstrate how first-stage estimation of $\boldsymbol{\theta}_0$ can affect the asymptotic distribution and show why ignoring this problem leads to incorrect standard errors even when pooled QLD is asymptotically normal. I briefly discuss why the standard errors in Westerlund et al. (2019) do not account for this problem. The full proof of asymptotic normality is given in the Appendix, so I will only sketch the problem here.

Let $\boldsymbol{A}_P = E(\boldsymbol{V}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{V}_i)$. I show in the Appendix that

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}_{QLDP} - \boldsymbol{\beta}_0) = \boldsymbol{A}_P^{-1}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\boldsymbol{X}_i'\widehat{\boldsymbol{H}}\widehat{\boldsymbol{H}}'(\boldsymbol{F}_0\boldsymbol{\gamma}_i + \boldsymbol{u}_i)\right) + o_p(1)$$

After a mean value expansion about $\boldsymbol{\theta}_0$, and using the results from Theorem 3.3.1, the normalized estimator is

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}_{QLDP} - \boldsymbol{\beta}_0) = \boldsymbol{A}_P^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\boldsymbol{V}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{u}_i + \boldsymbol{G}_P\boldsymbol{r}_i(\boldsymbol{\theta}_0)\right) + o_p(1)$$

where $\boldsymbol{r}_i(\boldsymbol{\theta}_0)$ is derived from Theorem 1 and $\boldsymbol{G}_P = E(\nabla_{\boldsymbol{\theta}}\boldsymbol{X}_i'\boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{H}(\boldsymbol{\theta})'(\boldsymbol{F}_0\boldsymbol{\gamma}_i + \boldsymbol{u}_i))$ evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. $\boldsymbol{G}_P = \boldsymbol{0}$ when $E(\boldsymbol{u}_i \otimes \boldsymbol{V}_i) = \boldsymbol{0}$, $E(\boldsymbol{u}_i \otimes \boldsymbol{\Gamma}_i) = \boldsymbol{0}$, and $E(\boldsymbol{V}_i \otimes \boldsymbol{\gamma}_i) = \boldsymbol{0}$.

I only need exogeneity of $\boldsymbol{V}_i$ with respect to $\boldsymbol{u}_i$ for asymptotic normality, so the other assumptions only simplify the asymptotic variance. Westerlund et al. (2019) impose these assumptions which ignores the effect of first-stage estimation uncertainty. My result thus proves asymptotic normality of the pooled QLD under weaker assumptions than used in Westerlund et al. (2019) for the pooled CCE with an even more general asymptotic variance formula. In fact, one could only assume exogeneity on the last p_0 elements of the differenced quantities, but this assumption is difficult to interpret. I now state the general asymptotic normality result assuming $p = p_0$ is known due to Theorem 3.3.1.

Theorem 3.3.4. *Given Assumptions 1 and 2, suppose that*

1. $\boldsymbol{A}_P = E(\boldsymbol{V}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{V}_i)$ *has full rank.*

2. $E(\boldsymbol{V}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{u}_i) = \boldsymbol{0}$.

Then $\widehat{\boldsymbol{\beta}}_{QLDP} \xrightarrow{p} \boldsymbol{\beta}_0$ *and*

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}_{QLDP} - \boldsymbol{\beta}_0) \xrightarrow{p} N(\boldsymbol{0}, \boldsymbol{A}_P^{-1}\boldsymbol{B}_P\boldsymbol{A}_P^{-1})$$

where $\boldsymbol{B}_P = E((\boldsymbol{V}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{u}_i + \boldsymbol{G}_P\boldsymbol{r}_i(\boldsymbol{\theta}_0))(\boldsymbol{V}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{u}_i + \boldsymbol{G}_P\boldsymbol{r}_i(\boldsymbol{\theta}_0))')$*. If* $E(\boldsymbol{u}_i \otimes \boldsymbol{\Gamma}_i) = \boldsymbol{0}$ *and* $E(\boldsymbol{V}_i \otimes \boldsymbol{\gamma}_i) = \boldsymbol{0}$*, then* $\boldsymbol{G}_P = \boldsymbol{0}$*.*

Proof. See Appendix for proof and a derivation of $\boldsymbol{G}_P$ and $\boldsymbol{r}_i(\boldsymbol{\theta}_0)$. Condition (2) is not practically weaker than $E(\boldsymbol{u}_i|\boldsymbol{V}_i) = \boldsymbol{0}$ for linear estimation but I state it for completeness. □

Remark (Joint estimation): The two-step procedure is less efficient than joint GMM estimation using $E(\boldsymbol{V}_i'\boldsymbol{H}_0\boldsymbol{H}_0'(\boldsymbol{y}_i - \boldsymbol{V}_i\boldsymbol{\beta}_0)) = \boldsymbol{0}$ and $E(\boldsymbol{H}_0'\boldsymbol{Z}_i) = \boldsymbol{0}$ unless $p = K+1$; see Ahn and Schmidt (1997). However, the $p = K+1$ case confers the advantage of invariance to common variables from Theorem 3.3.3 and appears consistent even when $p_0 < p$. There are also optimization issues involved in joint estimation because the moments which identify $\boldsymbol{\beta}_0$ are nonlinear in $\boldsymbol{\theta}_0$. ■

Remark (Known factors): Eliminating known factors like random intercepts or polynomial time trends can make the QLD estimators more precise. Simply remove the known factors from $[\boldsymbol{y}_i, \boldsymbol{X}_i]$ by regressing it, unit-by-unit, onto the known factors, then estimate $\boldsymbol{\theta}_0$ as in Theorem 3.3.1 using the residuals. This procedure is equivalent to defining $\boldsymbol{M} = \boldsymbol{I}_T - \boldsymbol{F}_1(\boldsymbol{F}_1'\boldsymbol{F}_1)^{-1}\boldsymbol{F}_1'$, where $\boldsymbol{F}_1$ are the known factors, and running estimation based off of $(\boldsymbol{y}_i^*, \boldsymbol{X}_i^*) = (\boldsymbol{I}_N \otimes \boldsymbol{M})(\boldsymbol{y}_i, \boldsymbol{X}_i)$. ■

Remark (Bootstrap): While I provide analytic inference below, the standard errors can be quite complicated in general. Regardless of any additional restrictions which can simplify the calculation of standard errors, $\sqrt{N}(\widehat{\boldsymbol{\beta}}_{QLDP} - \boldsymbol{\beta}_0)$ is asymptotically normal so that one can instead do inference via the nonparametric bootstrap. Just resample over $(\boldsymbol{y}_i, \boldsymbol{X}_i)$, with $\widehat{\boldsymbol{H}}$ estimated for each new sample to account for the first-stage estimation in the final standard errors. This procedure contrasts to Section 2 of the Supplement to Westerlund et al. (2019) which does not estimate $\widehat{\boldsymbol{F}}$ with each new sample. I do not provide a proof of consistency because the problem is standard;

Westerlund et al. (2019) needed a proof because of the CCE projection matrix has a reduced-rank limit. ■

The asymptotic variance can be estimated by $\widehat{\boldsymbol{A}}_p^{-1}\widehat{\boldsymbol{B}}_P\widehat{\boldsymbol{A}}_P^{-1}$ where

$$\widehat{\boldsymbol{A}}_P = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{X}_i'\widehat{\boldsymbol{H}}\widehat{\boldsymbol{H}}'\boldsymbol{X}_i$$

$$\widehat{\boldsymbol{B}}_P = \frac{1}{N}\sum_{i=1}^{N}\widehat{\boldsymbol{v}}_i\widehat{\boldsymbol{v}}_i{}'$$

Here, $\widehat{\boldsymbol{v}}_i = \boldsymbol{X}_i'\widehat{\boldsymbol{H}}\widehat{\boldsymbol{H}}'\widehat{\boldsymbol{\epsilon}}_i + \boldsymbol{G}_P(\widehat{\boldsymbol{\theta}})\boldsymbol{r}_i(\widehat{\boldsymbol{\theta}})$ where $\widehat{\boldsymbol{\epsilon}}_i = \boldsymbol{y}_i - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}}_{QLDP}$ is the full pooled QLD residual. The gradient is

$$\widehat{\boldsymbol{G}}_P = \frac{1}{N}\sum_{i=1}^{N}\left((\boldsymbol{I}_K \otimes \widehat{\boldsymbol{\epsilon}}_i'\widehat{\boldsymbol{H}})\begin{pmatrix}{\boldsymbol{x}_i}_1^{*\prime} \otimes \boldsymbol{I}_{T-p_0}\\ \vdots \\ {\boldsymbol{x}_i}_K^{*}{}' \otimes \boldsymbol{I}_{T-p_0}\end{pmatrix} + \boldsymbol{X}_i'\widehat{\boldsymbol{H}}(\widehat{\boldsymbol{\epsilon}}_i^{*\prime} \otimes \boldsymbol{I}_{T-p_0})\right) \tag{3.3.5}$$

$$\boldsymbol{r}_i(\widehat{\boldsymbol{\theta}}) = (\widehat{\boldsymbol{D}}_{\boldsymbol{\theta}}'\widehat{\boldsymbol{A}}_{\boldsymbol{\theta}}^{-1}\widehat{\boldsymbol{D}}_{\boldsymbol{\theta}})^{-1}\widehat{\boldsymbol{D}}_{\boldsymbol{\theta}}'\widehat{\boldsymbol{A}}_{\boldsymbol{\theta}}^{-1}\text{vec}(\widehat{\boldsymbol{H}}'\boldsymbol{Z}_i) \tag{3.3.6}$$

where a '$*$' denotes the last p_0 elements of a $T \times 1$ vector. The form for $\boldsymbol{r}_i(\widehat{\boldsymbol{\theta}})$ comes from Theorem 3.3.1 and is derived in the proof of Theorem 3.3.4. The matrix $\boldsymbol{G}_P$ appears because of correlation between the full error $\boldsymbol{\epsilon}_i = \boldsymbol{F}_0\boldsymbol{\gamma}_i + \boldsymbol{u}_i$ and the covariates $\boldsymbol{X}_i$, and the vector $\boldsymbol{r}_i$ comes from error in estimating $\boldsymbol{\theta}_0$ in the first stage. The regular cluster-robust standard errors for a pooled regression are only valid if $\boldsymbol{G}_P = \boldsymbol{0}$. Assuming factor loadings are independent of the errors causes this matrix to be zero, like in the classical CCE treatments of Pesaran (2006) and Westerlund et al. (2019).

Though the loadings are meant to model the correlation between $\boldsymbol{X}_i$ and all unobservables, they may still correlate with the errors due to misspecification. If there are additional factors in $\boldsymbol{y}_i$ not in $\boldsymbol{X}_i$, we can still estimate $\boldsymbol{\beta}_0$ but the asymptotic variances will depend on first-stage estimation of $\boldsymbol{\theta}_0$. In fact, if we allow for uncorrelated loadings, the CCE and QLD estimators exclude relevant information for estimation. Additionally assuming $E(\boldsymbol{V}_i|\boldsymbol{\gamma}_i) = \boldsymbol{0}$ like in Westerlund et al. (2019), we have:

$$E((\boldsymbol{H}_0'\boldsymbol{V}_i) \otimes \boldsymbol{H}_0'(\boldsymbol{y}_i - \boldsymbol{V}_i\boldsymbol{\beta}_0)) = \boldsymbol{0} \tag{3.3.7}$$

$$E((\boldsymbol{H}_0'\boldsymbol{V}_i) \otimes (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_0)) = \boldsymbol{0} \tag{3.3.8}$$

$$E(\boldsymbol{X}_i \otimes \boldsymbol{H}_0'(\boldsymbol{y}_i - \boldsymbol{V}_i\boldsymbol{\beta}_0)) = \boldsymbol{0} \tag{3.3.9}$$

$$E(\boldsymbol{H}_0'(\boldsymbol{y}_i - \boldsymbol{V}_i\boldsymbol{\beta}_0)) = \boldsymbol{0} \tag{3.3.10}$$

$$E(\boldsymbol{H}_0'\boldsymbol{V}_i) = \boldsymbol{0} \tag{3.3.11}$$

Equations (3.3.7)-(3.3.11) list $(T - p_0)((T - p_0)K + 2TK + K + 1)$ moment conditions which displays the strength of the CCE assumptions made in current applications. Without at least theoretically justifying $E(\boldsymbol{V}_i \otimes \boldsymbol{\gamma}_i) = \boldsymbol{0}$, CCE-based inference needs a modern treatment which accounts for first-stage estimation as in Brown et al. (2021). To summarize, if the loadings are allowed to be correlated, then the pooled CCE standard errors from Pesaran (2006) and Westerlund et al. (2019) are incorrect. If the loadings are assumed uncorrelated, then we have a significant number of unused moment restrictions. In fact, if first-stage estimation does not affect the asymptotic distribution, and the conditional covariance $E(\boldsymbol{u}_i\boldsymbol{u}_i'|\boldsymbol{X}_i)$ is estimable, the feasible version of the GLS estimator from Section 3.2 of Brown (2021) is $\sqrt{N}$-consistent and efficient among all estimators based off of $E(\boldsymbol{M}_{\boldsymbol{F}_0}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_0)) = \boldsymbol{0}$ in which case all the moments in equations (3.3.7)-(3.3.11) are redundant.

3.4 Heterogeneous Slopes

I now consider a generalization of the population model in equation (3.2.1) which allows for random slopes.

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta}_i + \boldsymbol{F}_0\boldsymbol{\gamma}_i + \boldsymbol{u}_i \tag{3.4.1}$$

$$\boldsymbol{\beta}_i = \boldsymbol{\beta}_0 + \boldsymbol{b}_i \tag{3.4.2}$$

$$\boldsymbol{b}_i \sim (\boldsymbol{0}, \boldsymbol{\Sigma}) \tag{3.4.3}$$

The random slopes model is identical to the forms in Wooldridge (2005) and Pesaran (2006) though the former assumes $\boldsymbol{F}_0$ is observable. Neither Ahn et al. (2013) nor Westerlund (2019) consider random slopes in their fixed-T analyses. I summarize this model in the following assumption:

Assumption 3 (Random slopes):

1. $\boldsymbol{y}_i = \boldsymbol{X}_i(\boldsymbol{\beta}_0 + \boldsymbol{b}_i) + \boldsymbol{F}_0\boldsymbol{\gamma}_i + \boldsymbol{u}_i$.

2. $(\boldsymbol{X}_i, \boldsymbol{b}_i, \boldsymbol{\gamma}_i, \boldsymbol{u}_i)$ are independent and identically distributed across i with finite fourth moments.

3. $E(\boldsymbol{b}_i) = \boldsymbol{0}$.

■

The iid sampling assumption on $\boldsymbol{b}_i$ does not rule out correlation between $\boldsymbol{b}_i$ and the other stochastic components of the model. Similarly, Assumption 3(3) places no restrictions on the correlation between $\boldsymbol{b}_i$ and $\boldsymbol{X}_i$. It only states that $\boldsymbol{b}_i$ is the heterogeneous, unobserved deviation from the population parameters $\boldsymbol{\beta}_0$.

Most fixed-T treatments of random slope models either exclude factors altogether or simplify the factor structure as in a fixed effects analysis. Examples of fixed effects treatments include Juhl and Lugovskyy (2014) Campello et al. (2019), and Breitung and Salish (2021). Though Pesaran (2006), Chudik and Pesaran (2015), Neal (2015), and Norkutė et al. (2021) allow for random slopes and arbitrary factors, they require T to grow to infinity and make strong exogeneity conditions which I avoid.

Before continuing with the analysis, I want to address how the random slopes model changes first-stage estimation of $\boldsymbol{\theta}_0$. The pure factor model for $\boldsymbol{Z}_i$ in equation (3.2.4) now takes the form

$$E(\boldsymbol{Z}_i) = \boldsymbol{F}_0 E(\boldsymbol{C}_i\boldsymbol{Q}_i) + E(\boldsymbol{U}_i\boldsymbol{Q}_i)$$

where $\boldsymbol{U}_i = [\boldsymbol{u}_i, \boldsymbol{V}_i]$. In order for the identification result in Lemma 3.3.1 to hold, we need two additional conditions. First, $\text{Rk}(E(\boldsymbol{C}_i\boldsymbol{Q}_i)) = p_0$, which is reasonable given Assumption 1. We also need $E(\boldsymbol{Q}_i\boldsymbol{U}_i) = \boldsymbol{0}$ which necessitates $E(\boldsymbol{\beta}_i'\boldsymbol{v}_{it}) = \boldsymbol{0}$ for each t, implying that $\boldsymbol{b}_i$ and $\boldsymbol{v}_{it}$ are uncorrelated but allows arbitrary correlation between $\boldsymbol{b}_i$ and $(\boldsymbol{\gamma}_i, \boldsymbol{\Gamma}_i)$. We could instead estimate $\boldsymbol{\theta}_0$ based off of $E(\boldsymbol{H}_0'\boldsymbol{X}_i) = E(\boldsymbol{H}_0'\boldsymbol{V}_i) = \boldsymbol{0}$ and require $p_0 \leq K$ instead of $K + 1$. The robustness result of Theorem 3.3.3(1) holds for $p = K$ but parts (2) and (3) are not necessarily true.

Remark (Testing for random slopes): Assumption 2 allows us to test for correlated random slopes. Assuming that $p_0 < K + 1$, we can test the model $E(\boldsymbol{H}_0'\boldsymbol{Z}_i) = \boldsymbol{0}$ using the standard overidentifying restrictions test. The moments are zero under Assumptions 2 and 3 only when $\boldsymbol{\beta}_i$ is uncorrelated with $\boldsymbol{V}_i$. ■

The remainder of this section assumes $\boldsymbol{\theta}_0$ is derived from the reduced form moments $E(\boldsymbol{H}_0'\boldsymbol{V}_i) = \boldsymbol{0}$ with an analogous result to Theorem 1 to avoid uncertainty related to the overidentifying restrictions test. I first consider the Ahn et al. (2013) estimator in the presence of random slopes. The GMM estimator cannot estimate the individual random slopes due to the well-known incidental parameters problem. As such, I consider estimation which ignores the random slopes so that $\boldsymbol{X}_i\boldsymbol{b}_i$ is absorbed into the error. The Ahn et al. (2013) expected residual becomes

$$E(\text{vec}(\boldsymbol{X}_i) \otimes \boldsymbol{H}_0'(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_0)) = E(\text{vec}(\boldsymbol{X}_i) \otimes \boldsymbol{H}_0'\boldsymbol{X}_i\boldsymbol{b}_i) \tag{3.4.4}$$

Theorem 3.4.1. *Under Assumptions 1 and 3, $(\boldsymbol{\beta}_0', \boldsymbol{\theta}_0')'$ is identified by equation (3.4.4) if and only if*

$$E(vec(\boldsymbol{X}_i) \otimes \boldsymbol{H}_0'\boldsymbol{V}_i\boldsymbol{b}_i) = \boldsymbol{0}$$

Proof. The proof is a corollary of the identification result presented in Section 3.1 of Ahn et al. (2013). □

Murtazashvili and Wooldridge (2008) consider IV estimation with random slopes and known factors. The exogeneity condition in Theorem 3.4.1 can depend on the type of instruments available. If there is a vector $\boldsymbol{w}_i$ of outside instruments, one sufficient condition is

$$Cov(\boldsymbol{H}_0'\boldsymbol{X}_i, \boldsymbol{b}_i|\boldsymbol{w}_i) = Cov(\boldsymbol{H}_0'\boldsymbol{X}_i, \boldsymbol{b}_i) = \boldsymbol{0} \tag{3.4.5}$$

which is similar to Assumption 3.3 of Murtazashvili and Wooldridge (2008).

With strictly exogenous covariates, the exogeneity condition is more similar to equations (12) and (13) of Wooldridge (2005) who considers fixed effects OLS. Wooldridge shows that pooled OLS is robust to heterogeneous slopes which are uncorrelated with the matrix of second moments of the defactored covariates; that is $E(\boldsymbol{X}_i'\boldsymbol{M}_{\boldsymbol{F}_0}\boldsymbol{X}_i\boldsymbol{b}_i) = \boldsymbol{0}$ where he also assumes $\boldsymbol{F}_0$ is known. An even simpler sufficient condition would be $E(\boldsymbol{b}_i|\boldsymbol{X}_i) = \mathbf{0}$ which is in fact even weaker than the random slope assumption from Pesaran (2006) who assumes $\boldsymbol{b}_i$ is independent of all stochastic components of the model.

The Ahn et al. (2013) estimator requires stronger exogeneity and rank conditions than Wooldridge (2005) and Murtazashvili and Wooldridge (2008) because $\boldsymbol{\theta}_0$ needs to be estimated along with $\boldsymbol{\beta}_0$. If we add Assumption 2, we are able to obtain a first stage $\sqrt{N}$-consistent estimator of $\boldsymbol{\theta}_0$ by Theorem 3.3.1 and so joint identification of $(\boldsymbol{\beta}_0', \boldsymbol{\theta}_0')'$ is irrelevant. This first stage estimator allows us to substantially weaken the identification requirements for $\boldsymbol{\beta}_0$ which allows for estimation under a broader class of settings. Using the given estimator $\widehat{\boldsymbol{\theta}}$ from Theorem 3.3.1, I study the pooled QLD estimator in the context of heterogeneous slopes.

Theorem 3.4.2. *Given Assumptions 2 and 3, where $Rk(E(\boldsymbol{\Gamma}_i)) = p_0 \leq K$, suppose that*

1. $\boldsymbol{A}_P = E(\boldsymbol{V}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{V}_i)$ *has full rank.*

2. $E(\boldsymbol{V}_i'\boldsymbol{H}_0\boldsymbol{H}_0'(\boldsymbol{V}_i\boldsymbol{b}_i + \boldsymbol{u}_i)) = \mathbf{0}$.

Then $\widehat{\boldsymbol{\beta}}_{QLDP} \xrightarrow{p} \boldsymbol{\beta}_0$ *and*

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}_{QLDP} - \boldsymbol{\beta}_0) \xrightarrow{p} N(\mathbf{0}, \boldsymbol{A}_P^{-1}\boldsymbol{B}_P\boldsymbol{A}_P^{-1})$$

where $\boldsymbol{B}_P = E((\boldsymbol{V}_i'\boldsymbol{H}_0\boldsymbol{H}_0'(\boldsymbol{V}_i\boldsymbol{b}_i + \boldsymbol{u}_i) + \boldsymbol{G}_P\boldsymbol{r}_{x,i}(\boldsymbol{\theta}_0))(\boldsymbol{V}_i'\boldsymbol{H}_0\boldsymbol{H}_0'(\boldsymbol{V}_i\boldsymbol{b}_i + \boldsymbol{u}_i) + \boldsymbol{G}_P\boldsymbol{r}_{x,i}(\boldsymbol{\theta}_0))')$, $\boldsymbol{G}_P = E(\boldsymbol{\nabla}_{\boldsymbol{\theta}}\boldsymbol{V}_i'\boldsymbol{H}_0\boldsymbol{H}_0'(\boldsymbol{X}_i\boldsymbol{b}_i + \boldsymbol{F}_0\boldsymbol{\gamma}_i + \boldsymbol{u}_i))$, *and* $\boldsymbol{r}_{x,i}(\boldsymbol{\theta}_0)$ *is given in the Appendix. If* $E(\boldsymbol{u}_i \otimes \boldsymbol{\Gamma}_i) = \mathbf{0}$, $E(\boldsymbol{V}_i \otimes \boldsymbol{b}_i) = \mathbf{0}$, *and* $E(\boldsymbol{V}_i \otimes \boldsymbol{\gamma}_i) = \mathbf{0}$, *then* $\boldsymbol{G}_P = \mathbf{0}$.

Proof. The proof is identical to the proof of Theorem 3.3.4 with the full error $\boldsymbol{\epsilon}_i = \boldsymbol{X}_i\boldsymbol{b}_i + \boldsymbol{F}_0\boldsymbol{\gamma}_i + \boldsymbol{u}_i$. While $\boldsymbol{B}_P$ does not have the same form as in Theorem 3.3.4, the standard errors are calculated the same but with $\boldsymbol{r}_{x,i}$ instead of $\boldsymbol{r}_i$, and so I use the same notation. The additional rank assumption on $E(\boldsymbol{\Gamma}_i)$ allows us to estimate $\boldsymbol{\theta}_0$ via $E(\boldsymbol{H}_0'\boldsymbol{V}_i) = \mathbf{0}$ which overcomes the problems of correlation between $\boldsymbol{\beta}_i$ and $\boldsymbol{V}_i$. The asymptotic variance of $\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ and the computation of $\boldsymbol{r}_{i,x}$ are given in the Appendix. □

Consistency is not affected by the first stage estimates of $\boldsymbol{\theta}_0$ even with random slopes so that the exogeneity conditions needed are identical in spirit to Wooldridge (2005) who assumes known factors. I also do not require independence between $\boldsymbol{b}_i$ and $(\boldsymbol{X}_i, \boldsymbol{u}_i)$ like Pesaran (2006), but I still restrict the correlation between $\boldsymbol{X}_i$ and $\boldsymbol{b}_i$. This condition can be weakened via mean group estimation which allows an arbitrary conditional distribution $D(\boldsymbol{b}_i|\boldsymbol{X}_i)$ at the expense of much stronger rank and exogeneity conditions. I now state consistency and asymptotic normality for the mean group QLD estimator. Again, $\widehat{\boldsymbol{\theta}}$ is derived from $E(\boldsymbol{H}_0'\boldsymbol{V}_i) = \boldsymbol{0}$. Define $\mathcal{T}$ as the parameter space of $\boldsymbol{\theta}_0$. Finally, let $a_i(\boldsymbol{\theta}) = \sqrt{\sum_{i=1}^K \sigma_i \left((\boldsymbol{X}_i'\boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{X}_i)^{-1}\right)}$ where $\{\sigma_i(\boldsymbol{D})\}_{i=1}^K$ are the singular values of the $K \times K$ matrix $\boldsymbol{D}$.

Theorem 3.4.3. *Given Assumptions 2 and 3, where $Rk(E(\boldsymbol{\Gamma}_i)) = p_0 \leq K$, suppose that*

1. *The eigenvalues of $\boldsymbol{X}_i'\boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{X}_i$ are almost surely positive uniformly over $\mathcal{T}$.*

2. *Uniformly over $\mathcal{T}$,*

$$\max\left\{E\left(a_i(\boldsymbol{\theta})\,\|\boldsymbol{X}_i\|\,\|\boldsymbol{u}_i\|\right), E\left(a_i(\boldsymbol{\theta})^2\,\|\boldsymbol{X}_i\|^3\,\|\boldsymbol{u}_i\|\right)\right\} < \infty$$

3. *$\mathcal{T}$ is a compact subset of $\mathbb{R}^{(T-p_0)p_0}$.*

Then $\widehat{\boldsymbol{\beta}}_{QLDMG} \xrightarrow{p} \boldsymbol{\beta}_0$ and

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}_{QLDMG} - \boldsymbol{\beta}_0) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{B}_{MG})$$

where $\boldsymbol{B}_{MG} = E(\left((\boldsymbol{V}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{V}_i)^{-1}\boldsymbol{V}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{u}_i + \boldsymbol{G}_{MG}\boldsymbol{r}_{x,i}(\boldsymbol{\theta}_0)\right)\left((\boldsymbol{V}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{V}_i)^{-1}\boldsymbol{V}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{u}_i + \boldsymbol{G}_{MG}\boldsymbol{r}_{x,i}(\boldsymbol{\theta}_0)\right)')$. If $E(\boldsymbol{b}_i|\boldsymbol{V}_i) = \boldsymbol{0}$ and $E(\boldsymbol{V}_i \otimes \boldsymbol{\gamma}_i = \boldsymbol{0})$, then $\boldsymbol{G}_{MG} = \boldsymbol{0}$.

Proof. See Appendix for proof and the derivation of $\boldsymbol{G}_{MG}$. Note that Assumption 2 implies $E(\boldsymbol{u}_i|\boldsymbol{V}_i) = \boldsymbol{0}$. □

Standard errors are derived similarly to the pooled QLD estimator in Section 3.3.2. Let

$$\widehat{\boldsymbol{B}} = \frac{1}{N}\sum_{i=1}^{N}\left((\boldsymbol{X}_i'\widehat{\boldsymbol{H}}\widehat{\boldsymbol{H}}'\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'\widehat{\boldsymbol{H}}\widehat{\boldsymbol{H}}'\widehat{\boldsymbol{\epsilon}}_i\widehat{\boldsymbol{G}}_{MG}\boldsymbol{r}_{x,i}(\widehat{\boldsymbol{\theta}})\right)\left((\boldsymbol{X}_i'\widehat{\boldsymbol{H}}\widehat{\boldsymbol{H}}'\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'\widehat{\boldsymbol{H}}\widehat{\boldsymbol{H}}'\widehat{\boldsymbol{\epsilon}}_i\widehat{\boldsymbol{G}}_{MG}\boldsymbol{r}_{x,i}(\widehat{\boldsymbol{\theta}})\right)' \quad (3.4.6)$$

where $\widehat{\boldsymbol{\epsilon}}_i = \boldsymbol{y}_i - \boldsymbol{X}_i \widehat{\boldsymbol{\beta}}_{CCEMG}$ is the mean group QLD residual and $\boldsymbol{r}_{x,i}(\widehat{\boldsymbol{\theta}})$ comes from Lemma .0.2 in the Appendix. The gradient $\boldsymbol{G}_{MG}$ can be estimated via

$$\widehat{\boldsymbol{G}}_{MG} = \frac{1}{N} \sum_{i=1}^{N} - \left(\boldsymbol{I}_K \otimes \widehat{\boldsymbol{\epsilon}}_i' \widehat{\boldsymbol{H}} \widehat{\boldsymbol{H}}' \boldsymbol{X}_i \right) \left((\boldsymbol{X}_i' \widehat{\boldsymbol{H}} \widehat{\boldsymbol{H}}' \boldsymbol{X}_i)^{-1} \otimes (\boldsymbol{X}_i' \widehat{\boldsymbol{H}} \widehat{\boldsymbol{H}}' \boldsymbol{X}_i)^{-1} \right) (\boldsymbol{I}_{K^2} + \boldsymbol{K}_K)(\boldsymbol{I}_K \otimes \boldsymbol{X}_i' \widehat{\boldsymbol{H}}) * \\ * \begin{pmatrix} \boldsymbol{x}_{i_1}^{*\prime} \otimes \boldsymbol{I}_{T-p_0} \\ \vdots \\ \boldsymbol{x}_{i_K}^{*\prime} \otimes \boldsymbol{I}_{T-p_0} \end{pmatrix} + \\ + (\boldsymbol{X}_i' \widehat{\boldsymbol{H}} \widehat{\boldsymbol{H}}' \boldsymbol{X}_i)^{-1} \left(\left(\boldsymbol{I}_K \otimes \widehat{\boldsymbol{\epsilon}}_i' \widehat{\boldsymbol{H}} \right) \begin{pmatrix} \boldsymbol{x}_{i_1}^{*\prime} \otimes \boldsymbol{I}_{T-p_0} \\ \vdots \\ \boldsymbol{x}_{i_K}^{*\prime} \otimes \boldsymbol{I}_{T-p_0} \end{pmatrix} + \boldsymbol{X}_i' \widehat{\boldsymbol{H}} \left(\widehat{\boldsymbol{\epsilon}}_i^{*\prime} \otimes \boldsymbol{I}_{T-p_0} \right) \right)$$

where $\boldsymbol{K}_K$ is the $K^2 \times K^2$ commutation matrix.

As discussed in Section 3.3.2, Theorem 3.4.3 is the first fixed-T proof of asymptotic normality for a mean group estimator which allows for arbitrary random factors. While I believe the mean group CCE estimator can be adjusted to allow T fixed, it has yet to be proved, as Pesaran (2006) required $T \to \infty$. Further, it is likely that a modern proof using the methods of Karabiyik et al. (2017) and Westerlund et al. (2019) is required. Like with the pooled estimator, the $\sqrt{N}$-asymptotic normal convergence result in Theorem 3.4.3 implies that inference can be done via the usual nonparametric bootstrap, estimating $\widehat{\boldsymbol{\theta}}$ for each new bootstrap sample.

Remark (Order conditions): Similar to the pooled estimator, one advantage of the QLD transformation is that it allows for more variables than the CCE when p_0 is small. CCE uses $(\overline{\boldsymbol{y}}, \overline{\boldsymbol{X}})$ to control for the factors. The rank of $\boldsymbol{M}_{\widehat{\boldsymbol{F}}}$ is generally $T - (K + 1)$ in finite samples, regardless of the number of factors. The rank of $\widehat{\boldsymbol{H}} \widehat{\boldsymbol{H}}'$ is $T - p$ which is assumed to be greater than $T - (K + 1)$ in Westerlund et al. (2019). ■

One consequence of the strong rank conditions is that we cannot allow values which take zero for all t with positive probability. This rules out demographic dummy variables which are common in applied microeconometrics. Instead, we could just split the sample and run mean group estimation on each demographic sub sample. The estimator's precision will suffer, but this technique allows us to estimate different slope means for different groups in the population.

3.5 Simulations

This section considers the finite-sample performance of the QLD estimators compared to the GMM and CCE estimators of Ahn et al. (2013) and Pesaran (2006) respectively.

3.5.1 Main Results

The main model is

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta}_0 + \boldsymbol{F}_0\boldsymbol{\gamma}_i + \boldsymbol{u}_i$$

$$\boldsymbol{X}_i = \boldsymbol{F}_0\boldsymbol{\Gamma}_i + \boldsymbol{V}_i$$

as in Assumptions 1 and 2. There are two variables with slopes $\boldsymbol{\beta}_0 = (1,1)'$. I do not include random slopes as they would only serve to increase the amount of noise in the model and restrict the first-stage estimation of $\boldsymbol{\theta}_0$ for the QLD estimators and the cross-sectional averages for the CCE estimator. Theorems 3.4.2 and 3.4.3 dictate theoretically how the estimators should perform in given scenarios. I refer the reader to Campello et al. (2019) for simulation studies regarding the performance of pooled estimators when slopes are correlated with the variables of interest.

The two factors are generated as AR(1) random processes with initial value from a normal distribution with mean 1 and variance 1, having parameters 0.75 and −0.75 respectively. The factors are generated once then fixed over repeated replications. The simulations do not substantively change if factors are repeatedly drawn[4]. As described earlier, since T is small and fixed, it is the factor loadings which cause problems asymptotically and not the factors. The loadings on $\boldsymbol{X}_i$ are drawn as

$$\boldsymbol{\Gamma}_i \sim \begin{pmatrix} N(1,1) & N(0,1) \\ N(0,1) & N(1,1) \end{pmatrix}$$

[4] Additional simulations are available upon request.

so that $\boldsymbol{\theta}_0$ is identified from the reduced form moments. The loadings in $\boldsymbol{y}_i$ are drawn

$$\boldsymbol{\gamma}_i \sim \begin{pmatrix} N(\Gamma_{1,1}, 1) \\ N(\Gamma_{2,2}, 1) \end{pmatrix}$$

The errors $\boldsymbol{u}_i$ and $\boldsymbol{V}_{ik}$ ($k = 1, 2$) are drawn from a multivariate normal distribution with mean $\boldsymbol{0}_{T\times 1}$ and variance $\boldsymbol{C}$ where $\boldsymbol{C}$ is the correlation matrix from an AR(1) process with parameter 0.75. That is, the two errors in $\boldsymbol{V}_i = (\boldsymbol{V}_{i1}, \boldsymbol{V}_{i2})$ are both drawn from $MVN(\boldsymbol{0}_{T\times 1}, \boldsymbol{C})$ but are independent of each other and $\boldsymbol{u}_i$. Each simulation study includes 1000 replications.

Table 3.1 compares the Ahn et al. (2013) estimator both with and without the additional moments $E(\boldsymbol{H}_0'\boldsymbol{Z}_i) = \boldsymbol{0}$. Both estimators are computed as two-step estimators where the optimal weight matrix is calculated with a consistent first-step estimator. The first-step estimator uses an identity weight matrix.

Table 3.1: GMM estimators

		Bias		**SD**		**RMSE**	
		GMM1	**GMM2**	**GMM1**	**GMM2**	**GMM1**	**GMM2**
N = 50	**T = 3**	0.0328	-0.0107	0.2326	0.1812	0.2349	0.1815
		-0.0053	-0.0167	0.1719	0.1690	0.1720	0.1698
	T = 4	-0.0019	-0.0225	0.1444	0.1518	0.1444	0.1535
		0.0137	-0.0196	0.1626	0.1424	0.1632	0.1438
	T = 5	0.0170	-0.0249	0.1701	0.1694	0.1710	0.1712
		0.1375	-0.0055	0.3080	0.2057	0.3373	0.2058
N = 300	**T = 3**	0.0328	-0.0107	0.2326	0.1812	0.2349	0.1815
		-0.0053	-0.0167	0.1719	0.1690	0.1720	0.1698
	T = 4	-0.0019	-0.0225	0.1444	0.1518	0.1444	0.1535
		0.0137	-0.0196	0.1626	0.1424	0.1632	0.1438
	T = 5	0.0005	-0.0016	0.0363	0.0364	0.0363	0.0365
		0.0156	-0.0029	0.1014	0.0367	0.1026	0.0368

The GMM estimator based off of the Ahn et al. (2013) residual $E(\text{vec}(\boldsymbol{X}_i) \otimes \boldsymbol{H}_0'(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_0))$ only is GMM1, whereas the GMM estimator using the Ahn et al. residual and the additional moments $E(\boldsymbol{H}_0'\boldsymbol{Z}_i) = \boldsymbol{0}$ is GMM2. The GMM estimator using both sets of moments consistently outperforms the original Ahn et al. (2013) estimator in terms of both bias and standard deviation implying that the additional moments are practically relevant in finite samples.

Before turning to a comparison of the pooled QLD and CCE estimators, I first investigate the performance of QLDP when p_0 is misspecified in estimation of $\boldsymbol{\theta}_0$. The simulation setting implies $p_0 = 2$, so I look at the performance of QLDP for $p = 1, 2, 3$. I reiterate that p_0 is given by the DGP and p is the number of factors specified by the econometrician.

Table 3.2: Misspecifying p_0

		Bias			**SD**			**RMSE**		
		p = 1	**p = 2**	**p = 3**	**p = 1**	**p = 2**	**p = 3**	**p = 1**	**p = 2**	**p = 3**
N = 50	**T = 4**	0.2700	0.0078	0.0118	0.1677	0.1097	0.1466	0.3178	0.1100	0.1471
		0.4024	0.0029	0.0120	0.1814	0.1097	0.1561	0.4414	0.1098	0.1566
	T = 5	0.4662	0.0095	0.0154	0.3511	0.1005	0.1282	0.5836	0.1009	0.1291
		0.5372	0.0058	0.0119	0.4111	0.0950	0.1228	0.6764	0.0952	0.1234
	T = 6	0.1697	0.0074	0.0126	0.1534	0.0956	0.1239	0.2287	0.0959	0.1246
		0.5843	0.0132	0.0200	0.1516	0.1025	0.1222	0.6036	0.1034	0.1238
N = 300	**T = 4**	0.2748	-0.0003	0.0000	0.0657	0.0424	0.0559	0.2826	0.0424	0.0559
		0.4087	0.0024	0.0030	0.0746	0.0411	0.0587	0.4154	0.0411	0.0588
	T = 5	0.5267	0.0008	0.0032	0.2545	0.0382	0.0491	0.5849	0.0383	0.0492
		0.5993	0.0007	0.0038	0.2953	0.0369	0.0474	0.6681	0.0369	0.0476
	T = 6	0.1484	0.0015	0.0027	0.0646	0.0392	0.0470	0.1618	0.0392	0.0471
		0.6191	0.0013	0.0020	0.0596	0.0406	0.0480	0.6220	0.0406	0.0480

Table 3.2 gives the results for the QLDP under the different specifications. My results track with previous simulation evidence provided by Ahn et al. (2013) and Breitung and Hansen (2020). Underestimating p_0 leads to substantial bias which does not decrease with N. However, overestimating p_0 leads to only slightly worse performance than correct specification. The bias is larger but decreases with N; in fact, even $N = 300$ gives reasonable bias for the $p = 3$ estimator. The $p = 3$ estimator also performs worse than the correctly specified estimator in terms of standard deviation, which is not surprising. Overall, I find evidence that overestimation of p_0 does not lead to substantial bias in estimation, but underestimating p_0 can.

I now turn to comparison of the QLDP and CCEP estimators. Tables 3.3 and 3.4 look at the QLDP estimator compared to the CCEP estimator where the QLD transformation is estimated under $p = p_0 = 2$. Table 3.3 contains results for $K = 2$ and table 3.4 contains results for $K = 3$. I include $K = 3$ because it demonstrates how CCE removes more information as K grows but QLD does not. First note that the CCEP is biased when $T = 3$ as $K + 1 = 3$ and this order condition is not allowed.

However, the QLDP is still consistent here. Further, the QLD estimators takes p_0 as known while the CCE estimators “overestimates" p_0 with the cross-sectional averages, of which there are $K + 1$. One might suspect this overestimation leads to inefficiency which is demonstrated by the results of the simulations. The QLDP estimator consistently shows a 15‰-25‰ decline in standard deviation over the CCE estimator. Further, the CCE identifying condition requires $T > K + 1$ which causes severe bias when violated. The QLDP estimator significantly outperforms the CCEP estimator in every setting provided.

Table 3.3: Pooled estimators, $K = 2$

		Bias		**SD**		**RMSE**	
		CCEP	**QLDP**	**CCEP**	**QLDP**	**CCEP**	**QLDP**
N = 50	**T = 3**	-0.5525	0.0082	25.9618	0.1546	25.9676	0.1548
		1.2734	0.0034	12.5824	0.1555	12.6467	0.1556
	T = 4	0.0118	0.0078	0.1466	0.1097	0.1471	0.1100
		0.0120	0.0029	0.1561	0.1097	0.1566	0.1098
	T = 5	0.0197	0.0095	0.1220	0.1005	0.1236	0.1009
		0.0089	0.0058	0.1152	0.0950	0.1155	0.0952
N = 300	**T = 3**	0.0272	0.0024	2.7295	0.0580	2.7296	0.0581
		0.9400	0.0026	3.3976	0.0585	3.5253	0.0585
	T = 4	0.0000	-0.0003	0.0559	0.0424	0.0559	0.0424
		0.0030	0.0024	0.0587	0.0411	0.0588	0.0411
	T = 5	0.0050	0.0008	0.0464	0.0382	0.0467	0.0383
		0.0027	0.0007	0.0441	0.0369	0.0442	0.0369

Comparing table 3.3 to table 3.1, the QLDP performs much better than either of the GMM estimators despite the fact that we know they are using valid instruments. That the QLDP has better finite-sample performance than the overidentified systems from Ahn et al. (2013) is most likely due to the fact that it uses a smaller, just identified system of moments. See the Appendix for additional simulations including larger values of T.

Finally, I investigate the performance of the mean group quasi-long-differencing (QLDMG) and mean group common correlated effects (CCEMG) estimators. The QLDMG estimator is given by equation (3.3.4) and the CCEMG estimator is identical to the QLDMG estimator but with $\boldsymbol{M}_{\widehat{\boldsymbol{F}}}$ in place of $\widehat{\boldsymbol{H}}\widehat{\boldsymbol{H}}'$. Consistency is proved in Pesaran (2006) but, like the pooled estimator, will

Table 3.4: Pooled estimators, $K = 3$

		Bias		**SD**		**RMSE**	
		CCEP	**QLDP**	**CCEP**	**QLDP**	**CCEP**	**QLDP**
N = 50	**T = 3**	0.0875	0.0076	3.0883	0.1586	3.0895	0.1588
		1.0809	0.0094	2.2956	0.1594	2.5373	0.1597
		0.3240	-0.0018	7.6585	0.1560	7.6654	0.1560
	T = 4	0.1574	0.0041	3.1025	0.1105	3.1065	0.1106
		1.1709	0.0140	3.2437	0.1107	3.4486	0.1116
		-0.2552	-0.0047	6.7375	0.1089	6.7423	0.1090
	T = 5	0.0151	0.0066	0.1530	0.0986	0.1537	0.0988
		0.0039	0.0031	0.1495	0.0979	0.1495	0.0979
		-0.0072	-0.0041	0.1408	0.0958	0.1410	0.0959
N = 300	**T = 3**	1.9936	0.0030	61.6795	0.0580	61.7117	0.0581
		2.5873	0.0007	45.5170	0.0578	45.5905	0.0578
		-0.8012	0.0017	17.5764	0.0570	17.5947	0.0570
	T = 4	0.0011	0.0008	0.0601	0.0397	0.0601	0.0397
		0.0028	0.0001	0.0559	0.0394	0.0560	0.0394
		0.0035	0.0009	0.0571	0.0378	0.0572	0.0378
	T = 5	0.0064	0.0028	2.0502	0.0400	2.0502	0.0401
		1.0163	0.0020	0.9861	0.0414	1.4160	0.0414
		-0.0826	0.0006	3.6462	0.0400	3.6471	0.0400

eventually require a modern treatment which either controls for the asymptotic degeneracy in $\boldsymbol{M}_{\widehat{\boldsymbol{F}}}$ like Karabiyik et al. (2017) and Westerlund et al. (2019) or assumes full rank limits like Brown et al. (2021). Table 3.5 contains the results for the mean group estimators where the QLD transformation is estimated assuming $p = p_0 = 2$. I start at $T = 5$ so that $T - p_0 > p_0$ and the CCEMG estimator is well-defined.

Despite $T > 2K + 1$ for each setting, the CCEMG estimator exhibits substantial bias when $T = 6$, though the QLDMG estimator appears unbiased. The QLDMG outperforms the CCEMG in terms of RMSE for each N and T besides $N = 600$ and $T = 8$. We would expect the CCEMG to perform well relative to the QLDMG as T grows due to the incidental parameter problem in the first-stage QLD estimation. However, even for moderately low values of N and large values of T, the QLDMG has optimistic properties.

Table 3.5: Mean group estimators

		Bias		**SD**		**RMSE**	
		CCEMG	**QLDMG**	**CCEMG**	**QLDMG**	**CCEMG**	**QLDMG**
N = 50	**T = 5**	-1.5703	-0.0055	34.8038	0.4837	34.8392	0.4837
		-0.4832	0.0256	18.2402	0.6523	18.2466	0.6529
	T = 6	0.0324	0.0056	0.4630	0.1737	0.4641	0.1738
		0.0256	0.0044	0.3774	0.1820	0.3782	0.1820
	T = 7	0.0187	0.0156	0.1670	0.1658	0.1681	0.1665
		0.0113	0.0102	0.1628	0.1574	0.1632	0.1577
N = 300	**T = 5**	-1.2597	-0.0039	27.7644	0.1537	27.7929	0.1537
		1.1968	-0.0030	34.6115	0.1420	34.6322	0.1420
	T = 6	-0.0077	0.0039	0.2846	0.0767	0.2847	0.0768
		0.0116	-0.0004	0.1768	0.0745	0.1772	0.0745
	T = 7	0.0003	0.0000	0.0649	0.0641	0.0649	0.0641
		0.0010	0.0009	0.0677	0.0595	0.0677	0.0595

3.5.2 Comparison to TWFE

Theorem 3.3.3 suggests a certain robustness property for the QLDP estimator with respect to the traditional TWFE estimator. If the factor structure gives the traditional two-way error $\boldsymbol{f}_t'\boldsymbol{\gamma}_i + u_{it} = \gamma_i + f_t + u_{it}$, the QLDP can accommodate the time and individual fixed effects without Assumption 2 holding. If one regresses out a heterogeneous intercept and estimates $\widehat{\boldsymbol{\theta}}$ assuming $p = K + 1$, the QLDP estimator will be consistent even if it is nonlinear in the unobserved effects. I first demonstrate that TWFE is inconsistent in the presence of an arbitrary factor structure. The DGP is the same as Section 3.5.1 so that the QLDP results are identical to table 3.2.

TWFE performs poorly as expected. I now generate the data according to the two-way error model so that

$$y_{it} = x_{it1} + x_{it2} + t + \gamma_i + u_{it}$$

where t is the time effect and $\gamma_i \sim N(1, 1)$ is the individual effect. The covariates are generated as

$$x_{it1} \sim \text{Poisson}(|c_i + t|)$$
$$x_{it2} \sim U(0, \log((c_i + t)^2))$$

so that Assumption 2 does not hold. The simulation results in table 3.7 compare TWFE to QLDP

Table 3.6: AR(1) factor structure

		Bias		**SD**		**RMSE**	
K = 2		**TWFE**	**QLDP**	**TWFE**	**QLDP**	**TWFE**	**QLDP**
N = 50	**T = 3**	0.0791	0.0082	0.1366	0.1546	0.1578	0.1548
		0.8684	0.0034	0.1339	0.1555	0.8787	0.1556
	T = 4	0.1148	0.0078	0.1351	0.1097	0.1773	0.1100
		0.8321	0.0029	0.1330	0.1097	0.8427	0.1098
	T = 5	0.1116	0.0095	0.1290	0.1005	0.1706	0.1009
		0.8107	0.0058	0.1302	0.0950	0.8211	0.0952
N = 300	**T = 3**	0.0765	0.0024	0.0528	0.0580	0.0929	0.0581
		0.8851	0.0026	0.0513	0.0585	0.8865	0.0585
	T = 4	0.1089	-0.0003	0.0527	0.0424	0.1210	0.0424
		0.8321	0.0024	0.0527	0.0411	0.8337	0.0411
	T = 5	0.1119	0.0008	0.0529	0.0382	0.1238	0.0383
		0.8055	0.0007	0.0530	0.0369	0.8073	0.0369

when $\widehat{\boldsymbol{\theta}}$ is computed with $p = K + 1$ (despite the fact that $p_0 = 1$) and after removing a random intercept for $\boldsymbol{X}_i$ and $\boldsymbol{y}_i$ unit-by-unit. That is, let $\boldsymbol{M}$ be the $T \times T$ within transformation. I compute $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\beta}}_{QLDP}$ with $\boldsymbol{y}_i^*$ and $\boldsymbol{X}_i^*$ where $\boldsymbol{y}_i^* = \boldsymbol{M}\boldsymbol{y}_i$ and $\boldsymbol{X}_i^* = \boldsymbol{M}\boldsymbol{X}$. The time effects are irrelevant because the QLDP estimator is the same regardless of whether or not they are controlled for in the regression.

Table 3.7: TWFE specification

		Bias		**SD**		**RMSE**	
		TWFE	**QLDP**	**TWFE**	**QLDP**	**TWFE**	**QLDP**
N = 50	**T = 4**	-0.0004	-0.0044	0.0284	0.0388	0.0284	0.0390
		-0.0006	-0.0013	0.0184	0.0276	0.0184	0.0277
	T = 5	-0.0010	-0.0022	0.0240	0.0300	0.0240	0.0301
		0.0000	-0.0015	0.0142	0.0196	0.0142	0.0197
	T = 6	-0.0004	-0.0022	0.0199	0.0251	0.0199	0.0252
		0.0007	-0.0013	0.0126	0.0157	0.0127	0.0157
N = 300	**T = 4**	-0.0003	-0.0004	0.0106	0.0142	0.0106	0.0142
		0.0003	-0.0005	0.0061	0.0086	0.0061	0.0086
	T = 5	-0.0001	-0.0004	0.0092	0.0116	0.0092	0.0116
		-0.0002	-0.0001	0.0054	0.0072	0.0054	0.0072
	T = 6	0.0001	0.0001	0.0082	0.0105	0.0082	0.0105
		-0.0002	-0.0005	0.0048	0.0065	0.0048	0.0065

While the TWFE estimator is clearly superior in terms of both bias and standard deviation when N is small, the QLDP shows promising results. When $N = 300$, the two estimators are nearly indistinguishable in terms of their bias. The QLDP's RMSE is inflated because of its higher variance, but this result is unsurprising as it is a more conservative estimator which is trying to eliminate more heterogeneity. However, it performs comparably well even though it removes more variation from the data than is needed.

3.6 Application

I evaluate the effect of expenditure per student on standardized test performance. I consider school district-level data in the state of Michigan over the time periods 1995-2001. The state of Michigan reformed education expenditure in 1994 to bring poorly-funded schools to parity with wealthier schools. See Papke (2005) for a comprehensive discussion of the data and institutional details.

There are $N = 501$ school districts observed for $T = 7$ school years over 1995-2001. I present summary statistics and descriptions for the variables of interest.

Variable	**Mean**	**Standard Deviation**	**Description**
math4	0.6939	0.1515	Fraction of fourth graders who pass the MEAP math test.
lunch	0.2886	0.1616	Fraction of students eligible for free and reduced lunch.
enroll	3112.31	7965.49	Total enrollment.
avgrexp	6385.51	1034.94	Average real expenditure per pupil.

The outcome variable, *math4*, denotes the pass rate for fourth-grade students taking a standardized math test and stands as a measure of student achievement. Michigan students undertake a battery of standardized tests in elementary, junior, and secondary school. Like Papke (2005) and Papke and Wooldridge (2008), I focus on the fourth-grade math test because it has been consistently defined and measured over the observed time periods.

The primary variable of interest is average expenditure per pupil, as it represents the effect of additional expenditure on test scores. Starting in the 1994/1995 school year, the state of Michigan began awarding so-called “foundation grants" which were based on the per-student spending of the school district in the previous year. The goal was to eventually bring schools up to a benchmark “basic foundation" amount which increased over time. The state started by awarding foundation grants to increase expenditure to a minimum \$4200 per student or an additional \$250 per student, whichever was higher. By 2000, the minimum and benchmark amounts were equal at \$5700. Expenditures per pupil were averaged over the current year as well as the previous three, meaning average real expenditure per pupil in 1995 is an average of expenditure in 1992, 1993, 1994, and 1995.

The equation of interest is

$$math4_{it} = c_i + \log(avgrexp_{it})\beta_1 + lunch_{it}\beta_2 + \log(enroll_{it})\beta_3 + \boldsymbol{f}_t'\boldsymbol{\gamma}_i + e_{it} \tag{3.6.1}$$

which is similar to Papke (2005). I collect $lunch_{it}$, $\log(enroll)_{it}$, and $\log(avgrexp)_{it}$ and use the reduced form CCE equation from Assumption 2 to implement the pooled QLD estimator. This specification allows me to test for the number of factors. I also use the Ahn et al. (2013) GMM function to test for p_0, with and without the CCE equations.

Table 3.8 provides the p-values for testing the hypothesis $H_0 : p_0 = p$ versus $H_1 : p_0 > p$.

Table 3.8: Testing for p_0

		p-values	
	RF2	**GMM1**	**GMM2**
$\boldsymbol{p_0 = 0}$	0.0000	0.0000	0.0000
$\boldsymbol{p_0 = 1}$	0.0000	0.0000	0.0000
$\boldsymbol{p_0 = 2}$	0.0000	0.4852	0.0000
$\boldsymbol{p_0 = 3}$	0.0000	0.1157	0.0000

A rejection of the hypothesis suggests more factors than the tested value, and a failure to reject suggests the current value is correct. The titles 'GMM1', 'GMM2', and 'RF' (for reduced form) refer to the respective objective function used to test the relevant hypothesis. I stress that testing for p_0 comes from a long-established literature, briefly described in Ahn et al. (2013). The only new concept I introduce with respect to this specific specification test is using the reduced form moments $E(\boldsymbol{H}_0'\boldsymbol{Z}_i) = \boldsymbol{0}$.

GMM1 is just the Ahn et al. (2013) objective function from equation (3.2.7). GMM2 is the Ahn et al. objective function with the additional moments $E(\boldsymbol{H}_0'\boldsymbol{Z}_i) = \boldsymbol{0}$. Finally, RF is just the reduced form moments $E(\boldsymbol{H}_0'\boldsymbol{Z}_i) = \boldsymbol{0}$. GMM1 suggests that the correct number of factors is $p_0 = 2$. GMM2 and RF both reject $p_0 = 2$ at any reasonable confidence level, and GMM2 rejects $p_0 = 3$, though it uses a much larger set of moments than the other two which may decrease power. It may suffer from the same global identification problems discussed in Hayakawa (2016) which suggests the GMM1 test will perform better practically. I stop testing at $p_0 = 3$ because RF is just identified at $p_0 = 4$. Regardless of the tests, the moments $E(\boldsymbol{H}_0'\boldsymbol{Z}_i) = \boldsymbol{0}$ only allow me to estimate up to four factors. Even if $p_0 > 4$, the QLDP nets more unobserved heterogeneity than TWFE.

For the purpose of comparison with the pooled QLD estimator, I include the TWFE estimator and the pooled CCE estimator. As $T = 7$ and $K = 3$, the CCE estimator can accommodate both $\overline{\boldsymbol{X}}$, $\overline{\boldsymbol{y}}$, and a heterogeneous intercept in $\widehat{\boldsymbol{F}}$. Further, the pooled QLD estimator is computed with $p = K = 3$ after eliminating a heterogeneous intercept from $\boldsymbol{X}_i$ and $\boldsymbol{y}_i$, unit-by-unit. As such, QLDP is a natural comparison to TWFE. Theorem 3.3.3 tells us that $\widehat{\boldsymbol{\beta}}_{QLDP}$ is invariant to common variables when $p = K$. Since it also eliminates a heterogeneous intercept, it will be consistent if TWFE is consistent, assuming strictly exogenous covariates.

I present results in table 3.9 which shows estimation after eliminating a heterogeneous intercept. For CCE, this simply amounts to $\widehat{\boldsymbol{F}} = (\boldsymbol{1}, \overline{\boldsymbol{y}}, \overline{\boldsymbol{X}})$. For QLDP, I project out the intercept from each $\boldsymbol{X}_i$ and $\boldsymbol{y}_i$ via the within transformation before estimating. Standard errors are in parentheses while p-values are in brackets. The reported standard errors are generated via the panel nonparametric bootstrap.

The QLDP estimator suggests substantial estimates for the effect of per student expenditures. A 10% increase in the average expenditure per student is associated with an 8.3 percentage point increase in the math test pass rate, with a p-value of 0.0009. This estimate is more than twice as large as the TWFE estimate and more than three halves the CCEP estimate. These results suggest that TWFE is not adequately controlling for the heterogeneity present in the data set. Both the CCE and QLDP estimates are statistically significant at the 5% level. The TWFE standard errors are generally smaller than CCE and QLD because it removes less variation from the data.

I also considered estimation via the mean group QLD and CCE estimators. However, both

Table 3.9: Controlling for heterogeneous intercept

	TWFE	**CCEP**	**QLDP**
lunch	-0.0419	0.0398	-0.1576
	(0.0730)	(0.1367)	(0.1637)
	[0.5658]	[0.7709]	[0.3381]
log(enroll)	0.0021	-0.0592	0.0268
	(0.0487)	(0.1497)	(0.2152)
	[0.9663]	[0.6924]	[0.8838]
log(avgrexp)	0.3771	0.5409	0.8287
	(0.0704)	(0.2695)	(0.3785)
	[0.0000]	[0.0446]	[0.0303]

parameter estimates and standard errors were unreasonable compared to the other estimators. In fact, the p-values were significantly larger than any other reported case and suggested a critical lack of precision. Recall that the mean group estimators require much stronger exogeneity and identifying conditions than the pooled estimators.

3.7 Conclusion

This paper considers fixed-T estimation of linear panel data models where the errors have a general unknown factor structure. I use the quasi-long-difference transformation studied by Ahn et al. (2013) to eliminate the factor structure and provide moment conditions for estimation. For the purpose of comparison with the popular pooled common correlated effects estimator, I study the moments implied by assuming a pure factor structure in the covariates. Applying the QLD transformation to the independent variables improves efficiency of estimating the parameters of interest in the main equation which is information that pooled CCE does not use.

Current proofs of fixed-T asymptotic normality of the pooled CCE estimator assumes loadings which are strictly exogenous with respect to the idiosyncratic errors in the independent variables. I show that the uncorrelated loadings assumptions implies the existence of an even larger number of moments which CCE neglects. Ultimately, if one makes the strong assumptions sufficient for asymptotic normality of pooled CCE in Westerlund et al. (2019), one should fully consider the information available for efficient estimation. Regardless, I provide robust standard errors in a more general and appealing setting than the CCE models in Pesaran (2006) and Westerlund et al. (2019).

I apply the moment-based perspective to a heterogeneous slopes model similar to the original Pesaran (2006) setting. I prove consistency and asymptotic normality of pooled and mean group estimators based off of the QLD transformation which put no restrictions on the relationship between T and K in contrast to CCE. These estimators are shown to outperform CCE estimators in finite samples even when N is small. The pooled QLD estimator also has the desirable property of invariance to common variables, like time trends and macroeconomic indicators, when the estimated number of factors equals the number of regressors. I reexamine estimation of school district expenditures on standardized test performance and find significantly larger effects of educational spending compared to simple fixed effects regression. These estimates are also reported up to reasonable precision which suggests that applied researchers are not adequately controlling for heterogeneity in their data.

One important direction for future work concerns the overestimation of p_0. It is known that CCE is robust to $K+1 > p_0$. Moon and Weidner (2015) prove that principal components estimation is also robust to overestimating the number of factors, provided T is large. However, while there is ample simulation evidence suggesting the robustness of QLD to such a failure, a formal proof is lacking. It would also be useful to investigate the robustness of the QLDP estimators to failure of the reduced form equation in Assumption 2. Finally, the methods presented in this paper all assumed balanced panels. Missing data causes challenges to constructing the CCE and QLD transformations. It is not clear how even a complete cases estimator would work, as the cross sectional averages and first-stage estimator of $\widehat{\boldsymbol{\theta}}$ require all time periods for each unit in the sample.

APPENDIX

APPENDIX

PROOFS FOR CHAPTER 1

This Appendix collects together proofs of the formal results stated in the text.

Proof of Lemma 1.3.1

From equation (1.3.14), Assumptions WV.1 and WV.2 imply

$$\text{Var}\left(\mathbf{y}_i|\mathbf{x}_i, c_i\right) = \alpha c_i \mathbf{M}_i^{1/2}\mathbf{R}\mathbf{M}_i^{1/2}$$

By the law of total variance,

$$\begin{aligned}
\text{Var}\left(\mathbf{y}_i|\mathbf{x}_i\right) &= \text{E}\left[\text{Var}\left(\mathbf{y}_i|\mathbf{x}_i, c_i\right)|\mathbf{x}_i\right] + \text{Var}\left[\text{E}\left(\mathbf{y}_i|\mathbf{x}_i, c_i\right)|\mathbf{x}_i\right] \\
&= \text{E}\left(\alpha c_i \mathbf{M}_i^{1/2}\mathbf{R}\mathbf{M}_i^{1/2}\middle|\mathbf{x}_i\right) + \text{Var}\left(c_i\mathbf{m}_i|\mathbf{x}_i\right) \\
&= \alpha\mu_c\left(\mathbf{x}_i\right)\mathbf{M}_i^{1/2}\mathbf{R}\mathbf{M}_i^{1/2} + \sigma_c^2\left(\mathbf{x}_i\right)\mathbf{m}_i\mathbf{m}_i' \qquad (.0.1)
\end{aligned}$$

To simplify notation in what follows, write $\mu_i \equiv \mu_c\left(\mathbf{x}_i\right)$, $\sigma_i^2 \equiv \sigma_c^2\left(\mathbf{x}_i\right)$. To derive $\mathbf{\Omega}_i^{-1}$, we apply an implication of Sherman and Morrison (1950): For a nonsingular $T \times T$ matrix $\mathbf{A}$ and $T \times 1$ vector $\mathbf{b}$,

$$\left(\mathbf{A} + \mathbf{b}\mathbf{b}'\right)^{-1} = \mathbf{A}^{-1} - \frac{1}{1 + \mathbf{b}'\mathbf{A}^{-1}\mathbf{b}}\mathbf{A}^{-1}\mathbf{b}\mathbf{b}'\mathbf{A}^{-1} \qquad (.0.2)$$

which can be verified by direct multiplication. Take $\mathbf{A} \equiv \alpha\mu_i\mathbf{M}_i^{1/2}\mathbf{R}\mathbf{M}_i^{1/2}$ and $\mathbf{b} \equiv \sigma_i\mathbf{m}_i$ in (.0.2) and note that $\left[\alpha\mu_i\mathbf{M}_i^{1/2}\mathbf{R}\mathbf{M}_i^{1/2}\right]^{-1} = \mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\mathbf{M}_i^{-1/2}/(\alpha\mu_i)$ and $\mathbf{M}_i^{-1/2}\mathbf{m}_i = \sqrt{\mathbf{m}_i}$.

Therefore,

$$\begin{aligned}
\boldsymbol{\Omega}_i^{-1} &= \frac{1}{\alpha\mu_i}\mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\mathbf{M}_i^{-1/2} \\
&\quad - \frac{1}{1+\left[\sigma_i^2\sqrt{\mathbf{m}_i}'\mathbf{R}^{-1}\sqrt{\mathbf{m}_i}\right]/(\alpha\mu_i)}\sigma_i^2\mathbf{R}^{-1}\sqrt{\mathbf{m}_i}\sqrt{\mathbf{m}_i}'\mathbf{R}^{-1}/(\alpha\mu_i)^2 \\
&= \frac{1}{\alpha\mu_i}\mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\mathbf{M}_i^{-1/2} \\
&\quad - \frac{\sigma_i^2}{\alpha\mu_i+\sigma_i^2\sqrt{\mathbf{m}_i}'\mathbf{R}^{-1}\sqrt{\mathbf{m}_i}}\sigma_i^2\mathbf{R}^{-1}\sqrt{\mathbf{m}_i}\sqrt{\mathbf{m}_i}'\mathbf{R}^{-1}/(\alpha\mu_i) \\
&= \frac{1}{\alpha\mu_i}\mathbf{M}_i^{-1/2}\left\{\mathbf{R}^{-1} - \frac{\sigma_i^2}{\left[\alpha\mu_i+\sigma_i^2\sqrt{\mathbf{m}_i}'\mathbf{R}^{-1}\sqrt{\mathbf{m}_i}\right]}\mathbf{R}^{-1}\sqrt{\mathbf{m}_i}\sqrt{\mathbf{m}_i}'\mathbf{R}^{-1}\right\}\mathbf{M}_i^{-1/2}
\end{aligned}$$

□

Proof of Theorem 1.3.1

Simplify the notation by defining $\mathbf{D}_i \equiv \mathbf{D}_o(\mathbf{x}_i)$, $\mathbf{V}_i \equiv \mathbf{V}_o(\mathbf{x}_i)$, $\mu_i \equiv \mu_c(\mathbf{x}_i)$, $\sigma_i^2 \equiv \sigma_c^2(\mathbf{x}_i)$, and drop dependences on $\boldsymbol{\beta}_0$. With this simplified notation,

$$\mathbf{V}_i^- = \boldsymbol{\Omega}_i^{-1} - \boldsymbol{\Omega}_i^{-1}\mathbf{m}_i\left(\mathbf{m}_i'\boldsymbol{\Omega}_i^{-1}\mathbf{m}_i\right)^{-1}\mathbf{m}_i'\boldsymbol{\Omega}_i^{-1}$$

and, from Lemma 1.3.1,

$$\boldsymbol{\Omega}_i^{-1} = \frac{1}{\alpha\mu_i}\mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\mathbf{M}_i^{-1/2} - \frac{\sigma_i^2}{\alpha\mu_i\left(\alpha\mu_i+a_i\sigma_i^2\right)}\mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\sqrt{\mathbf{m}_i}\sqrt{\mathbf{m}_i}'\mathbf{R}^{-1}\mathbf{M}_i^{-1/2}$$

where $a_i \equiv \sqrt{\mathbf{m}_i}'\mathbf{R}_i^{-1}\sqrt{\mathbf{m}_i}$. Therefore, because $\mathbf{M}_i^{-1/2}\mathbf{m}_i = \sqrt{\mathbf{m}_i}$,

$$\begin{aligned}
\boldsymbol{\Omega}_i^{-1}\mathbf{m}_i &= \frac{1}{\alpha\mu_i}\mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\sqrt{\mathbf{m}_i} - \frac{\sigma_i^2}{\alpha\mu_i\left(\alpha\mu_i+a_i\sigma_i^2\right)}\mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\sqrt{\mathbf{m}_i}\sqrt{\mathbf{m}_i}'\mathbf{R}^{-1}\sqrt{\mathbf{m}_i} \\
&= \frac{1}{\alpha\mu_i}\mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\sqrt{\mathbf{m}_i} - \frac{a_i\sigma_i^2}{\alpha\mu_i\left(\alpha\mu_i+a_i\sigma_i^2\right)}\mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\sqrt{\mathbf{m}_i} \\
&= \left[\frac{1}{\alpha\mu_i} - \frac{a_i\sigma_i^2}{\alpha\mu_i\left(\alpha\mu_i+a_i\sigma_i^2\right)}\right]\mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\sqrt{\mathbf{m}_i} \\
&= \frac{\left[\left(\alpha\mu_i+a_i\sigma_i^2\right)-a_i\sigma_i^2\right]}{\alpha\mu_i\left(\alpha\mu_i+a_i\sigma_i^2\right)}\mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\sqrt{\mathbf{m}_i} \\
&= \frac{1}{\left(\alpha\mu_i+a_i\sigma_i^2\right)}\mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\sqrt{\mathbf{m}_i}
\end{aligned}$$

Also,

$$\mathbf{m}_i'\boldsymbol{\Omega}_i^{-1}\mathbf{m}_i = \frac{1}{\left(\alpha\mu_i + a_i\sigma_i^2\right)}\sqrt{\mathbf{m}_i}'\mathbf{R}^{-1}\sqrt{\mathbf{m}_i} = \frac{a_i}{\alpha\mu_i + a_i\sigma_i^2}$$

It follows that

$$\boldsymbol{\Omega}_i^{-1}\mathbf{m}_i\left(\mathbf{m}_i'\boldsymbol{\Omega}_i^{-1}\mathbf{m}_i\right)^{-1}\mathbf{m}_i'\boldsymbol{\Omega}_i^{-1} = \frac{1}{a_i\left(\alpha\mu_i + a_i\sigma_i^2\right)}\mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\sqrt{\mathbf{m}_i}\sqrt{\mathbf{m}_i}'\mathbf{R}^{-1}\mathbf{M}_i^{-1/2}$$

Plugging into $\mathbf{V}_i^-$ gives

$$\begin{aligned}
\mathbf{V}_i^- &= \frac{1}{\alpha\mu_i}\mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\mathbf{M}_i^{-1/2} - \frac{\sigma_i^2}{\alpha\mu_i\left(\alpha\mu_i + a_i\sigma_i^2\right)}\mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\sqrt{\mathbf{m}_i}\sqrt{\mathbf{m}_i}'\mathbf{R}^{-1}\mathbf{M}_i^{-1/2}\\
&\quad - \frac{1}{a_i\left(\alpha\mu_i + a_i\sigma_i^2\right)}\mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\sqrt{\mathbf{m}_i}\sqrt{\mathbf{m}_i}'\mathbf{R}^{-1}\mathbf{M}_i^{-1/2}\\
&= \frac{1}{\alpha\mu_i}\mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\mathbf{M}_i^{-1/2} - \left[\frac{\alpha\mu_i + a_i\sigma_i^2}{a_i\alpha\mu_i\left(\alpha\mu_i + a_i\sigma_i^2\right)}\right]\mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\sqrt{\mathbf{m}_i}\sqrt{\mathbf{m}_i}'\mathbf{R}^{-1}\mathbf{M}_i^{-1/2}\\
&= \frac{1}{\alpha\mu_i}\left[\mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\mathbf{M}_i^{-1/2} - \frac{1}{a_i}\mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\sqrt{\mathbf{m}_i}\sqrt{\mathbf{m}_i}'\mathbf{R}^{-1}\mathbf{M}_i^{-1/2}\right]
\end{aligned}$$

which completes the result for $\mathbf{V}_i^-$. From (1.3.10), the optimal IVs are

$$\mathbf{D}_i'\mathbf{V}_i^- = -\mu_i\nabla_{\boldsymbol{\beta}}\mathbf{m}_i'\mathbf{V}_i^- = -\frac{1}{\alpha}\nabla_{\boldsymbol{\beta}}\mathbf{m}_i'\left[\mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\mathbf{M}_i^{-1/2} - \frac{1}{a_i}\mathbf{M}_i^{-1/2}\mathbf{R}^{-1}\sqrt{\mathbf{m}_i}\sqrt{\mathbf{m}_i}'\mathbf{R}^{-1}\mathbf{M}_i^{-1/2}\right]$$

and we can drop $-1/\alpha$ and factor out $\mathbf{M}_i^{-1/2}$ to get the result. □

Proof of Corollary 1.3.1

Putting $\mathbf{R} = \mathbf{I}_T$ into (1.3.17) and using simple algebra gives the optimal IVs as

$$\mathbf{Z}^*\left(\mathbf{x}_i\right)' = \nabla_{\boldsymbol{\beta}}\mathbf{m}_i\left(\boldsymbol{\beta}_0\right)'\left(\mathbf{M}_i^{-1} - \frac{1}{\sum_{r=1}^T m_{ir}}\mathbf{1}_T\mathbf{1}_T'\right)$$

We show that this choice of instruments leads to the FEP first order condition, as expressed by Wooldridge (1999), using the definition of $\mathbf{W}_i$ given in Section 1.2:

$$\nabla_{\boldsymbol{\beta}}\mathbf{p}_i\left(\boldsymbol{\beta}_0\right)'\mathbf{W}_i = \nabla_{\boldsymbol{\beta}}\mathbf{m}_i\left(\boldsymbol{\beta}_0\right)'\left[\mathbf{I}_T - \mathbf{1}_T\mathbf{p}_i\left(\boldsymbol{\beta}_0\right)'\right]\mathbf{M}_i^{-1}$$

To see the equivalence, note that

$$\mathbf{1}_T \mathbf{p}_i\left(\boldsymbol{\beta}_0\right)' \mathbf{M}_i^{-1} = \frac{1}{\left(\sum_{r=1}^{T} m_{ir}\right)} \begin{pmatrix} \mathbf{m}_i \\ \mathbf{m}_i \\ \vdots \\ \mathbf{m}_i \end{pmatrix} \mathbf{M}_i^{-1} = \frac{1}{\sum_{r=1}^{T} m_{ir}} \mathbf{1}_T \mathbf{1}_T'$$

and so

$$\nabla_{\boldsymbol{\beta}} \mathbf{p}_i\left(\boldsymbol{\beta}_0\right)' \mathbf{W}_i = \nabla_{\boldsymbol{\beta}} \mathbf{m}_i\left(\boldsymbol{\beta}_0\right)' \left(\mathbf{M}_i^{-1} - \frac{1}{\sum_{r=1}^{T} m_{ir}} \mathbf{1}_T \mathbf{1}_T'\right) = \mathbf{Z}^*\left(\mathbf{x}_i\right)'$$

□

APPENDIX

PROOFS FOR CHAPTER 2

Proof of Lemma 2.2.3

Let $p_{it}(\boldsymbol{\beta}) = m_t(\boldsymbol{x}_{it}, \boldsymbol{\beta}) \left(\sum_{s=1}^{T} m_s(\boldsymbol{x}_{is}, \boldsymbol{\beta})\right)^{-1}$, $\boldsymbol{p}_i(\boldsymbol{\beta}) = (p_{i1}(\boldsymbol{\beta}), ..., p_{iT}(\boldsymbol{\beta}))'$, and $n_i = \sum_{s=1}^{T} y_{is}$. Let $\mathbf{1}$ be a $T \times 1$ vector of ones. First I directly show the conclusion holds for $\boldsymbol{I}_T - \boldsymbol{p}(\boldsymbol{\beta})\mathbf{1}'$ which satisfies the lemma's assumption. It also satisfies Assumption MAT, which is made clear in Section 2.3. I need the following derivation:

$$\begin{aligned}\nabla_{\boldsymbol{\beta}} p_{it} &= (\sum_{r=1}^{T} m_{ir}(\boldsymbol{x}_{ir}, \boldsymbol{\beta}))^{-2}(\nabla_{\boldsymbol{\beta}} m_{it}(\boldsymbol{x}_{it}, \boldsymbol{\beta}) \sum_{r=1}^{T} m_{ir}(\boldsymbol{x}_{ir}, \boldsymbol{\beta}) - m_{it}(\boldsymbol{x}_{it}, \boldsymbol{\beta}) \sum_{r=1}^{T} \nabla_{\boldsymbol{\beta}} m_{ir}(\boldsymbol{x}_{ir}, \boldsymbol{\beta})) \\ &= (\sum_{r=1}^{T} m_{ir}(\boldsymbol{x}_{ir}, \boldsymbol{\beta}))^{-1}(\nabla_{\boldsymbol{\beta}} m_{it}(\boldsymbol{x}_{it}, \boldsymbol{\beta}) - p_{it}(\boldsymbol{\beta})(\sum_{r=1}^{T} \nabla_{\boldsymbol{\beta}} m_{ir}(\boldsymbol{x}_{ir}, \boldsymbol{\beta})))\end{aligned}$$

Stacking the T equations gives

$$\begin{aligned}\nabla_{\boldsymbol{\beta}} \boldsymbol{p}_i(\boldsymbol{\beta}) &= (\sum_{r=1}^{T} m_{ir}(\boldsymbol{x}_{ir}, \boldsymbol{\beta}))^{-1}(\nabla_{\boldsymbol{\beta}} \boldsymbol{m}_i(\boldsymbol{\beta}) - \boldsymbol{p}_i(\boldsymbol{\beta})\mathbf{1}'\nabla_{\boldsymbol{\beta}} \boldsymbol{m}_i(\boldsymbol{\beta})) \\ &= (\sum_{r=1}^{T} m_{ir}(\boldsymbol{x}_{ir}, \boldsymbol{\beta}))^{-1}(\boldsymbol{I}_T - \boldsymbol{p}_i(\boldsymbol{\beta})\mathbf{1}')\nabla_{\boldsymbol{\beta}} \boldsymbol{m}_i(\boldsymbol{\beta})\end{aligned}$$

As $E(-n_i|\boldsymbol{x}_i) = -\mu_c(\boldsymbol{x}_i) \sum_{r=1}^{T} m_{ir}(\boldsymbol{x}_{ir}, \boldsymbol{\beta}_0)$, evaluating the derivative at $\boldsymbol{\beta}_0$ and multiplying by $E(-n_i|\boldsymbol{x}_i)$ yields the final result.

Now let $\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta})$ be an $L \times T$ matrix satisfying the assumption of the lemma. $\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta})(\boldsymbol{I}_T - \boldsymbol{p}_i(\boldsymbol{\beta})\mathbf{1}') = \boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta})$ for all $\boldsymbol{\beta}$ near $\boldsymbol{\beta}_0$. Then writing $\boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\beta}) = (\boldsymbol{I}_T - \boldsymbol{p}_i(\boldsymbol{\beta})\mathbf{1}')\boldsymbol{y}_i$, we have for all $\boldsymbol{\beta}$ near $\boldsymbol{\beta}_0$

$$\begin{aligned}E(\nabla_{\boldsymbol{\beta}}(\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta})\boldsymbol{y}_i)|\boldsymbol{x}_i) &= E(\nabla_{\boldsymbol{\beta}}(\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta})\boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\beta}))|\boldsymbol{x}_i) \\ &= \nabla_{\boldsymbol{\beta}} \boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta})E(\boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\beta})|\boldsymbol{x}_i) + \boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta})E(\nabla_{\boldsymbol{\beta}} \boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\beta})|\boldsymbol{x}_i)\end{aligned}$$

Evaluating at $\boldsymbol{\beta}_0$ yields $E(\nabla_{\boldsymbol{\beta}} \boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)\boldsymbol{y}_i|\boldsymbol{x}_i) = \boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)\nabla_{\boldsymbol{\beta}} \boldsymbol{m}_i(\boldsymbol{\beta}_0)$ since $E(\boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)|\boldsymbol{x}_i) = \mathbf{0}$ and $E(\nabla_{\boldsymbol{\beta}} \boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)|\boldsymbol{x}_i) = (\boldsymbol{I}_T - \boldsymbol{p}_i(\boldsymbol{\beta}_0)\mathbf{1}')\nabla_{\boldsymbol{\beta}} \boldsymbol{m}_i(\boldsymbol{\beta}_0)$. □

Proof of Lemma 2.3.1

Write $E(\boldsymbol{y}_i\boldsymbol{y}_i'|\boldsymbol{x}_i) = \boldsymbol{\Sigma}_i$. Then for any $T-1\times T$ transformation $\boldsymbol{A}(\boldsymbol{x}_i,\boldsymbol{\beta}_0)$ with rank $T-1$,

$$\begin{aligned} Rank(\boldsymbol{A}(\boldsymbol{x}_i,\boldsymbol{\beta}_0)\boldsymbol{\Sigma}_i\boldsymbol{A}(\boldsymbol{x}_i,\boldsymbol{\beta}_0)) &= Rank((\boldsymbol{A}(\boldsymbol{x}_i,\boldsymbol{\beta}_0)\boldsymbol{\Sigma}_i^{1/2})((\boldsymbol{A}(\boldsymbol{x}_i,\boldsymbol{\beta}_0)\boldsymbol{\Sigma}_i^{1/2})') \\ &= Rank(\boldsymbol{A}(\boldsymbol{x}_i,\boldsymbol{\beta}_0)\boldsymbol{\Sigma}_i^{1/2}) \\ &= Rank(\boldsymbol{A}(\boldsymbol{x}_i,\boldsymbol{\beta}_0)) = T-1 \end{aligned}$$

as $\boldsymbol{\Sigma}_i^{1/2}$ is $T\times T$ and full rank. Thus the conditional variance is nonsingular and (2.2.4) holds with a proper inverse. Any generalized differencing residual with transformation satisfying Assumption RK.1 has a nonsingular conditional variance. This result goes for $\boldsymbol{Q}(\boldsymbol{I}_T - \boldsymbol{p}_i(\boldsymbol{\beta}_0)\mathbf{1}')$ and $\boldsymbol{Q}(\boldsymbol{I}_T - \boldsymbol{m}_i(\boldsymbol{\beta}_0)(\boldsymbol{m}_i(\boldsymbol{\beta}_0)'\boldsymbol{m}_i(\boldsymbol{\beta}_0))^{-1}\boldsymbol{m}_i(\boldsymbol{\beta}_0)$ since their full transformations have rank $T-1$. Lemma 1 of Verdier (2018) shows $Rank((\boldsymbol{I}_T - \boldsymbol{p}_i(\boldsymbol{\beta}_0)\mathbf{1}')) = T-1$; the rank of the residual maker transformation is a well-known result.

First note that $\boldsymbol{V}_i^-\boldsymbol{m}_i(\boldsymbol{\beta}_0) = \mathbf{0}$ by construction. As

$$\begin{aligned} &\boldsymbol{p}_i(\boldsymbol{\beta}_0)\mathbf{1}'(\boldsymbol{I}_T - \frac{1}{a_i}\boldsymbol{m}_i(\boldsymbol{\beta}_0)\boldsymbol{m}_i(\boldsymbol{\beta}_0)'\boldsymbol{\Sigma}_i^{-1}) = \mathbf{0} \\ &(\boldsymbol{I}_T - \boldsymbol{m}_i(\boldsymbol{\beta}_0)(\boldsymbol{m}_i(\boldsymbol{\beta}_0)'\boldsymbol{m}_i(\boldsymbol{\beta}_0))^{-1}\boldsymbol{m}_i(\boldsymbol{\beta}_0)')\boldsymbol{m}_i(\boldsymbol{\beta}_0) = \mathbf{0} \end{aligned}$$

the conditional gradients are given as

$$\begin{aligned} &(\boldsymbol{I}_T - \boldsymbol{p}_i(\boldsymbol{\beta}_0)\mathbf{1}')\nabla_{\boldsymbol{\beta}}\boldsymbol{m}_i(\boldsymbol{\beta}_0) \\ &(\boldsymbol{I}_T - \boldsymbol{m}_i(\boldsymbol{\beta}_0)(\boldsymbol{m}_i(\boldsymbol{\beta}_0)'\boldsymbol{m}_i(\boldsymbol{\beta}_0))^{-1}\boldsymbol{m}_i(\boldsymbol{\beta}_0)')\nabla_{\boldsymbol{\beta}}\boldsymbol{m}_i(\boldsymbol{\beta}_0) \end{aligned}$$

by Lemma 2.2.3. Then the systems defined by Assumption SYS for both transformations are consistent with $\boldsymbol{F}(\boldsymbol{x}_i) = \boldsymbol{V}_i^-\nabla_{\boldsymbol{\beta}}\boldsymbol{m}_i(\boldsymbol{\beta}_0)$ and the singularity assumption in Assumption RK.2 guarantees both efficiency bounds exist. □

Proof of Theorem 2.3.1

As mentioned in the text, Assumptions CM, RK.1, RK.2, and the positive definiteness of $E(\boldsymbol{y}_i\boldsymbol{y}_i'|\boldsymbol{x}_i)$ are sufficient for each of the transformations studied to satisfy Assumptions SYS and ORTH (and thus MAT) so that their asymptotic efficiency bounds are well-defined and given by

(2.2.8). Let $\boldsymbol{B}_i$ be one of the full rank $T - 1 \times T$ transformation (evaluated at $\boldsymbol{x}_i$ and $\boldsymbol{\beta}_0$) studied. $\boldsymbol{B}_i$ could be the generalized within transformation, or either the generalized within or residual maker transformation with any arbitrary row deleted. I will prove the theorem by showing each of these transformations are information equivalent to the full generalized within transformation via Theorem 1, and noting that a similar proof holds for the full residual maker transformation. Write $\boldsymbol{\Sigma}_i = E(\boldsymbol{y}_i\boldsymbol{y}_i'|\boldsymbol{x}_i)$. Since each of the potential $\boldsymbol{B}_i$ matrices satisfy Assumption ORTH, its efficiency bound is given by (2.2.8):

$$E(\nabla_{\boldsymbol{\beta}}\boldsymbol{m}_i(\boldsymbol{\beta}_0)'\boldsymbol{B}_i'(\boldsymbol{B}_i\boldsymbol{\Sigma}_i\boldsymbol{B}_i')^{-1}\boldsymbol{B}_i\nabla_{\boldsymbol{\beta}}\boldsymbol{m}_i(\boldsymbol{\beta}_0))^{-1}$$

In the notation of Theorem 1, let $\boldsymbol{V}_i = (\boldsymbol{I}_T - \boldsymbol{p}_i(\boldsymbol{\beta}_0)\boldsymbol{1}')\boldsymbol{\Sigma}_i(\boldsymbol{I}_T - \boldsymbol{1}\boldsymbol{p}_i(\boldsymbol{\beta}_0)')$ and $\boldsymbol{M}_i = (\boldsymbol{I}_T - \boldsymbol{p}_i(\boldsymbol{\beta}_0)\boldsymbol{1}')$.

$\boldsymbol{B}_i\boldsymbol{M}_i = \boldsymbol{B}_i$ as $\boldsymbol{B}_i\boldsymbol{p}_i(\boldsymbol{\beta}_0) = \boldsymbol{0}$ by Assumption CM. Also $Rank(\boldsymbol{M}_i\boldsymbol{V}_i\boldsymbol{M}_i') = Rank(\boldsymbol{V}_i) = T - 1 = Rank(\boldsymbol{M}_i)$, so Assumption GR.1 holds for the same $\boldsymbol{M}_i$ regardless of $\boldsymbol{B}_i$. As $\boldsymbol{B}_i\boldsymbol{V}_i\boldsymbol{B}_i' = \boldsymbol{B}_i\boldsymbol{\Sigma}_i\boldsymbol{B}_i'$, we have $Rank(\boldsymbol{B}_i\boldsymbol{V}_i\boldsymbol{B}_i) = T - 1 = Rank(\boldsymbol{B}_i)$, so Assumption GR.2 holds. Thus by Theorem 1 $\boldsymbol{B}_i'(\boldsymbol{B}_i\boldsymbol{\Sigma}_i\boldsymbol{B}_i')^{-1}\boldsymbol{B}_i = \boldsymbol{M}_i'(\boldsymbol{M}_i\boldsymbol{\Sigma}_i\boldsymbol{M}_i')^{-}\boldsymbol{M}_i$. The information bound for the generalized within transformation is

$$E(\nabla_{\boldsymbol{\beta}}\boldsymbol{m}_i(\boldsymbol{\beta}_0)\boldsymbol{M}_i'(\boldsymbol{M}_i\boldsymbol{\Sigma}_i\boldsymbol{M}_i')^{-}\boldsymbol{M}_i\nabla_{\boldsymbol{\beta}}\boldsymbol{m}_i(\boldsymbol{\beta}_0))^{-1}$$

This expression is equal to the expression in (2.2.6) by Theorem 2.2.1, so the generalized within transformation is information equivalent to $\boldsymbol{B}_i$. The proof for the residual maker transformation is similar with $\boldsymbol{M}_i = (\boldsymbol{I}_T - \boldsymbol{m}_i(\boldsymbol{\beta}_0)(\boldsymbol{m}_i(\boldsymbol{\beta}_0)'\boldsymbol{m}_i(\boldsymbol{\beta}_0))^{-1}\boldsymbol{m}_i(\boldsymbol{\beta}_0)')$ and $\boldsymbol{V}_i$ being the respective conditional covariance matrix. □

APPENDIX

PROOFS FOR CHAPTER 3

Proof of Lemma 3.3.1

Separate the estimated parameters into the respective $T - p \times p - p_0$ and $T - p \times p_0$ matrices $(\mathbf{\Theta}^1|\mathbf{\Theta}^2)$. Separate the true regularized parameters by rows $(\mathbf{\Theta}_0^{1\prime}|\mathbf{\Theta}_0^{2\prime})'$, which are then $T - p \times p_0$ and $p - p_0 \times p_0$ matrices, respectively. Then for $p > p_0$, $\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{F}_0 = \mathbf{\Theta}_0^1 + \mathbf{\Theta}_1\mathbf{\Theta}_0^2 - \mathbf{\Theta}_2$. Set $\mathbf{\Theta}_2 = \mathbf{\Theta}_0^1 + \mathbf{\Theta}_1\mathbf{\Theta}_0^2$ for any value of $\mathbf{\Theta}_1$, so that there are infinitely many solutions which make equation (3.3.1) zero. Finally when $p < p_0$ there are too many parameters than can be consistently estimated. Thus there are no values of $\mathbf{\Theta}$ which cause (3.3.1) to be zero. These order conditions for estimation of $\boldsymbol{\theta}_0$ are identical to Ahn et al. (2013). □

Proof of Theorem 3.3.2

I first state the Identifying Assumption (IA) which comes from Ahn et al. (2013)'s Basic Assumptions:

Identifying Assumption: $\text{Rk}(E(\boldsymbol{\gamma}_i\boldsymbol{\gamma}_i')) = p_0 < T$. For any $T \times (T - p_0)$ matrix $\boldsymbol{H}_0$ such that $\text{Rk}(\boldsymbol{F}_0, \boldsymbol{H}_0) = T$, the following matrix has full column rank:

$$\left(E(\boldsymbol{H}_0'\boldsymbol{X}_i \otimes \text{vec}(\boldsymbol{X}_i)), \boldsymbol{I}_{T-p_0} \otimes E(\text{vec}(\boldsymbol{X}_i)\boldsymbol{\gamma}_i')\right) \blacksquare$$

The two equations under consideration are equations (3.2.7) and (3.2.8),

$$E(\boldsymbol{w}_i \otimes \boldsymbol{H}_0'(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_0)) = \boldsymbol{0}$$
$$E(\boldsymbol{H}_0'\boldsymbol{V}_i) = \boldsymbol{0}$$

I appeal to the partial redundancy results given in Section 4 of Breusch et al. (1999). In this setting, partial redundancy of two sets of moment conditions means that the asymptotic variance of the GMM estimator of $\boldsymbol{\beta}_0$ based off of both sets of moment conditions is the same as that of the GMM estimator which only uses the first set. See Section 1 of Breusch et al. (1999) for examples.

Write $\boldsymbol{\lambda} = (\boldsymbol{\beta}_0', \boldsymbol{\theta}_0')'$ and let $\boldsymbol{\lambda}_1 = \boldsymbol{\beta}_0$ and $\boldsymbol{\lambda}_2 = \boldsymbol{\theta}_0$. Then $\boldsymbol{\lambda}_1$ is identified by equation (3.2.7) under IA[1] and $\boldsymbol{\lambda}_2$ is identified by equation (3.2.8), both facts I use in the proof. They consider a general vector of moment conditions

$$E(\boldsymbol{g}(\boldsymbol{\lambda}, \boldsymbol{\eta}_i)) = \begin{bmatrix} \boldsymbol{g}_1(\boldsymbol{\lambda}, \boldsymbol{\eta}_i)) \\ \boldsymbol{g}_2(\boldsymbol{\lambda}, \boldsymbol{\eta}_i)) \end{bmatrix} = \boldsymbol{0}$$

where in my notation $\boldsymbol{\eta}_i = (\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{\gamma}_i, \boldsymbol{\Gamma}_i)$, $\boldsymbol{g}_1 = \boldsymbol{H}(\boldsymbol{\theta})'(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_0 + \boldsymbol{F}\boldsymbol{\gamma}_i)$, and $\boldsymbol{g}_2 = \boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{V}_i$. I partition the gradient and covariances matrices as

$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{D}_{11} & \boldsymbol{D}_{12} \\ \boldsymbol{D}_{21} & \boldsymbol{D}_{22} \end{bmatrix}$$

$$\boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{bmatrix}$$

where $\boldsymbol{D}_{mn} = E(\nabla_{\boldsymbol{\lambda}_n}\boldsymbol{g}_m(\boldsymbol{\lambda}, \boldsymbol{\eta}_i))$ and $\boldsymbol{\Omega}_{mn} = E(\boldsymbol{g}_m(\boldsymbol{\lambda}, \boldsymbol{\eta}_i)\boldsymbol{g}_n(\boldsymbol{\lambda}, \boldsymbol{\eta}_i)')$. Equation (3.2.8) is partially redundant for estimating $\boldsymbol{\beta}_0$ if and only if

$$\boldsymbol{D}_{21} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{D}_{11} = (\boldsymbol{D}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{D}_{12})(\boldsymbol{D}_{12}'\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{D}_{12})^{-1}(\boldsymbol{D}_{12}'\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{D}_{11})$$

by Theorem 7 of Breusch et al. (1999). As $\boldsymbol{u}_i$ is mean independent of $\boldsymbol{X}_i$, $\boldsymbol{\Omega}_{21} = \boldsymbol{0}$ and $\boldsymbol{\Omega}_{12} = \boldsymbol{0}$ so that the necessary and sufficient condition of partial redundancy is

$$\boldsymbol{D}_{21} = \boldsymbol{D}_{22}(\boldsymbol{D}_{12}'\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{D}_{12})^{-1}(\boldsymbol{D}_{12}'\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{D}_{11})$$

Since $\boldsymbol{g}_2(\boldsymbol{\lambda}, \boldsymbol{\eta}_i)$ is not a function of $\boldsymbol{\beta}_0$, we also have $\boldsymbol{D}_{21} = \boldsymbol{0}$. Assumption PF gives that $\boldsymbol{D}_{22}$ has full column rank so that $\boldsymbol{D}_{22}(\boldsymbol{D}_{12}'\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{D}_{12})^{-1}$ is left-invertible. Therefore the redundancy condition becomes

$$\boldsymbol{D}_{12}'\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{D}_{11} = \boldsymbol{0}$$

□

Proof of Theorem 3.3.4

[1] See Section 3 of Ahn et al. (2013).

I start with the proof of consistency. The centered QLDP estimator is written as

$$\widehat{\boldsymbol{\beta}}_{QLDP} - \boldsymbol{\beta}_0 = \left(\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{X}_i'\widehat{\boldsymbol{H}}\widehat{\boldsymbol{H}}'\boldsymbol{X}_i\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{X}_i'\widehat{\boldsymbol{H}}\widehat{\boldsymbol{H}}'(\boldsymbol{F}_0\boldsymbol{\gamma}_i + \boldsymbol{u}_i)\right)$$

The denominator equals its infeasible counterpart $\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{V}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{V}_i$ up to a $O_p(N^{-1/2})$ term by Theorem 1 and the moment bounds from BASE. The inverse exists with probability approaching one by condition (1) of the theorem. Thus the denominator is a $O_p(1)$ term so consistency depends on the numerator.

The difference between the numerator and its infeasible counterpart is

$$\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{X}_i'(\widehat{\boldsymbol{H}}\widehat{\boldsymbol{H}}' - \boldsymbol{H}_0\boldsymbol{H}_0')(\boldsymbol{F}_0\boldsymbol{\gamma}_i + \boldsymbol{u}_i) = \left(\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{F}_0\boldsymbol{\gamma}_i + \boldsymbol{u}_i)' \otimes \boldsymbol{X}_i'\right)\text{vec}(\widehat{\boldsymbol{H}}\widehat{\boldsymbol{H}}' - \boldsymbol{H}_0\boldsymbol{H}_0') = O_p(1)o_p(1)$$

The sum converges to its finite expectation by the moment bounds from Assumption 2(2). $\text{vec}(\widehat{\boldsymbol{H}}\widehat{\boldsymbol{H}}' - \boldsymbol{H}_0\boldsymbol{H}_0') = O_p(N^{-1/2})$ by Theorem 3.3.1. The infeasible numerator, $\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{X}_i'\boldsymbol{H}_0\boldsymbol{H}_0'(\boldsymbol{F}_0\boldsymbol{\gamma}_i + \boldsymbol{u}_i)$, is $o_p(1)$ as $\boldsymbol{H}_0'\boldsymbol{F}_0 = \boldsymbol{0}$ and $\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{X}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{u}_i = o_p(1)$ by condition (3), so we have $\widehat{\boldsymbol{\beta}}_{QLDP} - \boldsymbol{\beta}_0 = o_p(1)$.

Before deriving the asymptotic distribution of the QLDP, I need the following lemma:

Lemma .0.1. *Let $\boldsymbol{\epsilon}_i = \boldsymbol{F}_0\boldsymbol{\gamma}_i + \boldsymbol{u}_i$. Then*

$$\nabla_{\boldsymbol{\theta}}(\boldsymbol{X}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{\epsilon}_i) = \left(\boldsymbol{I}_K \otimes \boldsymbol{u}_i'\boldsymbol{H}_0\right)\begin{pmatrix} {\boldsymbol{x}_i}_1^{*\prime} \otimes \boldsymbol{I}_{T-p_0} \\ \vdots \\ {\boldsymbol{x}_i}_K^{*}{}' \otimes \boldsymbol{I}_{T-p_0} \end{pmatrix} + \boldsymbol{V}_i'\boldsymbol{H}_0\left(\boldsymbol{\epsilon}_i^{*\prime} \otimes \boldsymbol{I}_{T-p_0}\right) \tag{.0.1}$$

where $\boldsymbol{x}_{ij}$ is the j'th column of $\boldsymbol{X}_i$ and $\boldsymbol{v}^* = (v_{T-p_0+1}, ..., v_T)'$ is the last p_0 elements of the $T \times 1$ vector $\boldsymbol{v}$.

Proof. I omit the pure factor notation for simplicity and work with the full matrix $\boldsymbol{X}_i$. Proposition 5.4 of Dhrymes (2013) gives

$$\nabla_{\boldsymbol{\theta}}(\boldsymbol{X}_i'\boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{\epsilon}_i) = (\boldsymbol{\epsilon}_i'\boldsymbol{H}(\boldsymbol{\theta}) \otimes \boldsymbol{I}_K)\nabla_{\boldsymbol{\theta}}(\boldsymbol{X}_i'\boldsymbol{H}(\boldsymbol{\theta})) + \boldsymbol{X}_i'\boldsymbol{H}(\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}}(\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{\epsilon}_i) \tag{.0.2}$$

where I follow standard notation in writing the derivative of the $n \times m$ matrix $\boldsymbol{A}$ with respect to the $k \times 1$ vector $\boldsymbol{\alpha}$ as $\boldsymbol{\nabla_\alpha A} = \nabla_{\alpha}\text{vec}(\boldsymbol{A})$. The row vectors of $\nabla_{\alpha}\boldsymbol{A}$ are then the $1 \times k$ gradient vectors of the elements of vec$(\boldsymbol{A})$ with respect to $\boldsymbol{\alpha}$.

In order to derive the various derivatives, I first start with the case of an arbitrary $T \times 1$ vector $\boldsymbol{v} = (v_1, ..., v_T)'$. As described in Section 3.1, $\boldsymbol{H}(\boldsymbol{\theta})' = (\boldsymbol{I}_{T-p_0}, \boldsymbol{\Theta})$ where $\boldsymbol{\theta} = \text{vec}(\boldsymbol{\Theta})$. I write the p_0 column vectors of $\boldsymbol{\Theta}$ as $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_{p_0})$ where each column can be written as $\boldsymbol{\theta}_j = (\theta_{j1}, ..., \theta_{j,T-p_0})'$. These definitions give the expression

$$\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{v} = \begin{pmatrix} v_1 + \theta_{11}v_{T-p_0+1} + ... + \theta_{p1}v_T \\ \vdots \\ v_{T-p_0} + \theta_{1,T-p_0}v_{T-p_0+1} + ... + \theta_{p,T-p_0}v_T \end{pmatrix} \tag{.0.3}$$

The expression above is similar to that derived below equation (4) of Ahn et al. (2013). They write the terms as the dot product between the rows of $\boldsymbol{H}(\boldsymbol{\theta})'$ and $\boldsymbol{v}^*$. However, I expand the sums so that the gradient is easier to see. Taking the gradient of the r'th element of $\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{v}$ with respect to $\boldsymbol{\theta}_j$ gives

$$\nabla_{\boldsymbol{\theta}_j}(v_r + \theta_{1r}v_{T-p_0+1} + ... + \theta_{p_0r}v_T) = (0, ..., 0, v_{T-p_0+j}, 0, ..., 0)$$

where the only nonzero term is in the r'th column. Thus differentiating with respect to the j'th vector gives

$$\nabla_{\boldsymbol{\theta}_j}\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{v} = \begin{pmatrix} v_{T-p_0+j} & 0 & \dots & 0 \\ 0 & v_{T-p_0+j} & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \dots & \dots & v_{T-p_0+j} \end{pmatrix} = v_{T-p_0+j}\boldsymbol{I}_{T-p_0}$$

Putting together the $T - p_0$ gradients gives

$$\nabla_{\boldsymbol{\theta}}\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{v} = \left(v_{T-p_0+1}\boldsymbol{I}_{T-p_0}, ..., v_T\boldsymbol{I}_{T-p_0}\right) = \boldsymbol{v}^{*\prime} \otimes \boldsymbol{I}_{T-p_0} \tag{.0.4}$$

Equation (.0.4) implies $\nabla_{\boldsymbol{\theta}}\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{\epsilon}_i = \boldsymbol{\epsilon}_i^{*\prime} \otimes \boldsymbol{I}_{T-p_0}$. Handling $\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{X}_i$ is done similarly. Writing the covariates in terms of its column vectors $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{iK})$ where now the subscript on $\boldsymbol{x}_{ik}$ denotes the $T \times 1$ vector of observations for variable k of individual i, we can see that

$$\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{X}_i = (\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{x}_{i1}, ..., \boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{x}_{iK})$$

which implies that

$$\text{vec}(\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{X}_i) = \begin{pmatrix} \boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{x}_{i1} \\ \vdots \\ \boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{x}_{iK} \end{pmatrix}$$

$\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{x}_{ik}$ is a $(T - p_0) \times 1$ vector so its gradient follow the same form as equation (.0.4). Thus

$$\nabla_{\boldsymbol{\theta}}\text{vec}(\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{X}_i) = \begin{pmatrix} \boldsymbol{x}_{i1}^{*\prime} \otimes \boldsymbol{I}_{T-p_0} \\ \vdots \\ \boldsymbol{x}_{iK}^{*\prime} \otimes \boldsymbol{I}_{T-p_0} \end{pmatrix}$$

Filling in the gradient in equation (.0.1) gives our final answer. □

Returning to the main proof of asymptotic normality, the pooled QLD estimator can be written as

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}_{QLDP} - \boldsymbol{\beta}_0) = \left(\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{X}_i'\widehat{\boldsymbol{H}}\widehat{\boldsymbol{H}}'\boldsymbol{X}_i\right)^{-1} \left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \boldsymbol{X}_i'\widehat{\boldsymbol{H}}\widehat{\boldsymbol{H}}'(\boldsymbol{F}_0\boldsymbol{\gamma}_i + \boldsymbol{u}_i)\right)$$

As before, he denominator equals $\boldsymbol{A}_P$ up to a $O_p(N^{-1/2})$. The inverse exists with probability approaching one by condition (1) of the theorem. Thus asymptotic normality depends on the numerator.

Write the full error as $\boldsymbol{\epsilon}_i = \boldsymbol{F}_0\boldsymbol{\gamma}_i + \boldsymbol{u}_i$ so that we study the asymptotic distribution of $\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \boldsymbol{X}_i'\widehat{\boldsymbol{H}}\widehat{\boldsymbol{H}}'\boldsymbol{\epsilon}_i$. Mean value expansion about $\boldsymbol{\theta}_0$ gives

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \boldsymbol{X}_i'\widehat{\boldsymbol{H}}\widehat{\boldsymbol{H}}'\boldsymbol{\epsilon}_i = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \boldsymbol{V}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{u}_i + \boldsymbol{G}_P\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_p(1)$$

where $\boldsymbol{G}_P = E(\nabla_{\boldsymbol{\theta}}\boldsymbol{X}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{\epsilon}_i)$ which is derived explicitly in Lemma .0.1. The estimator $\widehat{\boldsymbol{\theta}}$ is derived in Theorem 3.3.1 as based on the moments $E(\text{vec}(\boldsymbol{H}_0'\boldsymbol{Z}_i) = \boldsymbol{0}$. It is a GMM estimator using the optimal weight matrix $\widehat{\boldsymbol{A}}_{\boldsymbol{\theta}} = \frac{1}{N}\sum_{i=1}^{N}\text{vec}(\tilde{\boldsymbol{H}}'\boldsymbol{Z}_i)\text{vec}(\tilde{\boldsymbol{H}}'\boldsymbol{Z}_i)'$ where $\tilde{\boldsymbol{H}} = \boldsymbol{H}(\tilde{\boldsymbol{\theta}})$ uses an initial estimator. The first order conditions of the GMM optimization problem give

$$\left(\sum_{i=1}^{N}\nabla_{\boldsymbol{\theta}}\text{vec}(\widehat{\boldsymbol{H}}'\boldsymbol{Z}_i)\right)' \widehat{\boldsymbol{A}}_{\boldsymbol{\theta}}^{-1}\left(\sum_{i=1}^{N}\text{vec}(\widehat{\boldsymbol{H}}'\boldsymbol{Z}_i)\right) = \boldsymbol{0}$$

where $\nabla_{\boldsymbol{\theta}}\text{vec}(\widehat{\boldsymbol{H}}'\boldsymbol{Z}_i) = (\boldsymbol{z}_{i,1}^{*} \otimes \boldsymbol{I}_{T-p_0}, ..., \boldsymbol{z}_{i,K+1}^{*} \otimes \boldsymbol{I}_{T-p_0})'$ comes from Lemma 3.3.1. Interestingly, this gradient is free of any parameters and thus the same regardless of the estimator.

Write $\boldsymbol{D}_{\boldsymbol{\theta}} = E(\nabla_{\boldsymbol{\theta}}\text{vec}(\boldsymbol{H}_0'\boldsymbol{Z}_i))$ and $\boldsymbol{A}_{\boldsymbol{\theta}} = E(\text{vec}(\boldsymbol{H}_0'\boldsymbol{Z}_i)\text{vec}(\boldsymbol{H}_0'\boldsymbol{Z}_i)')$, the notation from Theorem 3.3.1. Using another standard mean value expansion gives

$$\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\boldsymbol{D}_{\boldsymbol{\theta}}'\boldsymbol{A}_{\boldsymbol{\theta}}^{-1}\boldsymbol{D}_{\boldsymbol{\theta}})^{-1}\boldsymbol{D}_{\boldsymbol{\theta}}'\boldsymbol{A}_{\boldsymbol{\theta}}^{-1}\text{vec}(\boldsymbol{H}_0'\boldsymbol{Z}_i) + o_p(1) \qquad (.0.5)$$

which allows us to write the estimator as

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}_{QLDP} - \boldsymbol{\beta}_0) = \boldsymbol{A}_P^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\boldsymbol{V}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{u}_i + \boldsymbol{G}_P\boldsymbol{r}_i(\boldsymbol{\theta}_0)\right) + o_p(1) \qquad (.0.6)$$

where $\boldsymbol{r}_i(\boldsymbol{\theta}_0) = (\boldsymbol{D}_{\boldsymbol{\theta}}'\boldsymbol{A}_{\boldsymbol{\theta}}^{-1}\boldsymbol{D}_{\boldsymbol{\theta}})^{-1}\boldsymbol{D}_{\boldsymbol{\theta}}'\boldsymbol{A}_{\boldsymbol{\theta}}^{-1}\text{vec}(\boldsymbol{H}_0'\boldsymbol{Z}_i)$. Thus we have

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}_{QLDP} - \boldsymbol{\beta}_0) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{A}_P^{-1}\boldsymbol{B}_P\boldsymbol{A}_P^{-1}) \qquad (.0.7)$$

where $\boldsymbol{B}_P = E((\boldsymbol{V}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{u}_i + \boldsymbol{G}_P\boldsymbol{r}_i(\boldsymbol{\theta}_0))(\boldsymbol{V}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{u}_i + \boldsymbol{G}_P\boldsymbol{r}_i(\boldsymbol{\theta}_0))')$. □

Proof of Theorem 3.4.2

Now the asymptotic variance depends only on the moments $E(\boldsymbol{H}_0'\boldsymbol{V}_i) = \boldsymbol{0}$.

Lemma .0.2. *Suppose Assumption 2 holds and $Rk(E(\boldsymbol{\Gamma}_i)) = p_0$ and let $\widehat{\boldsymbol{\theta}}$ be the GMM estimator based off of $E(vec(\boldsymbol{H}_0'\boldsymbol{X}_i)) = E(vec(\boldsymbol{H}_0'\boldsymbol{V}_i) = \boldsymbol{0}$ using a consistent estimator of the optimal weight matrix. Then*

$$\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(\boldsymbol{0}, \left(\boldsymbol{D}_{x,\boldsymbol{\theta}}'\boldsymbol{A}_{x,\boldsymbol{\theta}}^{-1}\boldsymbol{D}_{x,\boldsymbol{\theta}}\right)^{-1}).$$

and $\boldsymbol{r}_{x,i}(\boldsymbol{\theta}_0) = (\boldsymbol{D}_{x,\boldsymbol{\theta}}'\boldsymbol{A}_{x,\boldsymbol{\theta}}^{-1}\boldsymbol{D}_{x,\boldsymbol{\theta}})^{-1}\boldsymbol{D}_{x,\boldsymbol{\theta}}'\boldsymbol{A}_{x,\boldsymbol{\theta}}^{-1}vec(\boldsymbol{H}_0'\boldsymbol{V}_i)$, where $\boldsymbol{A}_{x,\boldsymbol{\theta}} = E(vec(\boldsymbol{H}_0'\boldsymbol{V}_i)vec(\boldsymbol{H}_0'\boldsymbol{V}_i)')$ and $\boldsymbol{D}_{x,\boldsymbol{\theta}} = E(\nabla_{\boldsymbol{\theta}}vec(\boldsymbol{H}_0'\boldsymbol{V}_i))$ is derived in Lemma .0.1.

□

Proof of Theorem 3.4.3

I first consider the proof of consistency. Facts about uniform convergence shown for consistency will be taken for granted in the proof of asymptotic normality.

As a technical aside, I do not differentiate between the Euclidean vector norm and the Frobenius matrix norm in terms of notation. It does not affect the proof as the two norms are compatible in the sense that $\|\boldsymbol{A}\boldsymbol{x}\|_E \leq \|\boldsymbol{A}\|_F \|\boldsymbol{x}\|_E$ where $\boldsymbol{A}$ is a $n \times m$ matrix, $\boldsymbol{x}$ is a $m \times 1$ vector, and the

F and E subscripts refer to Frobenius and Euclidean respectively. Further, since both norms are submultiplicative, it does not matter for the point of this proof. As such the notation should be clear from the context. Finally, all statements involving random quantities are assumed to hold almost surely unless stated otherwise.

The QDMG estimator can be written as

$$\begin{aligned}(\widehat{\boldsymbol{\beta}}_{QLDMG} - \boldsymbol{\beta}_0) &= \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{X}_i'\widehat{\boldsymbol{H}}\widehat{\boldsymbol{H}}'\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'\widehat{\boldsymbol{H}}\widehat{\boldsymbol{H}}'(\boldsymbol{F}_0\boldsymbol{\gamma}_i + \boldsymbol{u}_i) + \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{b}_i \\ &= \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{X}_i'\widehat{\boldsymbol{H}}\widehat{\boldsymbol{H}}'\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'\widehat{\boldsymbol{H}}\widehat{\boldsymbol{H}}'(\boldsymbol{F}_0\boldsymbol{\gamma}_i + \boldsymbol{u}_i) + O_p(N^{-1/2})\end{aligned}$$

where $\widehat{\boldsymbol{H}} = \boldsymbol{H}(\widehat{\boldsymbol{\theta}})$, $\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$ by Theorem 1. As $\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{b}_i = O_p(N^{-1/2})$ by the CLT, consistency of the QLDMG does not depend on the correlation between $\boldsymbol{b}_i$ and $(\boldsymbol{X}_i, \boldsymbol{\gamma}_i, \boldsymbol{u}_i)$. However, since the rate of convergence is $\sqrt{N}$, it will affect the asymptotic distribution. This fact is handled later in the proof.

I write $\boldsymbol{Z}_i(\boldsymbol{\theta}) = (\boldsymbol{X}_i'\boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'\boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{H}(\boldsymbol{\theta})'(\boldsymbol{F}_0\boldsymbol{\gamma}_i + \boldsymbol{u}_i)$ for convenience. The goal of this section is to show that

$$\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{Z}_i(\widehat{\boldsymbol{\theta}}) \xrightarrow{p} E(\boldsymbol{Z}_i(\boldsymbol{\theta}_0)) = \boldsymbol{0} \tag{.0.8}$$

By Theorem 21.6 of Davidson (1994), the convergence result in equation (.0.8) is implied by conditions:

$$\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0 \tag{.0.9}$$

$$\sup_{\theta \in B_0}\left\|\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{Z}_i(\boldsymbol{\theta}) - E(\boldsymbol{Z}_i(\boldsymbol{\theta}))\right\| = o_p(1) \text{ where } \boldsymbol{B}_0 \text{ is some open set about } \boldsymbol{\theta}_0. \tag{.0.10}$$

where $||.||$ denotes the Euclidean L^2 norm for vectors and Frobenius norm for matrices. Consistency of $\widehat{\boldsymbol{\theta}}$ holds by Theorem 1 so that uniform convergence is the only condition which needs to be verified. I show uniform convergence via a traditional argument which demonstrates both pointwise convergence in probability and stochastic equicontinuity (SE).

Pointwise convergence in probability follows from the WLLN by the moment bounds and sampling assumptions in Assumption 3(2). $\{\boldsymbol{X}_i'\boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{X}_i\}_{i\geq 1}$ is a sequence of positive definite random matrices for all possible values of $\boldsymbol{\theta}$ by condition (1) of the theorem. Thus for each $\boldsymbol{\theta}$, $\{\boldsymbol{Z}_i(\boldsymbol{\theta})\}_{i\geq 1}$ is well-defined and iid. By the WLLN, $\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{Z}_i(\boldsymbol{\theta}) \xrightarrow{p} E(\boldsymbol{Z}_i(\boldsymbol{\theta}))$ which is $\boldsymbol{0}$ when $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

For the purpose of verifying SE of the random sequence, I show that the following Lipschitz condition of Theorem 21.11 from Davidson (1994) holds: for some random sequence $\{B_{Ni}\}_{i\geq 1}$ with bounded expectations and real function h such that $h(x) \longrightarrow 0$ as $x \longrightarrow 0$, there exists $n \in \mathbb{N}$ such that

$$\frac{1}{N}\left\|(\boldsymbol{Z}_i(\boldsymbol{\theta}) - E(\boldsymbol{Z}_i(\boldsymbol{\theta}))) - (\boldsymbol{Z}_i(\dot{\boldsymbol{\theta}}) - E(\boldsymbol{Z}_i(\dot{\boldsymbol{\theta}})))\right\| \leq B_{Ni}h(\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|) \tag{.0.11}$$

for all $\boldsymbol{\theta}, \dot{\boldsymbol{\theta}} \in \mathcal{T}$ and $N \geq n$, where all stated inequalities hold almost surely as stated above.

I start with the stochastic component $\boldsymbol{Z}_i(\boldsymbol{\theta}) - \boldsymbol{Z}_i(\dot{\boldsymbol{\theta}})$. It will make sense to write $\boldsymbol{Z}_i(\boldsymbol{\theta}) = \boldsymbol{A}(\boldsymbol{\theta})^{-1}\boldsymbol{B}(\boldsymbol{\theta})$ where

$$\boldsymbol{A}_i(\boldsymbol{\theta}) = \boldsymbol{X}_i'\boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{X}_i$$
$$\boldsymbol{B}_i(\boldsymbol{\theta}) = \boldsymbol{X}_i'\boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{H}(\boldsymbol{\theta})'(\boldsymbol{F}_0\boldsymbol{\gamma}_i + \boldsymbol{u}_i)$$

We then have

$$\begin{aligned}\left\|\boldsymbol{Z}_i(\boldsymbol{\theta}) - \boldsymbol{Z}_i(\dot{\boldsymbol{\theta}})\right\| &= \left\|\boldsymbol{A}_i(\boldsymbol{\theta})^{-1}\boldsymbol{B}_i(\boldsymbol{\theta}) - \boldsymbol{A}_i(\dot{\boldsymbol{\theta}})^{-1}\boldsymbol{B}_i(\dot{\boldsymbol{\theta}})\right\| \\ &\leq \left\|\boldsymbol{A}_i(\boldsymbol{\theta})^{-1}\boldsymbol{B}_i(\boldsymbol{\theta}) - \boldsymbol{A}_i(\dot{\boldsymbol{\theta}})^{-1}\boldsymbol{B}_i(\boldsymbol{\theta})\right\| + \left\|\boldsymbol{A}_i(\dot{\boldsymbol{\theta}})^{-1}\boldsymbol{B}(\boldsymbol{\theta}) - \boldsymbol{A}_i(\dot{\boldsymbol{\theta}})^{-1}\boldsymbol{B}(\dot{\boldsymbol{\theta}})\right\|\end{aligned}$$

We can bound the second normed value on the right-hand side. Let $\boldsymbol{D}(\boldsymbol{\theta}, \dot{\boldsymbol{\theta}}) = \boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{H}(\boldsymbol{\theta})' - \boldsymbol{H}(\dot{\boldsymbol{\theta}})\boldsymbol{H}(\dot{\boldsymbol{\theta}})'$. The Frobenius norm of a matrix is equal to the square root of the sum of its squared singular values (see, for example, Horn and Johnson (2013)). Thus $\left\|\boldsymbol{A}(\boldsymbol{\theta})^{-1}\right\| = a_i(\boldsymbol{\theta}) > 0$ and we have

$$\begin{aligned}\left\|\boldsymbol{A}_i(\dot{\boldsymbol{\theta}})^{-1}\boldsymbol{B}_i(\boldsymbol{\theta}) - \boldsymbol{A}_i(\dot{\boldsymbol{\theta}})^{-1}\boldsymbol{B}_i(\dot{\boldsymbol{\theta}})\right\| &= \left\|\boldsymbol{A}_i(\dot{\boldsymbol{\theta}})^{-1}(\boldsymbol{B}_i(\boldsymbol{\theta}) - \boldsymbol{B}_i(\dot{\boldsymbol{\theta}}))\right\| \\ &\leq a_i(\dot{\boldsymbol{\theta}})\left\|\boldsymbol{X}_i'\boldsymbol{D}(\boldsymbol{\theta}, \dot{\boldsymbol{\theta}})(\boldsymbol{F}\boldsymbol{\gamma}_i + \boldsymbol{u}_i)\right\| \\ &\leq a_i(\dot{\boldsymbol{\theta}})\|\boldsymbol{X}_i\|\,\|\boldsymbol{F}\boldsymbol{\gamma}_i + \boldsymbol{u}_i\|\left\|\boldsymbol{D}(\boldsymbol{\theta}, \dot{\boldsymbol{\theta}})\right\|\end{aligned}$$

Turning now to the other term from the triangle inequality, note that condition (1) of the theorem implies $\boldsymbol{A}(\boldsymbol{\theta})$ is nonsingular for any $\boldsymbol{\theta}$ in the parameter space. Then

$$\begin{aligned}\left\|\boldsymbol{A}_i(\boldsymbol{\theta})^{-1}\boldsymbol{B}_i(\boldsymbol{\theta}) - \boldsymbol{A}_i(\dot{\boldsymbol{\theta}})^{-1}\boldsymbol{B}_i(\boldsymbol{\theta})\right\| &= \left\|\left(\boldsymbol{A}_i(\boldsymbol{\theta})^{-1} - \boldsymbol{A}_i(\dot{\boldsymbol{\theta}})^{-1}\right)\boldsymbol{B}_i(\boldsymbol{\theta})\right\| \\ &= \left\|\left(\boldsymbol{A}_i(\dot{\boldsymbol{\theta}})^{-1}\boldsymbol{A}_i(\dot{\boldsymbol{\theta}})\boldsymbol{A}_i(\boldsymbol{\theta})^{-1} - \boldsymbol{A}_i(\dot{\boldsymbol{\theta}})^{-1}\boldsymbol{A}_i(\boldsymbol{\theta})\boldsymbol{A}_i(\boldsymbol{\theta})^{-1}\right)\boldsymbol{B}_i(\boldsymbol{\theta})\right\| \\ &= \left\|\boldsymbol{A}_i(\dot{\boldsymbol{\theta}})^{-1}\left(\boldsymbol{A}_i(\dot{\boldsymbol{\theta}}) - \boldsymbol{A}_i(\boldsymbol{\theta})\right)\boldsymbol{A}_i(\boldsymbol{\theta})^{-1}\boldsymbol{B}_i(\boldsymbol{\theta})\right\| \\ &\leq \left\|\boldsymbol{A}_i(\dot{\boldsymbol{\theta}})^{-1}\right\|\left\|\boldsymbol{A}_i(\dot{\boldsymbol{\theta}}) - \boldsymbol{A}_i(\boldsymbol{\theta})\right\|\left\|\boldsymbol{A}_i(\boldsymbol{\theta})^{-1}\right\|\|\boldsymbol{B}_i(\boldsymbol{\theta})\|\end{aligned}$$

As before, $\left\|\boldsymbol{A}_i(\dot{\boldsymbol{\theta}})^{-1}\right\|\left\|\boldsymbol{A}_i(\boldsymbol{\theta})^{-1}\right\| = a_i(\dot{\boldsymbol{\theta}})a_i(\boldsymbol{\theta})$. $\|\boldsymbol{B}_i(\boldsymbol{\theta})\| = \left\|\boldsymbol{X}_i'\boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{H}(\boldsymbol{\theta})'(\boldsymbol{F}\boldsymbol{\gamma}_i + \boldsymbol{u}_i)\right\|$ where $\left\|(\boldsymbol{F}\boldsymbol{\gamma}_i + \boldsymbol{u}_i)\boldsymbol{X}_i'\right\|$ is bounded in expectation.

Condition (3) implies that $\sup_{\boldsymbol{\theta}\in\mathcal{T}}\|\boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{H}(\boldsymbol{\theta})'\| < \tau$ for some $\tau < \infty$. Finally note that

$$\begin{aligned}\left\|\boldsymbol{A}_i(\dot{\boldsymbol{\theta}}) - \boldsymbol{A}_i(\boldsymbol{\theta})\right\| &= \left\|\boldsymbol{X}_i'\boldsymbol{D}(\dot{\boldsymbol{\theta}}, \boldsymbol{\theta})\boldsymbol{X}_i\right\| \\ &\leq \|\boldsymbol{X}_i\|^2\left\|\boldsymbol{D}(\boldsymbol{\theta}, \dot{\boldsymbol{\theta}})\right\|\end{aligned}$$

as $\boldsymbol{D}(\boldsymbol{\theta}, \dot{\boldsymbol{\theta}}) = -\boldsymbol{D}(\dot{\boldsymbol{\theta}}, \boldsymbol{\theta})$. Putting everything together yields

$$\frac{1}{N}\left\|\boldsymbol{Z}_i(\boldsymbol{\theta}) - \boldsymbol{Z}_i(\dot{\boldsymbol{\theta}})\right\| \leq \frac{1}{N}\left(a_i(\dot{\boldsymbol{\theta}})\|\boldsymbol{X}_i\|\|(\boldsymbol{F}_0\boldsymbol{\gamma}_i + \boldsymbol{u}_i)\| + \tau a_i(\dot{\boldsymbol{\theta}})a_i(\boldsymbol{\theta})\|\boldsymbol{X}_i\|^3\|(\boldsymbol{F}_0\boldsymbol{\gamma}_i + \boldsymbol{u}_i)\|\right)\left\|\boldsymbol{D}(\boldsymbol{\theta}, \dot{\boldsymbol{\theta}})\right\|$$

Clearly $\left\|\boldsymbol{D}(\boldsymbol{\theta}, \dot{\boldsymbol{\theta}})\right\| \to 0$ as $\left\|\boldsymbol{\theta} - \dot{\boldsymbol{\theta}}\right\| \to 0$. In the language of Davidson's Theorem 21.11,

$$\sum_{i=1}^{N} B_{Ni} = \frac{1}{N}\sum_{i=1}^{N}\|\boldsymbol{X}_i\|\|(\boldsymbol{F}_0\boldsymbol{\gamma}_i + \boldsymbol{u}_i)\|a_i(\dot{\boldsymbol{\theta}})\left(1 + \tau a_i(\boldsymbol{\theta})\|\boldsymbol{X}_i\|\right)$$

The random variables here have identical moments by Assumption 2(2) and the bound on $\boldsymbol{a}_i(\boldsymbol{\theta})$ holds uniformly over $\mathcal{T}$ by Condition (2) so that

$$\begin{aligned}E(\sum_{i=1}^{N} B_{Ni}) &= E\left(\|\boldsymbol{X}_i\|\|(\boldsymbol{F}_0\boldsymbol{\gamma}_i + \boldsymbol{u}_i)\|a_i(\dot{\boldsymbol{\theta}})\left(1 + \tau a_i(\boldsymbol{\theta})\|\boldsymbol{X}_i\|\right)\right) \\ &= O(1)\end{aligned}$$

as the expectation is finite. Looking at equation (.0.11), we have

$$\left\|(\boldsymbol{Z}_i(\boldsymbol{\theta}) - E(\boldsymbol{Z}_i(\boldsymbol{\theta}))) - \left(\boldsymbol{Z}_i(\dot{\boldsymbol{\theta}}) - E(\boldsymbol{Z}_i(\dot{\boldsymbol{\theta}}))\right)\right\| \leq \left\|\boldsymbol{Z}_i(\boldsymbol{\theta}) - \boldsymbol{Z}_i(\dot{\boldsymbol{\theta}})\right\| + \left\|E(\boldsymbol{Z}_i(\boldsymbol{\theta}) - \boldsymbol{Z}_i(\dot{\boldsymbol{\theta}}))\right\|$$

As norms are convex, $\left\|E((\boldsymbol{Z}_i(\boldsymbol{\theta}) - \boldsymbol{Z}_i(\dot{\boldsymbol{\theta}}))\right\| \leq E(\left\|\boldsymbol{Z}_i(\boldsymbol{\theta}) - \boldsymbol{Z}_i(\dot{\boldsymbol{\theta}})\right\|)$ which is bounded by the same argument as above. I have thus verified SE and so $\widehat{\boldsymbol{\beta}}_{QLDMG} - \boldsymbol{\beta}_0 = o_p(1)$.

Turning to asymptotic normality, I need a lemma on the mean value expansion of the QLDMG estimator like in Theorem 3.3.4.

Lemma .0.3. *Let $\boldsymbol{\epsilon}_i = \boldsymbol{X}_i\boldsymbol{b}_i + \boldsymbol{F}_0\boldsymbol{\gamma}_i + \boldsymbol{u}_i$. Then*

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}}(\boldsymbol{X}_i\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{\epsilon}_i = &-\left(\boldsymbol{I}_K \otimes \boldsymbol{\epsilon}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{V}_i\right)\left((\boldsymbol{V}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{V}_i)^{-1} \otimes (\boldsymbol{V}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{V}_i)^{-1}\right) \\
&* (\boldsymbol{I}_{K^2} + \boldsymbol{K}_K)(\boldsymbol{I}_K \otimes \boldsymbol{V}_i'\boldsymbol{H}_0)\begin{pmatrix} {\boldsymbol{x}_i}_1^{*\prime} \otimes \boldsymbol{I}_{T-p_0} \\ \vdots \\ {\boldsymbol{x}_i}_K^{*\prime} \otimes \boldsymbol{I}_{T-p_0} \end{pmatrix} + \\
&+ (\boldsymbol{V}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{V}_i)^{-1}(\boldsymbol{I}_K \otimes \boldsymbol{\epsilon}_i'\boldsymbol{H}_0)\begin{pmatrix} {\boldsymbol{x}_i}_1^{*\prime} \otimes \boldsymbol{I}_{T-p_0} \\ \vdots \\ {\boldsymbol{x}_i}_K^{*\prime} \otimes \boldsymbol{I}_{T-p_0} \end{pmatrix} + \\
&+ (\boldsymbol{V}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{V}_i)^{-1}\boldsymbol{V}_i'\boldsymbol{H}_0\left(\boldsymbol{\epsilon}_i^{*\prime} \otimes \boldsymbol{I}_{T-p_0}\right)
\end{aligned}$$

where $\boldsymbol{K}_K$ is the $K^2 \times K^2$ commutation matrix.

Proof. Like in Lemma .0.1, I omit the factor structure $\boldsymbol{X}_i = \boldsymbol{F}_0\boldsymbol{\Gamma}_i + \boldsymbol{V}_i$ and derive the above form with respect to just $\boldsymbol{X}_i$. The factor structure is substituted in later after the lemma. Assumption 2 and conditions (1) and (2) imply that the inverse of $\boldsymbol{X}_i'\boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{X}_i$ is differentiable about $\boldsymbol{\theta}_0$. Proposition 5.16 of Dhrymes (2013) gives

$$\nabla_{\boldsymbol{\theta}}(\boldsymbol{X}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{X}_i)^{-1} = -\left((\boldsymbol{X}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{X}_i)^{-1} \otimes (\boldsymbol{X}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{X}_i)^{-1}\right)\left(\nabla_{\boldsymbol{\theta}}\boldsymbol{X}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{X}_i\right)$$

The differential of the $\boldsymbol{X}_i'\boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{X}_i$ can be worked out via 13.19(b) of Abadir and Magnus (2013):

$$d\text{vec}(\boldsymbol{X}_i'\boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{X}_i) = (\boldsymbol{I}_{K^2} + \boldsymbol{K}_K)(\boldsymbol{I}_K \otimes \boldsymbol{X}_i'\boldsymbol{H}(\boldsymbol{\theta}))d\text{vec}(\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{X}_i)$$

The associated gradient was worked out in the proof of Theorem 3.3.4. Thus we have

$$\nabla_{\boldsymbol{\theta}}(\boldsymbol{X}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{X}_i)^{-1} = -\left((\boldsymbol{X}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{X}_i)^{-1} \otimes (\boldsymbol{X}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{X}_i)^{-1}\right)(\boldsymbol{I}_{K^2}+\boldsymbol{K}_K)(\boldsymbol{I}_K \otimes \boldsymbol{X}_i'\boldsymbol{H}_0)\begin{pmatrix} {\boldsymbol{x}_i}_1^{*\prime} \otimes \boldsymbol{I}_{T-p_0} \\ \vdots \\ {\boldsymbol{x}_i}_K^{*\prime} \otimes \boldsymbol{I}_{T-p_0} \end{pmatrix}$$

The product rule of the gradient is given in Proposition 5.4 of Dhrymes (2013) and the gradient $\nabla_{\boldsymbol{\theta}}\boldsymbol{X}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{\epsilon}_i$ comes from Lemma .0.1 in the proof of Theorem 3.3.4. □

The $\sqrt{N}$-normalized estimator is

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}_{QLDMG} - \boldsymbol{\beta}_0) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\boldsymbol{X}_i'\widehat{\boldsymbol{H}}\widehat{\boldsymbol{H}}'\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'\widehat{\boldsymbol{H}}\widehat{\boldsymbol{H}}'\boldsymbol{\epsilon}_i$$

where $\boldsymbol{\epsilon}_i = \boldsymbol{X}_i\boldsymbol{b}_i + \boldsymbol{F}_0\boldsymbol{\gamma}_i + \boldsymbol{u}_i$. I write the estimator in terms of its full error because the asymptotic variance generally depends on the correlation between $\boldsymbol{b}_i$ and the other terms. I derive the asymptotic variance in full, with a simpler form under stronger exogeneity conditions. I apply a mean value expansion to the above sum and get

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\boldsymbol{X}_i'\widehat{\boldsymbol{H}}\widehat{\boldsymbol{H}}'\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'\widehat{\boldsymbol{H}}\widehat{\boldsymbol{H}}'\epsilon_i = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\boldsymbol{V}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{V}_i)^{-1}\boldsymbol{V}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{\epsilon}_i + \boldsymbol{G}_{MG}\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_p(1)$$

where $\boldsymbol{G}_{MG}$ comes from Lemma .0.3. Thus

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}_{QLDMG} - \boldsymbol{\beta}_0) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left((\boldsymbol{V}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{V}_i)^{-1}\boldsymbol{V}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\epsilon_i + \boldsymbol{G}_{MG}\boldsymbol{r}_{x,i}(\boldsymbol{\theta}_0)\right) + o_p(1) \qquad (.0.12)$$

where $\boldsymbol{r}_{x,i}(\boldsymbol{\theta}_0) = (\boldsymbol{D}_{x,\boldsymbol{\theta}}'\boldsymbol{A}_{x,\boldsymbol{\theta}}^{-1}\boldsymbol{D}_{x,\boldsymbol{\theta}})^{-1}\boldsymbol{D}_{x,\boldsymbol{\theta}}'\boldsymbol{A}_{x,\boldsymbol{\theta}}^{-1}\text{vec}(\boldsymbol{H}_0'\boldsymbol{V}_i)$ comes from Lemma .0.2. We then have

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}_{QLDMG} - \boldsymbol{\beta}_0) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{B}_{MG}) \qquad (.0.13)$$

where $\boldsymbol{B}_{MG} = Var\left((\boldsymbol{V}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{V}_i)^{-1}\boldsymbol{V}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{\epsilon}_i + \boldsymbol{G}_{MG}\boldsymbol{r}_{x,i}(\boldsymbol{\theta}_0)\right)$. □

APPENDIX

ADDITIONAL TABLES FOR CHAPTER 3

I now present additional simulations comparing the pooled CCE and QLD estimators. Table .1 gives results for $K = 2$ and $p_0 = 2$ but for larger values of T.

Table .1: Pooled estimator, $K = 2$

		Bias		**SD**		**RMSE**	
		CCEP	**QLDP**	**CCEP**	**QLDP**	**CCEP**	**QLDP**
N = 50	**T = 6**	0.0128	0.0074	0.1028	0.0956	0.1036	0.0959
		0.0128	0.0132	0.1019	0.1025	0.1027	0.1034
	T = 7	0.0146	0.0102	0.0994	0.1222	0.1004	0.1226
		0.0150	0.0096	0.0910	0.1191	0.0922	0.1194
	T = 8	0.0105	0.0061	0.0873	0.0886	0.0879	0.0888
		0.0166	0.0086	0.0855	0.0852	0.0871	0.0856
N = 300	**T = 6**	0.0029	0.0015	0.0405	0.0392	0.0406	0.0392
		0.0039	0.0013	0.0416	0.0406	0.0418	0.0406
	T = 7	0.0016	0.0001	0.0376	0.0477	0.0377	0.0477
		0.0021	-0.0001	0.0374	0.0450	0.0374	0.0450
	T = 8	0.0020	0.0009	0.0344	0.0348	0.0344	0.0349
		0.0010	0.0001	0.0345	0.0344	0.0346	0.0344

Both estimators perform poorly when $N = 50$ with CCEP typically outperforming the QLDP in terms of SD for all N and T. Interestingly, QLDP seems to decrease in bias as T gets larger despite the fact that the number of parameters increases linearly in T for fixed p_0. Generally, the differences in bias are small, and CCEP has a smaller RMSE dues to its reduced SD. Table .2 performs the same simulations but for $K = 3$. In these cases, the QLDP has the smaller SD, most likely due to the fact that the additional covariates provide information which the QLD transformation can exploit.

Table .2: Pooled estimators, $K = 3$

		Bias		**SD**		**RMSE**	
K = 3		**CCE**	**QLDP**	**CCE**	**QLDP**	**CCE**	**QLDP**
N = 50	**T = 6**	0.0115	0.0055	0.1174	0.1010	0.1179	0.1012
		0.0207	0.0131	0.1143	0.1024	0.1161	0.1032
		-0.0041	-0.0009	0.1151	0.1001	0.1151	0.1001
	T = 7	0.0184	0.0127	0.0991	0.1255	0.1008	0.1261
		0.0218	0.0079	0.1009	0.1245	0.1033	0.1247
		-0.0054	-0.0022	0.0998	0.1157	0.0999	0.1157
	T = 8	0.0151	0.0122	0.0883	0.0867	0.0896	0.0875
		0.0095	0.0084	0.0896	0.0873	0.0901	0.0877
		0.0015	-0.0041	0.0895	0.0870	0.0895	0.0871
N = 300	**T = 6**	0.0034	0.0024	0.0451	0.0374	0.0452	0.0375
		-0.0001	0.0007	0.0468	0.0404	0.0468	0.0404
		0.0001	-0.0016	0.0440	0.0391	0.0440	0.0391
	T = 7	0.0038	0.0021	0.0385	0.0468	0.0387	0.0468
		0.0048	0.0010	0.0381	0.0448	0.0384	0.0448
		0.0005	0.0016	0.0382	0.0461	0.0382	0.0461
	T = 8	0.0005	-0.0002	0.0352	0.0347	0.0352	0.0347
		0.0042	0.0015	0.0364	0.0336	0.0367	0.0336
		0.0000	0.0012	0.0351	0.0344	0.0351	0.0344

BIBLIOGRAPHY

BIBLIOGRAPHY

Abadir, K. M., & Magnus, J. R. (2005). *Matrix algebra* (Vol. 1). Cambridge University Press.

Andrews, D. W. K. (2005). Cross-section regression with common shocks. *Econometrica*, *73*, 1551–1585.

Ahn, S. C. (2015). Comment on 'iv estimation of panels with factor residuals' by d. robertson and v. sarafidis. *Journal of Econometrics*, *185*, 542–544. https://doi.org/10.1016/j.jeconom.2014.12.002

Ahn, S. C., Lee, Y. H., & Schmidt, P. (2013). Panel data models with multiple time-varying individual effects. *Journal of Econometrics*, *174*, 1–14. https://doi.org/10.1016/j.jeconom.2012.12.002

Ahn, S. C., & Schmidt, P. (1997). Efficient estimation of dynamic panel data models: Alternative assumptions and simplified estimation. *Journal of Econometrics*, *76*, 309–321.

Amsler, C., Lee, Y. H., & Schmidt, P. (2009). A survey of stochastic frontier models and likely future developments. *Seoul Journal of Economics*, *22*(1).

Arellano, M., & Bover, O. (1995). Another look at the instrumental variable estimation of error-components models. *Journal of Econometrics*, *68*(1), 29–51.

Arellano, M., Hahn, J. et al. (2005). *Understanding bias in nonlinear panel models: Some recent developments* (tech. rep.). Mimeo, CEMFI.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, *71*(1), 135–171.

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica*, *77*(4), 1229–1279.

Breitung, J., & Hansen, P. (2021). Alternative estimation approaches for the factor augmented panel data model with small t. *Empirical Economics*, *60*, 327–351. https://doi.org/10.1007/s00181-020-01948-7

Breitung, J., & Salish, N. (2021). Estimation of heterogeneous panels with systematic slope variations. *Journal of Econometrics*, *220*, 399–415. https://doi.org/10.1016/j.jeconom.2020.04.007

Breusch, T., Qian, H., Schmidt, P., & Wyhowski, D. J. (1997). Redundancy of moment conditions. *Journal of Econometrics*, *91*.

Brown, N. (2021). Information equivalence among transformations of semiparametric nonlinear panel data models *. https://www.researchgate.net/publication/344047637_Information-equivalence_among_transformations_of_semiparametric_nonlinear_panel_data_models

Brown, N., Schmidt, P., & Wooldridge, J. M. (2021). Simple alternatives to the common correlated effects model. https://doi.org/10.13140/RG.2.2.12655.76969/1

Brown, N. L., & Wooldridge, J. M. (2021). More efficient estimation of multiplicative panel data models in the presence of serial correlation. *Manuscript submitted for publication*.

Campello, M., Galvao, A. F., & Juhl, T. (2019). Testing for slope heterogeneity bias in panel data models. *Journal of Business and Economic Statistics*, *37*, 749–760. https://doi.org/10.1080/07350015.2017.1421545

Castillo, J. C., Mejía, D., & Restrepo, P. (2020). Scarcity without leviathan: The violent effects of cocaine supply shortages in the mexican drug war. *Review of Economics and Statistics*, *102*(2), 269–286.

Chamberlain, G. (1980). Analysis of covariance with qualitative data. *Review of Economic Studies*, *47*, 225–238.

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

Chamberlain, G. (1992). Efficiency bounds for semiparametric regression. *Econometrica: Journal of the Econometric Society*, *60*(3), 567–596.

Chen, M., Fernández-Val, I., & Weidner, M. (2014). Nonlinear factor models for network and panel data. *arXiv preprint arXiv:1412.5647*.

Chudik, A., & Pesaran, M. H. (2015). Common correlated effects estimation of heterogeneous dynamic panel data models with weakly exogenous regressors. *Journal of Econometrics*, *188*, 393–420. https://doi.org/10.1016/j.jeconom.2015.03.007

Davidson, J. (1994, October). *Stochastic limit theory: An introduction for econometricians*. Oxford University Press. https://doi.org/10.1093/0198774036.001.0001

Vos, I. D., & Everaert, G. (2021). Bias-corrected common correlated effects pooled estimation in dynamic panels. *Journal of Business and Economic Statistics*, *39*, 294–306. https://doi.org/10.1080/07350015.2019.1654879

Vos, I. D., & Westerlund, J. (2019). On cce estimation of factor-augmented models when regressors are not linear in the factors. *Economics Letters*, *178*, 5–7. https://doi.org/10.1016/j.econlet.2019.02.001

Dhrymes, P. J. (2013). *Mathematics for econometrics*. Springer Science; Business Media.

Fernández-Val, I., & Weidner, M. (2018). Fixed effects estimation of large-t panel data models. *Annual Review of Economics*, *10*, 109–138.

Fischer, S., Royer, H., & White, C. (2018). The impacts of reduced access to abortion and family planning services on abortions, births, and contraceptive purchases. *Journal of Public Economics*, *167*, 43–68.

Hahn, J. (1997). A note on the efficient semiparametric estimation of some exponential panel models. *Econometric Theory*, *13*(4), 583–588.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, *50*, 1029–1054. https://doi.org/10.2307/1912775

Hardin, J. W., & Hilbe, J. M. (2012). *Generalized estimation equations* (2nd ed.). London: Chapman Hall.

Hausman, J., Hall, B. H., & Griliches, Z. (1984). Econometric models for count data with an application to the patents-r&d relationship. *Econometrica: Journal of the Econometric Society*, *52*(4), 909–938.

Hayakawa, K. (2012). Gmm estimation of short dynamic panel data models with interactive fixed effects. *J. Japan Statist. Soc*, *42*, 109–123.

Hayakawa, K. (2016). Identification problem of gmm estimators for short panel data models with interactive fixed effects. *Economics Letters*, *139*, 22–26. https://doi.org/10.1016/j.econlet.2015.12.012

Horn, R. A., & Johnson, C. R. (2012). *Matrix analysis*. Cambridge University Press.

Hsiao, C. (2018). Panel models with interactive effects. *Journal of Econometrics*, *206*, 645–673. https://doi.org/10.1016/j.jeconom.2018.06.017

Im, K. S., Ahn, S. C., Schmidt, P., & Wooldridge, J. M. (1999). Efficient estimation of panel data models with strictly exogenous explanatory variables. *Journal of Econometrics*, *93*(1), 177–201.

Juhl, T., & Lugovskyy, O. (2014). A test for slope heterogeneity in fixed effects models. *Econometric Reviews*, *33*, 906–935. https://doi.org/10.1080/07474938.2013.806708

Juodis, A., & Sarafidis, V. (2018). Fixed t dynamic panel data estimators with multifactor errors. *Econometric Reviews*, *37*, 893–929. https://doi.org/10.1080/00927872.2016.1178875

Juodis, A., & Sarafidis, V. (2020). A linear estimator for factor-augmented fixed-t panels with endogenous regressors. *Journal of Business and Economic Statistics*. https://doi.org/10.1080/07350015.2020.1766469

Juodis, A., & Sarafidis, V. (2021). An incidental parameters free inference approach for panels with common shocks. *Journal of Econometrics*. https://doi.org/10.1016/j.jeconom.2021.03.011

Karabiyik, H., Reese, S., & Westerlund, J. (2017). On the role of the rank condition in cce estimation of factor-augmented panel regressions. *Journal of Econometrics*, *197*(1), 60–64.

Krapf, M., Ursprung, H. W., & Zimmermann, C. (2017). Parenthood and productivity of highly skilled labor: Evidence from the groves of academe. *Journal of Economic Behavior & Organization*, *140*, 147–175.

Liang, Y., & Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*, 13–22.

McCabe, M. J., & Snyder, C. M. (2014). Identifying the effect of open access on citations using a panel of science journals. *Economic Inquiry*, *52*(4), 1284–1300.

McCabe, M. J., & Snyder, C. M. (2015). Does online availability increase citations? theory and evidence from a panel of economics and business journals. *Review of Economics and Statistics*, *97*(1), 144–165.

McCullagh, P., & Nelder, J. (1989). *Generalized linear models* (2nd ed.). London: Chapman Hall.

Moon, H. R., & Weidner, M. (2015). Linear regression for panel with unknown number of factors as interactive fixed effects. *Econometrica*, *83*, 1543–1579. https://doi.org/10.3982/ecta9382

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica*, *46*, 69–85.

Murtazashvili, I., & Wooldridge, J. M. (2008). Fixed effects instrumental variables estimation in correlated random coefficient panel data models. *Journal of Econometrics*, *142*, 539–552. https://doi.org/10.1016/j.jeconom.2007.09.001

Neal, T. (2015). Estimating heterogeneous coefficients in panel data models with endogenous regressors and common factors.

Newey, W. K. (2001). Conditional moment restrictions in censored and truncated regression models. *Econometric Theory*, *17*(5), 863–888.

Newey, K., & McFadden, D. (1994). Large sample estimation and hypothesis. *Handbook of Econometrics, IV, Edited by RF Engle and DL McFadden*, 2112–2245.

Norkutė, M., Sarafidis, V., Yamagata, T., & Cui, G. (2021). Instrumental variable estimation of dynamic linear panel data models with defactored regressors and a multifactor error structure. *Journal of Econometrics*, *220*, 416–446. https://doi.org/10.1016/j.jeconom.2020.04.008

Papke, L. E. (2005). The effects of spending on test pass rates: Evidence from michigan. *Journal of Public Economics*, *89*, 821–839. https://doi.org/10.1016/j.jpubeco.2004.05.008

Papke, L. E., & Wooldridge, J. M. (2008). Panel data methods for fractional response variables with an application to test pass rates. *Journal of Econometrics*, *145*, 121–133. https://doi.org/10.1016/j.jeconom.2008.05.009

Pesaran, M. H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica*, *74*, 967–1012.

Phillips, R. F. (2020). Quantifying the advantages of forward orthogonal deviations for long time series. *Computational Economics*, *55*(2), 653–672.

Rao, C. R., & Mitra, S. K. Generalized inverse of a matrix and its applications. In: *Proceedings of the sixth berkeley symposium on mathematical statistics and probability, volume 1: Theory of statistics*. The Regents of the University of California. 1972.

Robertson, D., & Sarafidis, V. (2015). Iv estimation of panels with factor residuals. *Journal of Econometrics*, *185*, 526–541. https://doi.org/10.1016/j.jeconom.2014.12.001

Schlenker, W., & Walker, W. R. (2016). Airports, air pollution, and contemporaneous health. *The Review of Economic Studies*, *83*(2), 768–809.

Schmidt, P., Ahn, S. C., & Wyhowski, D. (1992). On the estimation of panel-data models with serial correlation when instruments are not strictly exogenous: Comment. *Journal of Business Economic Statistics*, *10*, 10–14. https://doi.org/10.2307/1391796

Sherman, J, & Morrison, W. (1950). Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *Annals of Mathematical Statistics*, *21*, 124–127.

Verdier, V. (2018). Local semi-parametric efficiency of the poisson fixed effects estimator. *Journal of Econometric Methods*, *7*(1).

Westerlund, J. (2019). On estimation and inference in heterogeneous panel regressions with interactive effects. *Journal of Time Series Analysis*, *40*, 852–857. https://doi.org/10.1111/jtsa.12432

Westerlund, J. (2020). A cross-section average-based principal components approach for fixed-t panels. *Journal of Applied Econometrics*, *35*(6), 776–785.

Westerlund, J., Petrova, Y., & Norkutė, M. (2019). Cce in fixed-t panels. *Journal of Applied*

Econometrics, *34*, 746–761. https://doi.org/10.1002/jae.2707

Williams, M. L., Burnap, P., Javed, A., Liu, H., & Ozalp, S. (2020). Hate in the machine: Anti-black and anti-muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology*, *60*(1), 93–117.

Wooldridge, J. M. (1997). Multiplicative panel data models without the strict exogeneity assumption. *Econometric Theory*, *13*(5), 667–678.

Wooldridge, J. M. (1999). Distribution-free estimation of some nonlinear panel data models. *Journal of Econometrics*, *90*(1), 77–97.

Wooldridge, J. M. (2005). Fixed-effects and related estimators for correlated random-coefficient and treatment-effect panel data models. *Source: The Review of Economics and Statistics*, *87*, 385–390. https://about.jstor.org/terms

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data* (2nd ed., Vol. 1). MIT press.