Brown, Nicholas

Working Paper

Moment-based estimation of linear panel data models with factor-augmented errors

Queen's Economics Department Working Paper, No. 1498

Provided in Cooperation with:
Queen's University, Department of Economics (QED)

Suggested Citation: Brown, Nicholas (2023) : Moment-based estimation of linear panel data models with factor-augmented errors, Queen's Economics Department Working Paper, No. 1498, Queen's University, Department of Economics, Kingston (Ontario)

This Version is available at:
https://hdl.handle.net/10419/281102

Queen's Economics Department Working Paper No. 1498

Moment-Based Estimation of Linear Panel Data Models with Factor-Augmented Errors

Nicholas Brown
Queen's University

Department of Economics
Queen's University
94 University Avenue
Kingston, Ontario, Canada
K7L 3N6

2-2023

Moment-Based Estimation of Linear Panel Data Models with Factor-Augmented Errors*

Nicholas Brown†
Department of Economics
Michigan State University

Date of draft: February 3, 2023

Abstract

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 that 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 that allows for a tradeoff between identifying conditions. These estimators do not require the number of independent variables to be less than one minus the number of time periods, a strong restriction when the number of time periods is fixed in the asymptotic analysis. Finally, I investigate the effects of per-student expenditure on standardized test performance using data from the state of Michigan.

JEL classification codes: C36, C38

Keywords: Factor models, common correlated effects, quasi-long differencing, fixed effects, correlated random coefficients.

*I would like to thank Jeffrey Wooldridge and Peter Schmidt for their guidance and advice. I would also like to thank Ben Zou, Nicole Mason-Wardell, Joakim Westerlund, Seung Ahn, Vasilis Sarafidis, Hashem Pesaran, and all participants in the MSU Econometrics Seminar Series. All errors are my own. I have no conflicts of interest to report.

†Current affiliation: Department of Economics, Queen's University, 94 University Ave, Kingston, ON K7L 3N6
E-mail address: n.brown@queensu.ca

1 Introduction

The prevalence of panel data in modern economics has led theorists and practitioners to pay more attention to unobserved and interactive heterogeneity in linear models. 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". Except under highly-specific circumstances, the usual within transformation is insufficient in controlling for these unobserved effects. Previous 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 an additional reduced form model where 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 that estimates unit-specific slopes on the cross-sectional averages of the dependent and independent variables. CCE is similar to a fixed effects treatment that 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. 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 $N \to \infty$.

Despite its theoretical rigor and practicality, the CCE estimator is ad hoc in the sense that it is not derived from the fundamental moment conditions of the model. This fact implies that the pure factor structure in the covariates cannot improve efficiency for the CCE estimator because it is irrelevant from a population information perspective, despite being necessary for consistency. As other fixed-T $\sqrt{N}$-consistent estimators exist that do not require this assumption, it is worth investigating how else the CCE assumptions can be used in estimation. I use the quasi-long-differencing (QLD) transformation of Ahn et al. (2013) to explore the implications of this model and show that the additional ignored CCE pure factor moments are relevant for the estimation of the parameters of interest in the main equation[1].

Ahn et al. (2013) choose a particular normalization of the unobserved factors that induces a smaller set of estimable parameters. They include these parameters in their QLD transformation that can then asymptotically eliminate the space spanned by the factors. While they did not originally assume a pure factor structure in the covariates, I use their transformation to study the CCE model and estimator and demonstrate its shortcomings.

[1]The 'quasi-long-differencing' terminology was not present in the original paper but others have adopted it since; see Juodis and Sarafidis (2018) for example. The name comes from the fact that the transformation subtracts a linear combination of future variables from current ones.

I show that the reduced form pure factor model provides information for estimating the parameters of interest, which is ignored by the pooled CCE estimator. Further, the CCE estimator generally uses more factor proxies than necessary that can lead to inefficiency. Any attempt to reduce the number of proxies is essentially arbitrary. I also demonstrate how the literature's current understanding of the factor loadings causes problems for inference under model misspecification, which I correct for. Finally, the CCE estimator requires more time periods than one plus the number of covariates to be well-defined, a highly restrictive assumption 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. My estimation suggestions will not require this restriction and allow an arbitrary number of covariates.

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[2]. I prove fixed-T consistency and asymptotic normality of my 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 2 discusses the model of interest. Section 3 provides the assumptions that underlie the model and discusses implementation of the QLD-based estimators. Section 4 introduces random slopes. Section 5 provides simulation evidence for the finite-sample properties of the QLD estimators. Section 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 7 concludes with a brief summary and suggestions for future research.

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

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta_0} + \boldsymbol{F}_0\boldsymbol{\gamma}_i + \boldsymbol{u}_i \tag{1}$$

[2] Westerlund and Kaddoura (2022) prove asymptotic normality assuming T is fixed, but they make the same independent random slope assumptions as in Pesaran (2006).

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, and $\boldsymbol{u}_i$ is a $T \times 1$ vector of idiosyncratic shocks. A '0' subscript denotes the true or realized value of an unobserved parameter. The $K \times 1$ vector $\boldsymbol{\beta}_0$ is the object of interest and the factor structure $\boldsymbol{F}_0\boldsymbol{\gamma}_i$ is treated as a collection of nuisance parameters. p_0 is then unobserved because $\boldsymbol{F}_0$ and $\boldsymbol{\gamma}_i$ are unobserved. However, we can consistently test for p_0 so it will be treated as known. Simulation evidence from this paper and others also suggests that overestimating p_0 does not cause inconsistency in QLD estimation. p denotes the number of factors specified by the econometrician.

I define 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 $\boldsymbol{D}_0\boldsymbol{\rho}_i$ that 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 that are uncorrelated with $\boldsymbol{X}_i$. For this reason, I allow the loadings and errors to correlate. The factors are treated as constant. I could alternatively treat them as random and independent of the cross-sectional data like in Westerlund et al. (2019).

2.1 Common Correlated Effects

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

$$\boldsymbol{X}_i = \boldsymbol{F}_0\boldsymbol{\Gamma}_i + \boldsymbol{V}_i \tag{2}$$

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. 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 of $\boldsymbol{\beta}_0$ treats the cross-sectional averages as having unit-specific slopes and can be represented as

$$\widehat{\boldsymbol{\beta}}_{CCEP} = \left(\sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{M}_{\widehat{\boldsymbol{F}}} \boldsymbol{X}_i\right)^{-1} \sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{M}_{\widehat{\boldsymbol{F}}} \boldsymbol{y}_i \tag{3}$$

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 CCEP estimator under the following

intuition: first, write $\boldsymbol{Z}_i = (\boldsymbol{y}_i, \boldsymbol{X}_i)$. The two models in equations (1) and (2) imply

$$E(\boldsymbol{Z}_i) = \boldsymbol{F}_0 E(\boldsymbol{C}_i)\boldsymbol{Q} \tag{4}$$

where $\boldsymbol{C}_i = (\boldsymbol{\gamma}_i, \boldsymbol{\Gamma}_i)$ and $\boldsymbol{Q}$ is constant and positive definite. $\boldsymbol{M}_{\widehat{\boldsymbol{F}}}$ then asymptotically eliminates the space spanned by $\boldsymbol{F}_0$, including $\boldsymbol{F}_0\boldsymbol{\gamma}_i$. All moment conditions are written in terms of the general index i because I assume the data is randomly sampled.

Westerlund et al. (2019) show that $\boldsymbol{M}_{\widehat{\boldsymbol{F}}}$ generally converges to the space orthogonal to both $\boldsymbol{F}_0$ and a random term that is a function of the model's idiosyncratic errors. For the sake of simplicity, suppose that $\boldsymbol{M}_{\widehat{\boldsymbol{F}}} \xrightarrow{p} \boldsymbol{M}_{\boldsymbol{F}_0}$ as is the case when $p_0 = K + 1$. Then the CCEP estimator is based on 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 additional set of moments, they are uninformative for estimating $\boldsymbol{\beta}_0$. I use a parameterized QLD transformation to get value from additional CCE moments.

Pesaran (2006) assumes the idiosyncratic errors in both equations are mutually independent and independent over time. He also assumes random sampling of the factor loadings as well as independence between the loadings and the idiosyncratic errors. Westerlund et al. (2019) assume the errors are still mutually independent, but allow arbitrary unconditional serial correlation in both $\boldsymbol{u}_i$ and $\boldsymbol{V}_i$. However, their main departure comes in the factor loadings. They assume the loadings form a constant sequence with no restriction other than a full rank requirement on their sums. This assumption allows more general sampling schemes and an arbitrary relationship between $\{\boldsymbol{\gamma}_i\}_{i=1}^{\infty}$ and $\{\boldsymbol{\Gamma}_i\}_{i=1}^{\infty}$.

A particularly harsh restriction of the CCEP estimator is the rank condition required for the denominator. $\boldsymbol{M}_{\widehat{\boldsymbol{F}}}$ is a residual-maker matrix and so it has rank $T-(K+1)$. The restriction $T > K+1$ is practically binding regardless of the asymptotic analysis. Even if T is large in a given sample, it must still bound the number of covariates, which is often large in microeconometric applications. Also, when $K + 1 > p_0$, the CCEP estimator unnecessarily removes variation from the data which could improve precision of the estimator. I address both of these problems in Section 3.2.

2.2 Quasi-long-differencing

Ahn et al. (2013) do not assume the pure factor structure in $\boldsymbol{X}_i$. They start with equation (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 $\boldsymbol{F}_0$ 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 on a row-reduction rotation:

$$\boldsymbol{F}_0 = (\boldsymbol{\Theta}_0', -\boldsymbol{I}_{p_0})' \tag{5}$$

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 $\boldsymbol{F}_0$. It is also not unique as any p^2 normalization can be imposed. In this case, I only assume that the factors are full rank; the normalization chosen merely reflects this fact. The parameters $\boldsymbol{\Theta}_0$ are not interesting by themselves, but allow us to eliminate the factors via a convenient reparameterization.

Given the normalization of the general factor matrix $\boldsymbol{F}_0$ in equation (5), Ahn et al. (2013) define the **quasi-long-differencing (QLD)** matrix[3]

$$\boldsymbol{H}(\boldsymbol{\theta}_0) = \begin{pmatrix} \boldsymbol{I}_{T-p_0} \\ \boldsymbol{\Theta}_0' \end{pmatrix} \tag{6}$$

The pure factor structure in equation (4) can thus be used for estimating the parameters in equation (5). If we assume $\boldsymbol{X}_i = \boldsymbol{F}_0\boldsymbol{\Gamma}_i + \boldsymbol{V}_i$ where $E(\boldsymbol{V}_i) = \boldsymbol{0}$, then

$$E(\boldsymbol{H}(\boldsymbol{\theta}_0)'\boldsymbol{Z}_i) = \boldsymbol{0} \tag{7}$$

where $\boldsymbol{\theta}_0 = \text{vec}(\boldsymbol{\Theta})$. I also define $\boldsymbol{H}_0 = \boldsymbol{H}(\boldsymbol{\theta}_0)$ for notational convenience. I show explicitly in the following section how and when these additional moments are useful for the purpose of identification and efficiency, which demonstrates the usefulness of the QLD transformation in studying the CCE model.

The QLD transformation can also be used to exploit moment conditions implied by assumptions on the loadings. This paper takes a "fixed effects" approach in allowing the factor loadings to be arbitrarily correlated with each other and the idiosyncratic errors. If one wishes to maintain that the factor loadings are constant or

[3]The name comes from equation (4) of Ahn et al. (2013). The s'th element of $\boldsymbol{H}(\boldsymbol{\theta}_0)'\boldsymbol{u}_i$ subtracts a linear combination of the last $(T - p_0)$ elements of $\boldsymbol{u}_i$ from u_{is}.

independent as in Westerlund et al. (2019) and Pesaran (2006), we have the additional moment conditions:

$$E((\boldsymbol{H}_0'\boldsymbol{V}_i) \otimes \boldsymbol{H}_0'(\boldsymbol{y}_i - \boldsymbol{V}_i\boldsymbol{\beta}_0)) = \mathbf{0} \tag{8}$$

$$E((\boldsymbol{H}_0'\boldsymbol{V}_i) \otimes (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_0)) = \mathbf{0} \tag{9}$$

$$E(\boldsymbol{X}_i \otimes \boldsymbol{H}_0'(\boldsymbol{y}_i - \boldsymbol{V}_i\boldsymbol{\beta}_0)) = \mathbf{0} \tag{10}$$

$$E(\boldsymbol{H}_0'(\boldsymbol{y}_i - \boldsymbol{V}_i\boldsymbol{\beta}_0)) = \mathbf{0} \tag{11}$$

$$E(\boldsymbol{H}_0'\boldsymbol{V}_i) = \mathbf{0} \tag{12}$$

Equations (8)-(12) list $(T-p_0)((T-p_0)K+2TK+K+1)$ moment conditions that displays the strength of the CCE assumptions made in current applications. Again, CCEP only uses the moments $E(\boldsymbol{X}_i'\boldsymbol{M}_{\boldsymbol{F}_0}\boldsymbol{u}_i) = \mathbf{0}$. Even if one wanted to work with the CCE estimator, the first four sets of moment conditions are valid because they include $\boldsymbol{\beta}_0$. The last set of moments is irrelevant to CCE because there are no parameters in $E(\boldsymbol{M}_{\boldsymbol{F}_0}\boldsymbol{V}_i) = \mathbf{0}$ to estimate.

Another point of interest comes from the fact that the above equations can all be derived under varying identifying assumptions. Instead of assuming purely fixed (or independent) loadings, we may have reason to believe the loadings in the main equation are independent of the errors in the reduced form equation. Then the second equation is valid, but the third may not be. Further, if we believe the loadings are independent of each other, we could simply demean them and exploit a zero correlation restriction:

$$E((\boldsymbol{X}_i - E(\boldsymbol{X}_i)) \otimes ((\boldsymbol{y}_i - E(\boldsymbol{y}_i)) - (\boldsymbol{X}_i - E(\boldsymbol{X}_i))\boldsymbol{\beta}_0)) = \mathbf{0} \tag{13}$$

where the means can easily be estimated by the usual sample average[4]. This section demonstrates the point that the CCE model implies many unused moment conditions, some of which can be derived under even weaker assumptions than those made in fixed-T CCE analysis.

3 Estimation

I now state this paper's primary assumptions. The first assumption defines the model of interest. The second set specifies the pure factor structure in $\boldsymbol{X}_i$ similar to Westerlund et al. (2019).

Assumption 1 (Linear population model):

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

[4]This set of moments relies on the factor loadings having a common mean. Westerlund et al. (2019) does not make this restriction. However, all work in the current paper is done assuming random sampling.

Assumption 2 (CCE reduced form equations):

(i)$\boldsymbol{X}_i = \boldsymbol{F}_0\boldsymbol{\Gamma}_i + \boldsymbol{V}_i$.

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

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

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

Assumption 1 defines the relevant population model. Assumption 2 specifies the pure factor assumption similar to Pesaran (2006) and Westerlund et al. (2019). Unlike these CCE analyses, I do not require independence between the errors in the main or reduced form equations. In fact, I only restrict $E(\boldsymbol{u}_i|\boldsymbol{V}_i) = \mathbf{0}$ but place no assumptions on the conditional distribution $D(\boldsymbol{V}_i|\boldsymbol{u}_i)$. This assumption allows for heteroskedasticity conditional on both observables and unobservables in both sets of errors which, while common in the fixed-T GMM literature, is ruled out in the CCE approaches of Pesaran (2006) and Westerlund et al. (2019).

I assume the factor loadings are random and iid in the cross section. I could relax this assumption at the cost of notational complexity.Westerlund et al. (2019) show that more general sampling techniques are allowed in the asymptotic analysis[5]. For example, I could replace Assumption 2(iv) with Assumption C of Westerlund et al. (2019). However, I do not assume $(\boldsymbol{\gamma}_i, \boldsymbol{\Gamma}_i)$ is orthogonal to $(\boldsymbol{u}_i, \boldsymbol{V}_i)$. While this assumption is reasonable as the factor structure is supposed to represent correlation between $\boldsymbol{X}_i$ and the full error $\boldsymbol{F}\boldsymbol{\gamma}_i + \boldsymbol{u}_i$, it can fail if the model is misspecified. I show in Section 3.2 that $\sqrt{N}$-consistent estimation is possible even if $\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{V}_i \otimes \boldsymbol{\gamma}_i$ does not converge to zero due to model misspecification.

As discussed earlier, I do not require $T > K + 1$, unlike the CCEP estimator. I directly use the moments $E(\boldsymbol{H}_0'\boldsymbol{Z}_i) = \mathbf{0}$ to remove the factors and only require $K \geq p_0 + 1$, a restriction also made by Pesaran (2006) and Westerlund et al. (2019). As long as there are enough time periods to cover all the unobserved effect, my procedure can allow for an arbitrarily large number of covariates, subject to the usual bounds applied to non-regularized regressions. I also discuss in Section 3.2 how to include known factors like a heterogeneous intercept that decreases the number of relevant factors and makes the assumption even less restrictive.

3.1 CCE Moment Conditions

I now look at the moment conditions implied by Assumption 2. Equation (4) of Section 2, $E(\boldsymbol{H}_0'\boldsymbol{Z}_i) = \mathbf{0}$ where $\boldsymbol{Z}_i = (\boldsymbol{y}_i, \boldsymbol{X}_i)$, implies that Assumption 2 provides information on $\boldsymbol{\theta}_0$ that 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), p is the number of factors specified by the econometrician.

[5]I refer the reader to their online appendix

Lemma 1. *Under Assumption 2, $\boldsymbol{\theta}_0$ is identified by $E(\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{Z}_i) = \mathbf{0}$ if and only if $p = p_0$.*

All proofs are contained in the Appendix.

We can use Lemma 1 to provide an estimator of $\boldsymbol{\theta}_0$ based on the covariates alone. Let $\widehat{\boldsymbol{H}} = \boldsymbol{H}(\widehat{\boldsymbol{\theta}})$, $\boldsymbol{A}_{\boldsymbol{\theta}} = E(\text{vec}(\boldsymbol{H}_0'\boldsymbol{Z}_i)\text{vec}(\boldsymbol{H}_0'\boldsymbol{Z}_i)')$, and $\boldsymbol{D}_{\boldsymbol{\theta}} = E(\nabla_{\boldsymbol{\theta}}\text{vec}(\boldsymbol{H}_0'\boldsymbol{Z}_i))$ where $\nabla_{\boldsymbol{\theta}}$ is the gradient with respect to $\boldsymbol{\theta}$.

Theorem 1. *Suppose Assumption 2 holds, and let $\widehat{\boldsymbol{\theta}}$ be the GMM estimator based on $E(vec(\boldsymbol{H}_0'\boldsymbol{Z}_i)) = \mathbf{0}$ using a consistent estimator of the optimal weight matrix. Then*

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

Now suppose that $\widehat{\boldsymbol{A}}_{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{A}_{\boldsymbol{\theta}}$ using a consistent first-step estimator of $\boldsymbol{\theta}_0$.

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

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

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 that 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 1. This can occur if there are polynomial or interactive functions of the covariates in the estimating equation. 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 1. *When $p_0 = K + 1$, the estimator $\widehat{\boldsymbol{\theta}}$ solves*

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

Corollary 1 provides important robustness properties in Section 3. For now, I point out how Theorem 1 can help test for p_0. There are $(T - p_0)(K + 1)$ moments and $(T - p_0)p_0$ parameters; 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 (ii) of Theorem 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 on the moments in Theorem 1.

I now demonstrate that the additional reduced form moments generally improve efficiency of estimating $\boldsymbol{\beta}_0$ by providing non-redundant moment conditions. The following theorem completely characterizes when the moments $E(\boldsymbol{H}_0'\boldsymbol{X}_i) = E(\boldsymbol{H}_0'\boldsymbol{V}_i) = \mathbf{0}$ are partially redundant for estimating $\boldsymbol{\beta}_0$ using moments given by

applying QLD to the main population equation, meaning its asymptotic variance is the same with or without the additional moments. I do not include $E(\boldsymbol{H}_0'\boldsymbol{y}_i) = \mathbf{0}$ in the reduced form 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 $\boldsymbol{g}_{i2}(\boldsymbol{\theta}) = \boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{X}_i$. Define $\boldsymbol{D}_{11} = E(\nabla_{\boldsymbol{\beta}}\boldsymbol{g}_{i1}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0))$, $\boldsymbol{D}_{12} = E(\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_{i1}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0))$, and $\boldsymbol{\Omega}_{11} = Var(\boldsymbol{g}_{i1}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0))$.

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

$$\boldsymbol{D}_{12}'\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{D}_{11} = \mathbf{0} \tag{14}$$

*Proof.*See Appendix for proof. The extra assumption is only needed so that $(\boldsymbol{\beta}_0', \boldsymbol{\theta}_0')'$ are identified by $E(\boldsymbol{g}_{i1}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)) = \mathbf{0}$ and are equivalent to the Basic Assumptions of Ahn et al. (2013). I assume $E(\boldsymbol{u}_i|\boldsymbol{X}_i) = \mathbf{0}$ whereas Assumption 2 implies the weaker $E(\boldsymbol{u}_i|\boldsymbol{V}_i) = \mathbf{0}$. 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 (14) holds in general, and so the additional moments improve the efficiency of estimating $\boldsymbol{\beta}_0$ among the given class of estimators. Trivial cases where equation (14) holds includes $\boldsymbol{\theta}_0$ being known to the researcher and $p_0 = 0$.

Theorem 2 generally demonstrates that the reduced form equations can be used to improve efficiency for estimators that use just the equation of interest. This class of estimators includes CCEP, which uses $E(\boldsymbol{X}_i'\boldsymbol{M}_{\boldsymbol{F}_0}\boldsymbol{u}_i) = \mathbf{0}$. As mentioned earlier, the CCEP estimator cannot make use of the moments $E(\boldsymbol{M}_{\boldsymbol{F}_0}\boldsymbol{X}_i) = \mathbf{0}$ because there are no additional parameters[6]. While QLD introduces $(T - p_0)p_0$ additional parameters, the extra moments in $E(\boldsymbol{H}_0'\boldsymbol{X}_i) = \mathbf{0}$ will often lead to overidentification and provide benefits in terms of efficiency and testing.

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 on the QLD transformation. They allow for a variety of rank and exogeneity conditions that are especially useful when the researcher includes heterogeneous slopes in the model, like in Section 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

[6]Theorem 2 could easily be rewritten in terms of the CCE transformation, with the result in equation (14) being largely unchanged save for the notation.

the "defactored" data $\widehat{\boldsymbol{H}}'\boldsymbol{y}_i$ and $\widehat{\boldsymbol{H}}'\boldsymbol{X}_i$:

$$\widehat{\boldsymbol{\beta}}_{QLDP} = \left(\sum_{i=1}^{N} \boldsymbol{X}_i'\widehat{\boldsymbol{H}}\widehat{\boldsymbol{H}}'\boldsymbol{X}_i\right)^{-1} \sum_{i=1}^{N} \boldsymbol{X}_i'\widehat{\boldsymbol{H}}\widehat{\boldsymbol{H}}'\boldsymbol{y}_i \tag{15}$$

The **pooled quasi-long-differencing (QLDP)** estimator defined by equation (15) 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 (2021) but not formally studied. The **mean group quasi-long-differencing (QLDMG)** estimator 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:

$$\widehat{\boldsymbol{\beta}}_{QLDMG} = \frac{1}{N}\sum_{i=1}^{N} (\boldsymbol{X}_i'\widehat{\boldsymbol{H}}\widehat{\boldsymbol{H}}'\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'\widehat{\boldsymbol{H}}\widehat{\boldsymbol{H}}'\boldsymbol{y}_i \tag{16}$$

It should be noted that $\widehat{\boldsymbol{H}}'$ can be used to "defactor" any variables that 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 correlation between the random slopes and covariates at the cost of rank assumptions and precision. To see how, note that imposing iid random slopes $\boldsymbol{\beta}_i$ on the population model implies

$$\widehat{\boldsymbol{\beta}}_{QLDMG} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\beta}_i + \frac{1}{N}\sum_{i=1}^{N} (\boldsymbol{X}_i'\widehat{\boldsymbol{H}}\widehat{\boldsymbol{H}}'\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'\widehat{\boldsymbol{H}}\widehat{\boldsymbol{H}}'(\boldsymbol{F}_0\boldsymbol{\gamma}_i + \boldsymbol{u}_i) \tag{17}$$

Then given an appropriate uniform law of large numbers applies to $\widehat{\boldsymbol{H}}$ (as shown in the Appendix), the mean group QLD estimator is consistent for $E(\boldsymbol{\beta}_i)$ regardless of the correlation between $\boldsymbol{X}_i$ and $\boldsymbol{\beta}_i$.

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 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.

As discussed earlier, the QLD transformation does not remove more variation from the data than necessary. The CCE transformation, $\boldsymbol{M}_{\widehat{\boldsymbol{F}}}$, is the same even if the econometrician knows p_0. The QLD transformation efficiently uses information on the number of factors, which is consistently estimable. Simulations in Section 5 demonstrate that the QLDP estimator is often more efficient than the CCEP estimator.

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

(i) $\widehat{\boldsymbol{\beta}}_{QLDP} = \tilde{\boldsymbol{\beta}}$.

(ii) $\tilde{\boldsymbol{\alpha}} = \boldsymbol{0}$.

(iii) $\widehat{\boldsymbol{\epsilon}}_i = \tilde{\boldsymbol{\epsilon}}_i$.

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). The result follows explicitly because the first-stage FOC is $\widehat{\boldsymbol{H}}'[\overline{\boldsymbol{y}}, \overline{\boldsymbol{X}}]$. In fact, the proof in the appendix demonstrates we could drop $\boldsymbol{y}_i$ from the initial estimator of $\boldsymbol{\theta}_0$. This result is novel and only recently shown to apply to CCE estimators where it is crucial to use $\overline{\boldsymbol{X}}$ as a cross-sectional average; including $\overline{\boldsymbol{y}}$ is unnecessary (Brown et al. 2021).

It may appear that Theorem 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) that suggests the same thing when using their GMM estimator. Breitung and Hansen (2021) 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 that 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 5.2.

I now show asymptotic normality for the QLDP 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 the QLDP estimator is asymptotically normal. A similar problem occurs in CCE estimation; see Brown et al. (2021). 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 1, the normalized estimator is

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

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. 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 1.

Theorem 4. *Given Assumptions 1 and 2, suppose that*

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

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

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* $\boldsymbol{G}_P = \boldsymbol{0}$.

Remark (Joint estimation): The two-step procedure is less efficient than joint GMM estimation using $E(\boldsymbol{X}_i'\boldsymbol{H}_0\boldsymbol{H}_0'(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_0)) = \boldsymbol{0}$ and $E(\boldsymbol{H}_0'\boldsymbol{Z}_i) = \boldsymbol{0}$ unless $p = K + 1$; see Prokhorov and Schmidt (2009). However, the $p = K + 1$ case confers the advantage of invariance to common variables from Theorem 3 and appears consistent even when $p_0 < p$. There are also optimization issues involved in joint estimation because the moments that 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 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 (like a constant or time trend), and running estimation based on $(\boldsymbol{y}_i^*, \boldsymbol{X}_i^*) = \boldsymbol{M}(\boldsymbol{y}_i, \boldsymbol{X}_i)$. Further, removing known factors can make the QLDP

estimator more robust. According to Theorem 3, removing a random intercept and setting $p = K + 1$ explicitly nests the popular two-way error structure. ■

Remark (Bootstrap): While I provide analytic inference below, the standard errors can be quite complicated in general. $\sqrt{N}(\widehat{\boldsymbol{\beta}}_{QLDP} - \boldsymbol{\beta}_0)$ 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) that 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 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

$$\widehat{\boldsymbol{G}}_P = \frac{1}{N}\sum_{i=1}^{N}\left((\boldsymbol{I}_K \otimes \widehat{\boldsymbol{\epsilon}}_i'\widehat{\boldsymbol{H}})\begin{pmatrix}\boldsymbol{x}_{i1}^{*\prime}\otimes\boldsymbol{I}_{T-p_0}\\ \vdots \\ \boldsymbol{x}_{iK}^{*\prime}\otimes\boldsymbol{I}_{T-p_0}\end{pmatrix} + \boldsymbol{X}_i'\widehat{\boldsymbol{H}}(\widehat{\boldsymbol{\epsilon}}_i^{*\prime}\otimes\boldsymbol{I}_{T-p_0})\right) \tag{18}$$

$$\boldsymbol{r}_i(\widehat{\boldsymbol{\theta}}) = (\widehat{\boldsymbol{D}}_{\boldsymbol{\theta}}'\widehat{\boldsymbol{A}}_{\boldsymbol{\theta}}^{-1}\widehat{\boldsymbol{D}}_{\boldsymbol{\theta}})^{-1}\widehat{\boldsymbol{D}}_{\boldsymbol{\theta}}'\widehat{\boldsymbol{A}}_{\boldsymbol{\theta}}^{-1}\text{vec}(\widehat{\boldsymbol{H}}'\boldsymbol{Z}_i) \tag{19}$$

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 1 and is derived in the proof of Theorem 4. When $\boldsymbol{G}_P = \boldsymbol{0}$, the standard errors take the usual cluster-robust form, similar to the standard errors derived in Westerlund et al. (2019) where $\widehat{\boldsymbol{H}}\widehat{\boldsymbol{H}}'$ is replaced by $\boldsymbol{M}_{\widehat{\boldsymbol{F}}}$. However, whenever $E(\boldsymbol{\Gamma}_i \otimes \boldsymbol{u}_i) \neq \boldsymbol{0}$ or $E(\boldsymbol{\gamma}_i \otimes \boldsymbol{V}_i) \neq \boldsymbol{0}$ due to model misspecification that does not cause inconsistency, this additional term remains in the asymptotic variance[7].

Even if we assume $\boldsymbol{G}_P = \boldsymbol{0}$ along with conditional homoskedasticity and zero serial correlation in $\boldsymbol{u}_i$, the asymptotic variance will still take the sandwich form, suggesting it is less efficient than CCE[8]. Even in this case, the CCE estimator will also take the sandwich form if $K + 1 > p_0$ by the work in Westerlund et al. (2019). Further, consistency of a GLS-type estimator based on the QLD transformation would follow by an almost

[7]Suppose the researcher believes the outcome variable is a nonlinear function of the factors while the covariates exhibit a pure factor structure. Then the reduced form equations can identify the factors, and the additional assumption $E(\boldsymbol{V}_i|\boldsymbol{u}_i) = \boldsymbol{0}$ (assumed in Westerlund et al. (2019)) guarantees consistency of the QLDP. However, $\boldsymbol{H}_0'\boldsymbol{\epsilon}_i$ is still correlated with $\boldsymbol{X}_i$ and so we will need the fully robust analytic standard errors derived here.

[8]I thank an anonymous referee for pointing this out

identical argument to the proof of Theorem 4. This GLS estimator would be efficient when p_0 is asymptotically known by Theorem 3 of Brown (2022). I discuss these issues in the following section.

3.3 Rotation Invariance

The normalization of the factors in equation (5) is irrelevant with regards to consistency of estimating $\boldsymbol{\beta}_0$. However, they may play a significant role in the finite-sample properties of the resulting GMM and linear estimators. It is also clear from Theorem 4 that the QLDP asymptotic variance depends on $\boldsymbol{H}_0$ which itself depends on the normalization in equation (5). The GMM estimator in Ahn et al. (2013) also suffers from this problem. Harding et al. (2022) discuss the problem of selecting normalizations in empirical work. They provide an estimator that is invariant to the choice of identifying normalization. They also propose an estimator that averages over different normalizations to improve efficiency. I now discuss methods to achieve rotation indeterminacy in the GMM and linear estimators of $\boldsymbol{\beta}_0$.

Instead of using the QLD matrix to construct the QLDP estimator, we could instead use the QLD parameters to construct a direct estimator of the factor space. Let

$$\boldsymbol{M}_{\boldsymbol{F}(\boldsymbol{\theta})} = \boldsymbol{I}_T - \boldsymbol{F}(\boldsymbol{\theta})(\boldsymbol{F}(\boldsymbol{\theta})'\boldsymbol{F}(\boldsymbol{\theta}))^{-1}\boldsymbol{F}(\boldsymbol{\theta})' \tag{20}$$

where $\boldsymbol{F}(\boldsymbol{\theta})$ given by any normalization of $\boldsymbol{F}_0$. Theorem 2.2 of Brown and Butts (2022) demonstrates that as long as the normalization used to generate $\boldsymbol{\theta}$ can be written as $\boldsymbol{F}(\boldsymbol{\theta}) = \boldsymbol{F}_0\boldsymbol{A}$ where $\boldsymbol{A}$ is nonsingular (a rotation), then

$$\boldsymbol{M}_{\boldsymbol{F}(\boldsymbol{\theta})} = \boldsymbol{M}_{\boldsymbol{F}_0} \tag{21}$$

That is, the residual-maker matrices generated by $\boldsymbol{F}_0$ and $\boldsymbol{F}(\boldsymbol{\theta})$ are identical. This result is not surprising because $\boldsymbol{F}_0$ and $\boldsymbol{F}(\boldsymbol{\theta})$ span the same space by construction. However, it provides us with conditions under which the normalization used to identify the factors will not affect estimation of $\boldsymbol{\beta}_0$ if we use the proper transformation.

Consider the following estimator:

$$\tilde{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N} \boldsymbol{X}_i'\boldsymbol{M}_{\boldsymbol{F}(\widehat{\boldsymbol{\theta}})}\boldsymbol{X}_i\right)\left(\sum_{i=1}^{N} \boldsymbol{X}_i'\boldsymbol{M}_{\boldsymbol{F}(\widehat{\boldsymbol{\theta}})}\boldsymbol{y}_i\right) \tag{22}$$

where $\widehat{\boldsymbol{\theta}}$ is estimated as in Theorem 1. We are interested in the asymptotic variance of this estimator. Assuming $E(\boldsymbol{V}_i'\boldsymbol{M}_{\boldsymbol{F}_0}\boldsymbol{V}_i)$ is full rank, it is clear by the proof of Theorem 4 that

$$\sqrt{N}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = E(\boldsymbol{V}_i'\boldsymbol{M}_{\boldsymbol{F}_0}\boldsymbol{V}_i)^{-1}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \boldsymbol{X}_i'\boldsymbol{M}_{\boldsymbol{F}(\widehat{\boldsymbol{\theta}})}\boldsymbol{u}_i\right) + o_p(1) \tag{23}$$

where the denominator contains the infeasible $\boldsymbol{M}_{\boldsymbol{F}_0}$. Under the conditions in Theorem 4 that guarantee first-stage estimation does not affect the asymptotic distribution ($E(\boldsymbol{u}_i \otimes \boldsymbol{\Gamma}_i) = \mathbf{0}$ and $E(\boldsymbol{V}_i \otimes \boldsymbol{\gamma}_i) = \mathbf{0}$), we have

$$\sqrt{N}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = E(\boldsymbol{V}_i' \boldsymbol{M}_{\boldsymbol{F}_0} \boldsymbol{V}_i) \left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \boldsymbol{V}_i' \boldsymbol{M}_{\boldsymbol{F}_0} \boldsymbol{u}_i \right) + o_p(1) \tag{24}$$

which implies that $\tilde{\boldsymbol{\beta}}$ is asymptotically equivalent to the infeasible estimator that treats the factors as known.

One benefit of the QLDP over $\tilde{\boldsymbol{\beta}}$ comes from Theorem 3. There is nothing in the in the first order conditions of the the $\widehat{\boldsymbol{\theta}}$ that imply $\boldsymbol{M}_{\boldsymbol{F}(\widehat{\boldsymbol{\theta}})} \overline{\boldsymbol{X}} = \mathbf{0}$. We may hope to construct a QLD-based estimator that retains this property while also asymptotically invariant to the normalization inherent in estimating $\boldsymbol{\theta}$. Consider the following GLS-type estimator:

$$\widehat{\boldsymbol{\beta}}_{QLDGLS} = \left(\sum_{i=1}^{N} \boldsymbol{X}_i' \widehat{\boldsymbol{H}} (\widehat{\boldsymbol{H}}' \widehat{\boldsymbol{\Omega}}^{-1} \widehat{\boldsymbol{H}})^{-1} \widehat{\boldsymbol{H}}' \boldsymbol{X}_i \right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{X}_i' \widehat{\boldsymbol{H}} (\widehat{\boldsymbol{H}}' \widehat{\boldsymbol{\Omega}}^{-1} \widehat{\boldsymbol{H}})^{-1} \widehat{\boldsymbol{H}}' \boldsymbol{y}_i \right) \tag{25}$$

where $\widehat{\boldsymbol{\Omega}} = \frac{1}{N} \sum_{i=1}^{N} \widehat{\boldsymbol{u}}_i \widehat{\boldsymbol{u}}_i'$ is a consistent estimator of $E(\boldsymbol{u}_i \boldsymbol{u}_i')$ using residuals constructed from an initial first-stage estimator of $\boldsymbol{\beta}_0$. Theorem 4 demonstrates that first-stage estimation of $\boldsymbol{\theta}_0$ does not affect the final estimator under the additional exogeneity conditions assumed in Westerlund et al. (2019). Under the same argument as in Theorem 3, the QLDGLS estimator is the same whether or not unit-invariant variables are included. Further, Theorem 3 of Brown (2022) proves that the infeasible QLDGLS estimator that treats $\boldsymbol{\theta}_0$ as given is algebraically equivalent to the infeasible GLS estimator

$$\left(\sum_{i=1}^{N} \boldsymbol{V}_i' \boldsymbol{M}_{\boldsymbol{F}_0} (\boldsymbol{M}_{\boldsymbol{F}_0} \boldsymbol{\Omega}^{-1} \boldsymbol{M}_{\boldsymbol{F}_0})^{-} \boldsymbol{M}_{\boldsymbol{F}_0} \boldsymbol{V}_i \right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{V}_i' \boldsymbol{M}_{\boldsymbol{F}_0} (\boldsymbol{M}_{\boldsymbol{F}_0} \boldsymbol{\Omega}^{-1} \boldsymbol{M}_{\boldsymbol{F}_0})^{-} \boldsymbol{M}_{\boldsymbol{F}_0} \boldsymbol{V}_i \right) \tag{26}$$

which treats $\boldsymbol{F}_0$ as known. This form is different from a feasible CCE-based GLS estimator when $K + 1 > p_0$ because CCE "overestimates" the factor space. As such, QLDGLS is guaranteed to asymptotically reach the information bound under conditional homoskedasticity when initial estimation of $\boldsymbol{\theta}_0$ does not affect the asymptotic distribution[9].

I now turn to the GMM estimator that incorporates the additional CCE moments. The joint GMM estimator that incorporates the additional CCE moments is defined in Theorem 2, which I repeat here:

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

$$E(\boldsymbol{H}_0' \boldsymbol{Z}_i) = \mathbf{0}$$

where $\boldsymbol{Z}_i = [\boldsymbol{y}_i, \boldsymbol{X}_i]$. One approach to rotation invariance is to just start with the Ahn et al. (2013) moments

[9]This argument requires p_0 being known, which is true asymptotically by the arguments in Ahn et al. (2013).

and use the optimal instruments. They assume $E(\boldsymbol{u}_i|\boldsymbol{w}_i) = \mathbf{0}$ so that the instruments are strictly exogenous. Instead of using only the first moments of $\boldsymbol{w}_i$, we could instead build a GMM estimator based on the moments $E(\boldsymbol{H}_0'\boldsymbol{u}_i|\boldsymbol{w}_i) = \mathbf{0}$ using the instruments

$$E(\nabla \boldsymbol{H}_0'\boldsymbol{u}_i|\boldsymbol{w}_i)Var(\boldsymbol{H}_0'\boldsymbol{u}_i|\boldsymbol{w}_i)^{-1} \tag{27}$$

where ∇ implies the gradient taken with respect to both $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. These are the optimal instruments as derived in Chamberlain (1987). The work in Brown (2022) suggests that such an estimator would be invariant to the normalization chosen to remove the factors. One could either estimate the moments nonparametrically or introduce additional working assumptions to give the moments parametric forms. Using this set of moments implies the additional CCE moments are redundant, but requires stronger conditions to implement.

We could instead identify $\boldsymbol{\theta}_0$ from the CCE moments $E(\boldsymbol{H}_0'\boldsymbol{X}_i) = \mathbf{0}$, but use the residual-maker matrix in equation (20) to eliminate the factors from equation (1). This set of moments would take the form

$$E(\boldsymbol{w}_i \otimes \boldsymbol{M}_{\boldsymbol{F}(\boldsymbol{\theta}_0)}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_0)) = \mathbf{0} \tag{28}$$

where we know that $\boldsymbol{M}_{\boldsymbol{F}(\boldsymbol{\theta}_0)} = \boldsymbol{M}_{\boldsymbol{F}_0}$. Because $\boldsymbol{\theta}_0$ appears in a highly nonlinear fashion above, it ease computational burden to first estimate $\boldsymbol{\theta}_0$ using the CCE moments, then plug it in as a first-step estimator to the moments above. This method may save on computational time at the cost of efficiency[10].

4 Heterogeneous Slopes

I now consider a generalization of the population model in equation (1) that allows for random slopes.

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta}_i + \boldsymbol{F}_0\boldsymbol{\gamma}_i + \boldsymbol{u}_i \tag{29}$$

$$\boldsymbol{\beta}_i = \boldsymbol{\beta}_0 + \boldsymbol{b}_i \tag{30}$$

$$\boldsymbol{b}_i \sim (\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{b}}) \tag{31}$$

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 et al. (2019) consider random slopes in their fixed-T analyses. I summarize this model in the following assumption:

Assumption 3 (Random slopes):

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

[10]The joint and two-step estimators are numerically equivalent if both sets of moments are just-identified (Prokhorov and Schmidt 2009). However, there is no clear guidance on which moments to drop from either set of equations.

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

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

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(iii) 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), 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[11].

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 (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 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 $E(\boldsymbol{Q}_i \boldsymbol{U}_i) = \mathbf{0}$ which necessitates $E(\boldsymbol{\beta}_i' \boldsymbol{v}_{it}) = \mathbf{0}$ 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 on $E(\boldsymbol{H}_0' \boldsymbol{X}_i) = E(\boldsymbol{H}_0' \boldsymbol{V}_i) = \mathbf{0}$ and require $p_0 \leq K$ instead of $K+1$. The robustness result of Theorem 3(i) holds for $p = K$ but parts (ii) and (iii) 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 $E(\boldsymbol{H}_0' \boldsymbol{Z}_i) = \mathbf{0}$ 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 $E(\boldsymbol{H}_0' \boldsymbol{V}_i) = \mathbf{0}$ 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 that 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

$$E(\text{vec}(\boldsymbol{X}_i) \otimes \boldsymbol{H}_0'(\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta}_0)) = E(\text{vec}(\boldsymbol{X}_i) \otimes \boldsymbol{H}_0' \boldsymbol{X}_i \boldsymbol{b}_i) \tag{32}$$

and must now be zero for identification of $(\boldsymbol{\beta}_0', \boldsymbol{\theta}_0')'$.

[11] It should be noted that Chudik and Pesaran (2015), Neal (2015), Norkutė et al. (2021) consider dynamic models that will not translate to the fixed-T mean group analysis. Still, this paper is the first to consider the static case with a factor model in the error.

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 that 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 $E(\boldsymbol{b}_i|\boldsymbol{X}_i) = \boldsymbol{0}$, 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 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 $\boldsymbol{\beta}_0$, allowing for estimation under a broader class of settings. Using the given estimator $\widehat{\boldsymbol{\theta}}$ from Theorem 1, I study the pooled QLD estimator in the context of heterogeneous slopes.

Theorem 5. *Given Assumptions 2 and 3, where* $Rk(E(\boldsymbol{\Gamma}_i)) = p_0 \leq K$, *suppose that*

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

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

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

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

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

The proof is identical to the proof of Theorem 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 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 $E(\boldsymbol{H}_0'\boldsymbol{V}_i) = \boldsymbol{0}$, 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 that 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 $a_i(\boldsymbol{\theta}) = \sqrt{\sum_{k=1}^{K} \sigma_k \left((\boldsymbol{X}_i' \boldsymbol{H}(\boldsymbol{\theta}) \boldsymbol{H}(\boldsymbol{\theta})' \boldsymbol{X}_i)^{-1} \right)}$ where $\{\sigma_k(\boldsymbol{D})\}_{k=1}^{K}$ are the singular values of the $K \times K$ matrix $\boldsymbol{D}$.

Theorem 6. *Given Assumptions 2 and 3, where $Rk(E(\boldsymbol{\Gamma}_i)) = p_0 \leq K$, suppose that*

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

(ii) Uniformly over $\mathcal{T}$,

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

(iii) $\mathcal{T}$ 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 $\boldsymbol{G}_{MG} = \boldsymbol{0}$.

Standard errors are derived similarly to the pooled QLD estimator in Section 3.2. Let

$$\widehat{\boldsymbol{B}} = \frac{1}{N} \sum_{i=1}^{N} \left((\boldsymbol{X}_i' \widehat{\boldsymbol{H}} \widehat{\boldsymbol{H}}' \boldsymbol{X}_i)^{-1} \boldsymbol{X}_i' \widehat{\boldsymbol{H}} \widehat{\boldsymbol{H}}' \widehat{\boldsymbol{\epsilon}}_i \widehat{\boldsymbol{G}}_{MG} \boldsymbol{r}_{x,i}(\widehat{\boldsymbol{\theta}}) \right) \left((\boldsymbol{X}_i' \widehat{\boldsymbol{H}} \widehat{\boldsymbol{H}}' \boldsymbol{X}_i)^{-1} \boldsymbol{X}_i' \widehat{\boldsymbol{H}} \widehat{\boldsymbol{H}}' \widehat{\boldsymbol{\epsilon}}_i \widehat{\boldsymbol{G}}_{MG} \boldsymbol{r}_{x,i}(\widehat{\boldsymbol{\theta}}) \right)' \tag{33}$$

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 3 in the Appendix. The gradient $\boldsymbol{G}_{MG}$ can be estimated via

$$\begin{aligned}
\widehat{\boldsymbol{G}}_{MG} = \frac{1}{N} \sum_{i=1}^{N} & - \left(\boldsymbol{I}_K \otimes \widehat{\boldsymbol{\epsilon}}_i' \widehat{\boldsymbol{H}} \widehat{\boldsymbol{H}}' \boldsymbol{X}_i \right) \left((\boldsymbol{X}_i' \widehat{\boldsymbol{H}} \widehat{\boldsymbol{H}}' \boldsymbol{X}_i)^{-1} \otimes (\boldsymbol{X}_i' \widehat{\boldsymbol{H}} \widehat{\boldsymbol{H}}' \boldsymbol{X}_i)^{-1} \right) (\boldsymbol{I}_{K^2} + \boldsymbol{K}_K)(\boldsymbol{I}_K \otimes \boldsymbol{X}_i' \widehat{\boldsymbol{H}}) \times \\
& \times \begin{pmatrix} \boldsymbol{x}_{i1}^{*\prime} \otimes \boldsymbol{I}_{T-p_0} \\ \vdots \\ \boldsymbol{x}_{iK}^{*\prime} \otimes \boldsymbol{I}_{T-p_0} \end{pmatrix} + \\
& + (\boldsymbol{X}_i' \widehat{\boldsymbol{H}} \widehat{\boldsymbol{H}}' \boldsymbol{X}_i)^{-1} \left(\left(\boldsymbol{I}_K \otimes \widehat{\boldsymbol{\epsilon}}_i' \widehat{\boldsymbol{H}} \right) \begin{pmatrix} \boldsymbol{x}_{i1}^{*\prime} \otimes \boldsymbol{I}_{T-p_0} \\ \vdots \\ \boldsymbol{x}_{iK}^{*\prime} \otimes \boldsymbol{I}_{T-p_0} \end{pmatrix} + \boldsymbol{X}_i' \widehat{\boldsymbol{H}} \left(\widehat{\boldsymbol{\epsilon}}_i^{*\prime} \otimes \boldsymbol{I}_{T-p_0} \right) \right)
\end{aligned}$$

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

As discussed in Section 3.2, Theorem 6 is the first fixed-T proof of asymptotic normality for a mean group estimator that 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 6 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 $T-p$ and 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.

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. 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 $\boldsymbol{\beta}_0 = (1,1)'$, which was picked as a arbitrary value. I do not include random slopes as they would only serve to increase the amount of noise in the model. Theorems 6 and 7 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[12]. As described earlier, since T is small and fixed, it is the factor loadings that cause problems asymptotically and

[12] Additional simulations are available upon request.

not the factors. The loadings on $\boldsymbol{X}_i$ are drawn as

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

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

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

where $\Gamma_{1,1}$ and $\Gamma_{2,2}$ are the upper-left and bottom-right diagonal values of $\boldsymbol{\Gamma}_i$. The errors $\boldsymbol{u}_i$ and $\boldsymbol{V}_{ik}$ $(k = 1, 2)$ are drawn from a multivariate normal distribution with mean $\mathbf{0}_{T\times 1}$ 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 $MVN(\mathbf{0}_{T\times 1}, \boldsymbol{C})$ but are independent of each other and $\boldsymbol{u}_i$. Each simulation study includes 1000 replications.

Table 1 compares the Ahn et al. (2013) estimator both with and without the additional moments $E(\boldsymbol{H}_0'\boldsymbol{Z}_i) = \mathbf{0}$. 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. I report the results for each (N, T) pair for both sets of coefficients, which are equal to one in the DGP.

Table 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	0.0026	-0.0225	0.2997	0.1518	0.2997	0.1535
		0.0781	-0.0196	0.3184	0.1424	0.3279	0.1438
	T = 5	-0.0008	-0.0249	0.3702	0.1694	0.3702	0.1712
		0.2631	-0.0055	0.4922	0.2057	0.5581	0.2058
N = 300	**T = 3**	0.0111	0.0015	0.1057	0.0594	0.1063	0.0594
		0.0020	0.0015	0.0588	0.0597	0.0588	0.0597
	T = 4	0.0033	-0.0020	0.1187	0.0427	0.1188	0.0428
		0.0084	0.0001	0.0749	0.0414	0.0754	0.0414
	T = 5	-0.0126	-0.0016	0.1633	0.0364	0.1638	0.0365
		0.1903	-0.0029	0.4069	0.0367	0.4492	0.0368

Notes. This table presents a set of simulations with 1000 replications. Each table consists of a single data generating process where N and T vary. The two rows for a given pair of N and T are the values associated with estimators of each of the two coefficients. "SD" and "RMSE" are respectively the standard deviation and root mean squared error of the estimators over all replications for a given experiment. 'GMM1' refers to the Ahn et al. (2013) estimator using $\text{vec}(\boldsymbol{X}_i)$ as instruments. 'GMM2' uses these moments, as well as the reduced form moments in equation (7). Both estimators are computed using an optimal weight matrix that is a function of an initial consistent first-stage estimator.

The GMM estimator based on 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. (2013) residual and the additional moments $E(\boldsymbol{H}_0'\boldsymbol{Z}_i) = \mathbf{0}$

is GMM2. GMM1 uses $TK(T-2)$ moments while GMM2 uses an additional $(T-2)K$. The GMM estimator using both sets of moments generally outperforms the original Ahn et al. (2013) estimator in terms of root mean square error, 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 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

Notes. This table presents a set of simulations with 1000 replications. Each table consists of a single data generating process where N and T vary. The two rows for a given pair of N and T are the values associated with estimators of each of the two coefficients. The columns refer to the QLDP estimator that uses the given value of p in the estimation of the parameters $\boldsymbol{\theta}_0$. "SD" and "RMSE" are respectively the standard deviation and root mean squared error of the estimators over all replications for a given experiment.

Table 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 (2021). Underestimating p_0 leads to substantial bias that 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 also consider hypothesis testing for different specifications of p. Using the same model but setting $\boldsymbol{\beta}_0 = (0, 0)'$, I construct the QLDP estimators under $p = 1$, $p = 2$, and $p = 3$, when the true value is $p_0 = 2$. Table 3 includes the average rejection rate for the usual Wald statistics of the individual hypothesis tests $H_0 : \beta_1 = 0$ and $H_0 : \beta_2 = 0$ against the relevant two-sided alternative. I carry out the tests at the 5% level, so the test is considered a rejection if the p-value associated with the Wald statistic (evaluated with a standard normal distribution) is greater than or equal to 0.975. We can see that correct specification and overestimation of p_0

leads to reasonable rejection rates when the null hypothesis is true, especially as N increases for a given T. However, underestimating p_0 gives wildly unrealistic rejection rates due to the bias caused by underestimating p_0 as described by table 2. These results further bolster the simulation evidence of Breitung and Hansen (2021) who study the classical Ahn et al. (2013) GMM estimator.

Table 3: Inference with misspecified p_0

		Reject (x100)		
		p = 1	**p = 2**	**p = 3**
N = 50	**T = 4**	46.30	7.70	7.40
		81.70	7.70	10.20
	T = 5	79.70	7.60	9.60
		78.00	6.40	6.40
	T = 6	18.10	6.90	8.80
		99.00	8.30	8.40
N = 300	**T = 4**	98.10	5.10	4.80
		100.00	4.50	6.60
	T = 5	99.50	4.70	6.60
		99.90	4.90	4.00
	T = 6	41.60	6.20	5.00
		100.00	6.50	5.90

Notes. This table presents a set of simulations with 1000 replications. The DGP is identical to the DGP described at the beginning of the section but with $\boldsymbol{\beta}_0 = (0, 0)'$. The columns correspond to the average rejection rate of the Wald statistic for the hypothesis test $H_0 : \beta_1 = 0$ and $H_0 : \beta_2 = 0$ under the different specifications of p when $p_0 = 2$. The rows within the columns correspond to the rejection rates for the tests of the respective parameters associated with the two covariates, x_{it1} and x_{it2}. I calculate the Wald statistic using the standard errors in Theorem 4 but with $\boldsymbol{G}_P = \mathbf{0}$ due to the nature of the DGP. A p-value is calculated for each statistic using a standard normal cdf. The test is considered a rejection for the test of the given parameter if the p-value is greater than or equal to 0.975. The final value is multiplied by 100.

I now turn to comparison of the QLDP and CCEP estimators. I omit the GMM estimators from table 1 because they are outperformed by the just-identified QLDP in terms of root mean square error. Unexpectedly, the QLDP bias is significantly lower than both GMM estimators, and its standard deviation is often significantly lower, especially when N is smaller.

Table 4 looks at the QLDP estimator compared to the CCEP estimator where the QLD transformation is estimated under $p = p_0 = 2$ when $K = 2$ (returning the the original DGP with $\boldsymbol{\beta}_0 = (1, 1)'$). 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 born out by the SD of the simulations. The QLDP estimator consistently shows a 15%-25% decline in standard deviation over the CCEP 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.

Comparing table 4 to table 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

Table 4: Pooled estimators

		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

Notes. This table presents a set of simulations with 1000 replications. Each table consists of a single data generating process where N and T vary. The two rows for a given pair of N and T are the values associated with estimators of each of the two coefficients. "SD" and "RMSE" are respectively the standard deviation and root mean squared error of the estimators over all replications for a given experiment.

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. Simulations for larger values of T give similar results and are available upon request.

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 (16) 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 eventually require a modern treatment that 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 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 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

Notes. This table presents a set of simulations with 1000 replications. Each table consists of a single data generating process where N and T vary. The two rows for a given pair of N and T are the values associated with estimators of each of the two coefficients. "SD" and "RMSE" are respectively the standard deviation and root mean squared error of the estimators over all replications for a given experiment.

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.
avgrexp	6385.51	1034.94	Average real expenditure per pupil.
lunch	0.2886	0.1616	Fraction of students eligible for free and reduced lunch.
enroll	3112.31	7965.49	Total enrollment.

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" that 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 that 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

$$math4_{it} = c_i + \log(avgrexp_{it})\beta_1 + lunch_{it}\beta_2 + \log(enroll_{it})\beta_3 + \boldsymbol{f}_t'\boldsymbol{\gamma}_i + e_{it} \tag{34}$$

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.

The effect of changes in state-level policy are usually evaluated via difference-in-differences or synthetic control methods. However, these methods require the existence of control groups that share a common outcome variable. Standardized tests in the United States vary across states both in terms of the content they test and their evaluation methods. Therefore modeling and eliminating district-level heterogeneity via QLD and CCE techniques provides a compelling way to isolate the treatment effect of interest. These factor models also account for reasonable economic factors that affect the variable of interest. For example, districts with higher concentrations of a given industry will be uniquely affected by macroeconomic shocks. As school funding came primarily from local property taxes before the fulfillment of the new state policy, heterogeneous responses to economic changes would both affect the level of real spending in the district and correlate to underlying demographic characteristics.

Table 6 provides the p-values for testing the hypothesis $H_0 : p_0 = p$ versus $H_1 : p_0 > p$.

Table 6: 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

Notes. This table presents the p-values from GMM overidentifying tests (as in Theorem 1) using different moment conditions. "RF2" uses only the CCE moment conditions. "GMM1" uses only the Ahn et al. (2013) moments while "GMM2" combines both sets.

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 'RF2' (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 $E(\boldsymbol{H}_0'\boldsymbol{Z}_i) = \mathbf{0}$.

GMM1 is just the Ahn et al. (2013) objective function. GMM2 is the Ahn et al. (2013) objective function with the additional moments $E(\boldsymbol{H}_0'\boldsymbol{Z}_i) = \mathbf{0}$. Finally, RF is just the reduced form moments $E(\boldsymbol{H}_0'\boldsymbol{Z}_i) = \mathbf{0}$. 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 $E(\boldsymbol{H}_0'\boldsymbol{Z}_i) = \mathbf{0}$ 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 CCEP estimator. As $T = 7$ and $K = 3$, the CCEP 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 $p = K + 1 = 4$ 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 tells us that $\widehat{\boldsymbol{\beta}}_{QLDP}$ is invariant to common variables when $p = K + 1$ and simulation evidence in the previous section suggests that overestimating p_0 is not particularly problematic from the perspective of bias or inference. Since it also eliminates a heterogeneous intercept, it will be consistent if TWFE is consistent, assuming strictly exogenous covariates.

I present results in table 7 that show estimation after eliminating a heterogeneous intercept. For CCEP, this simply amounts to $\widehat{\boldsymbol{F}} = (\mathbf{1}, \overline{\boldsymbol{y}}, \overline{\boldsymbol{X}})$. 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, which is significant at the 5% level. 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 CCEP 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 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.

Table 7: 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]

Notes. This table presents results for the different estimators of the coefficients in equation (34). Standard errors for the respective estimates are in parentheses. The numbers in brackets are p-values for the test of significance of the respective coefficient estimates. The CCEP and QLDP estimators both explicitly control for a heterogeneous intercept for the sake of comparison with the TWFE estimator. The CCEP estimator also uses the cross-sectional averages of the outcome and regressors as factor proxies. The QLDP first-stage estimates using the within transformed outcome and regressors and sets $p = K + 1 = 4$.

I also conducted a simulation experiment comparing the pooled QLD estimator to the TWFE estimator under different specifications of an additive error model. These results can be found in the Appendix.

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 CCEP does not use.

Current proofs of fixed-T asymptotic normality of the CCEP estimator assume loadings that 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 CCEP 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 on the QLD transformation and 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, suggesting 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 A: Proofs

Proof of Lemma 1

Assumption 2(iii) implies

$$E(\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{Z}_i) = \boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{F}_0 E(\boldsymbol{C}_i)\boldsymbol{Q} \tag{35}$$

where $\boldsymbol{E}(C_i) = E([\boldsymbol{\gamma}_i, \boldsymbol{\Gamma}_i])$ and $\boldsymbol{Q}$ is given in Section 2.1. $\boldsymbol{Q}$ is nonsingular and $E(\boldsymbol{C}_i)$ has full row rank by Assumption 2(iv), so equation (35) is zero if and only if $\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{F}_0 = \boldsymbol{0}$. When $p = p_0$, $H(\boldsymbol{\theta})'\boldsymbol{F}_0 = \boldsymbol{\Theta}_0 - \boldsymbol{\Theta}$ which is zero if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

Now separate the estimated parameters into the respective $(T-p)\times(p-p_0)$ and $(T-p)\times p_0$ matrices $(\boldsymbol{\Theta}^1|\boldsymbol{\Theta}^2)$. Separate the true regularized parameters by rows $(\boldsymbol{\Theta}_0^{1\prime}|\boldsymbol{\Theta}_0^{2\prime})'$, which are then $(T-p)\times p_0$ and $(p-p_0)\times p_0$ matrices, respectively. Then for $p > p_0$, $\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{F}_0 = \boldsymbol{\Theta}_0^1 + \boldsymbol{\Theta}_1\boldsymbol{\Theta}_0^2 - \boldsymbol{\Theta}_2$. Set $\boldsymbol{\Theta}_2 = \boldsymbol{\Theta}_0^1 + \boldsymbol{\Theta}_1\boldsymbol{\Theta}_0^2$ for any value of $\boldsymbol{\Theta}_1$, so that there are infinitely many solutions that make equation (35) zero. Finally when $p < p_0$ there are too many parameters than can be consistently estimated. Thus there are no values of $\boldsymbol{\Theta}$ that cause (35) to be zero. These order conditions for estimation of $\boldsymbol{\theta}_0$ are identical to Ahn et al. (2013).

□

Proof of Theorem 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:

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

The two equations under consideration are

$$E(\boldsymbol{w}_i \otimes \boldsymbol{H}_0'(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_0)) = \boldsymbol{0} \tag{36}$$

$$E(\boldsymbol{H}_0'\boldsymbol{V}_i) = \boldsymbol{0} \tag{37}$$

I appeal to the partial redundancy results given in Section 4 of Breusch et al. (1997). 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 on 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. (1997) 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 $\boldsymbol{\lambda}$ is identified by equation (36) under IA[13] and $\boldsymbol{\lambda}_2$ is identified by equation (37), both facts I use in the proof. They consider a general vector of moment conditions

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

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 (37) 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. (1997). As $\boldsymbol{u}_i$ is mean independent of $\boldsymbol{X}_i$, $\boldsymbol{\Omega}_{21} = \mathbf{0}$ and $\boldsymbol{\Omega}_{12} = \mathbf{0}$ 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 $\boldsymbol{D}_{21} = \mathbf{0}$. 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

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

□

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

Proof of Theorem 3

By Corollary 1, the first-stage estimator $\widehat{\boldsymbol{\theta}}$ solves $\widehat{\boldsymbol{H}}'[\overline{\boldsymbol{y}}, \overline{\boldsymbol{X}}] = \mathbf{0}$.

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

by Corollary 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,

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

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

□

Proof of Theorem 4

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. The inverse exists with probability approaching one by condition (i) 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

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

The sum converges to its finite expectation by the moment bounds from Assumption 2(ii). $\operatorname{vec}(\widehat{\boldsymbol{H}} \widehat{\boldsymbol{H}}' - \boldsymbol{H}_0 \boldsymbol{H}_0') = O_p(N^{-1/2})$ by Theorem 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 $\boldsymbol{H}_0' \boldsymbol{F}_0 = \mathbf{0}$ 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 (iii), 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 2. *Let* $\boldsymbol{\epsilon}_i = \boldsymbol{F}_0\boldsymbol{\gamma}_i + \boldsymbol{u}_i$. *Then*

$$\nabla_{\boldsymbol{\theta}}(\boldsymbol{X}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{\epsilon}_i) = (\boldsymbol{I}_K \otimes \boldsymbol{u}_i'\boldsymbol{H}_0)\begin{pmatrix} \boldsymbol{x}_{i1}^{*\prime} \otimes \boldsymbol{I}_{T-p_0} \\ \vdots \\ \boldsymbol{x}_{iK}^{*\prime} \otimes \boldsymbol{I}_{T-p_0} \end{pmatrix} + \boldsymbol{V}_i'\boldsymbol{H}_0\left(\boldsymbol{\epsilon}_i^{*\prime} \otimes \boldsymbol{I}_{T-p_0}\right) \tag{38}$$

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

$$\nabla_{\boldsymbol{\theta}}(\boldsymbol{X}_i'\boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{\epsilon}_i) = (\boldsymbol{\epsilon}_i'\boldsymbol{H}(\boldsymbol{\theta}) \otimes \boldsymbol{I}_K)\nabla_{\boldsymbol{\theta}}(\boldsymbol{X}_i'\boldsymbol{H}(\boldsymbol{\theta})) + \boldsymbol{X}_i'\boldsymbol{H}(\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}}(\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{\epsilon}_i) \tag{39}$$

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 $\boldsymbol{\nabla}_{\boldsymbol{\alpha}}\boldsymbol{A} = \nabla_{\boldsymbol{\alpha}}\text{vec}(\boldsymbol{A})$. The row vectors of $\nabla_{\boldsymbol{\alpha}}\boldsymbol{A}$ 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

$$\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{v} = \begin{pmatrix} v_1 + \theta_{11}v_{T-p_0+1} + ... + \theta_{p1}v_T \\ \vdots \\ v_{T-p_0} + \theta_{1,T-p_0}v_{T-p_0+1} + ... + \theta_{p,T-p_0}v_T \end{pmatrix} \tag{40}$$

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

$$\nabla_{\boldsymbol{\theta}} \boldsymbol{H}(\boldsymbol{\theta})' \boldsymbol{v} = (v_{T-p_0+1} \boldsymbol{I}_{T-p_0}, ..., v_T \boldsymbol{I}_{T-p_0}) = \boldsymbol{v}^{*\prime} \otimes \boldsymbol{I}_{T-p_0} \tag{41}$$

Equation (41) 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 (41). 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 (39) 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 (i) 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 $\boldsymbol{G}_P = E(\nabla_{\boldsymbol{\theta}} \boldsymbol{X}_i' \boldsymbol{H}_0 \boldsymbol{H}_0' \boldsymbol{\epsilon}_i)$ and is derived explicitly in Lemma 2. The estimator $\widehat{\boldsymbol{\theta}}$ is derived in Theorem 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 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 1. Using another standard mean value expansion gives

$$\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\boldsymbol{D}_{\boldsymbol{\theta}}'\boldsymbol{A}_{\boldsymbol{\theta}}^{-1}\boldsymbol{D}_{\boldsymbol{\theta}})^{-1}\boldsymbol{D}_{\boldsymbol{\theta}}'\boldsymbol{A}_{\boldsymbol{\theta}}^{-1}\text{vec}(\boldsymbol{H}_0'\boldsymbol{Z}_i) + o_p(1) \tag{42}$$

The derivations above allow me to write the estimator as

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}_{QLDP} - \boldsymbol{\beta}_0) = \boldsymbol{A}_P^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\boldsymbol{V}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{u}_i + \boldsymbol{G}_P\boldsymbol{r}_i(\boldsymbol{\theta}_0)) + o_p(1) \tag{43}$$

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

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}_{QLDP} - \boldsymbol{\beta}_0) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{A}_P^{-1}\boldsymbol{B}_P\boldsymbol{A}_P^{-1}) \tag{44}$$

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 5

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

Lemma 3. *Suppose Assumption 2 holds and* $Rk(E(\boldsymbol{\Gamma}_i)) = p_0$ *and let* $\widehat{\boldsymbol{\theta}}$ *be the GMM estimator based on* $E(vec(\boldsymbol{H}_0'\boldsymbol{X}_i)) = E(vec(\boldsymbol{H}_0'\boldsymbol{V}_i) = \boldsymbol{0}$ *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 2.*

□

Proof of Theorem 6

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

$$\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{Z}_i(\widehat{\boldsymbol{\theta}}) \xrightarrow{p} E(\boldsymbol{Z}_i(\boldsymbol{\theta}_0)) = \boldsymbol{0} \tag{45}$$

By Theorem 21.6 of Davidson (1994), the convergence result in equation (45) is implied by conditions:

$$\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0 \tag{46}$$

$$\sup_{\boldsymbol{\theta}\in B_0}\left\|\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{Z}_i(\boldsymbol{\theta}) - E(\boldsymbol{Z}_i(\boldsymbol{\theta}))\right\| = o_p(1) \text{ where } \boldsymbol{B}_0 \text{ is some open set about } \boldsymbol{\theta}_0. \tag{47}$$

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 that needs to be verified. I show uniform convergence via a traditional argument that 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. $\{\boldsymbol{X}_i'\boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{X}_i\}_{i\geq 1}$ is a sequence of positive definite random matrices for all possible

values of $\boldsymbol{\theta}$ by condition (i) 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 $\mathbf{0}$ 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 $h(x)\to 0$ as $x\to 0$, there exists $n\in\mathbb{N}$ such that

$$\frac{1}{N}\left\|(\boldsymbol{Z}_i(\boldsymbol{\theta})-E(\boldsymbol{Z}_i(\boldsymbol{\theta})))-(\boldsymbol{Z}_i(\dot{\boldsymbol{\theta}})-E(\boldsymbol{Z}_i(\dot{\boldsymbol{\theta}})))\right\| \leq B_{Ni}h(\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|) \tag{48}$$

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 (2012)). Thus $\left\|\boldsymbol{A}(\boldsymbol{\theta})^{-1}\right\|=a_i(\boldsymbol{\theta})>0$ and we have

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

Turning now to the other term from the triangle inequality, note that condition (i) of the theorem implies $\boldsymbol{A}(\boldsymbol{\theta})$

is nonsingular for any $\boldsymbol{\theta}$ in the parameter space. Then

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

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

Condition (iii) 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 (1994)'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(ii) and the bound on $\boldsymbol{a}_i(\boldsymbol{\theta})$ holds uniformly over $\mathcal{T}$ by Condition (ii) 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 to equation (38), we have

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

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 4.

Lemma 4. *Let* $\boldsymbol{\epsilon}_i = \boldsymbol{X}_i\boldsymbol{b}_i + \boldsymbol{F}_0\boldsymbol{\gamma}_i + \boldsymbol{u}_i$. *Then*

$$\nabla_{\boldsymbol{\theta}}(\boldsymbol{X}_i\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{\epsilon}_i = -\left(\boldsymbol{I}_K \otimes \boldsymbol{\epsilon}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{V}_i\right)\left((\boldsymbol{V}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{V}_i)^{-1} \otimes (\boldsymbol{V}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{V}_i)^{-1}\right)(\boldsymbol{I}_{K^2} + \boldsymbol{K}_K)(\boldsymbol{I}_K \otimes \boldsymbol{V}_i'\boldsymbol{H}_0) *$$
$$* \begin{pmatrix} \boldsymbol{x}_{i1}^{*\prime} \otimes \boldsymbol{I}_{T-p_0} \\ \vdots \\ \boldsymbol{x}_{iK}^{*\prime} \otimes \boldsymbol{I}_{T-p_0} \end{pmatrix} +$$
$$+ (\boldsymbol{V}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{V}_i)^{-1}\begin{pmatrix} (\boldsymbol{I}_K \otimes \boldsymbol{\epsilon}_i'\boldsymbol{H}_0)\begin{pmatrix} \boldsymbol{x}_{i1}^{*\prime} \otimes \boldsymbol{I}_{T-p_0} \\ \vdots \\ \boldsymbol{x}_{iK}^{*\prime} \otimes \boldsymbol{I}_{T-p_0} \end{pmatrix} + \boldsymbol{V}_i'\boldsymbol{H}_0\left(\boldsymbol{\epsilon}_i^{*\prime} \otimes \boldsymbol{I}_{T-p_0}\right) \end{pmatrix}$$

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

*Proof.*Like in Lemma 2, 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 (i) and (ii) 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

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

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 (2005):

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

The associated gradient was worked out in the proof of Theorem 5. Thus we have

$$\nabla_{\boldsymbol{\theta}}(\boldsymbol{X}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{X}_i)^{-1} = -\left((\boldsymbol{X}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{X}_i)^{-1} \otimes (\boldsymbol{X}_i'\boldsymbol{H}_0\boldsymbol{H}_0'\boldsymbol{X}_i)^{-1}\right)(\boldsymbol{I}_{K^2} + \boldsymbol{K}_K)(\boldsymbol{I}_K \otimes \boldsymbol{X}_i'\boldsymbol{H}_0)\begin{pmatrix} \boldsymbol{x}_{i1}^{*\prime} \otimes \boldsymbol{I}_{T-p_0} \\ \vdots \\ \boldsymbol{x}_{iK}^{*\prime} \otimes \boldsymbol{I}_{T-p_0} \end{pmatrix}$$

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 form Lemma 2 in the proof of Theorem 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 4. Thus

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}_{QLDMG} - \boldsymbol{\beta}_0) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left((\boldsymbol{V}_i' \boldsymbol{H}_0 \boldsymbol{H}_0' \boldsymbol{V}_i)^{-1} \boldsymbol{V}_i' \boldsymbol{H}_0 \boldsymbol{H}_0' \boldsymbol{\epsilon}_i + \boldsymbol{G}_{MG} \boldsymbol{r}_{x,i}(\boldsymbol{\theta}_0) \right) + o_p(1) \tag{49}$$

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 3. We then have

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}_{QLDMG} - \boldsymbol{\beta}_0) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{B}_{MG}) \tag{50}$$

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 B: Comparison to TWFE

Theorem 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 5.1 so that the QLDP results are identical to table 2.

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

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

Table 8: 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

Notes. This table presents a set of simulations with 1000 replications. Each table consists of a single data generating process where N and T vary. The two rows for a given pair of N and T are the values associated with estimators of each of the two coefficients. "SD" and "RMSE" are respectively the standard deviation and root mean squared error of the estimators over all replications for a given experiment.

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 6 compare TWFE to QLDP 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.

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 that is trying to eliminate more heterogeneity. However, it performs comparably well even though it removes more variation from the data than is needed.

Table 9: 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

Notes. This table presents a set of simulations with 1000 replications. Each table consists of a single data generating process where N and T vary. The two rows for a given pair of N and T are the values associated with estimators of each of the two coefficients. "SD" and "RMSE" are respectively the standard deviation and root mean squared error of the estimators over all replications for a given experiment.

References

Abadir, Karim M., and Jan R. Magnus. 2005. *Matrix Algebra*. Volume 1. Cambridge University Press.

Ahn, Seung C., Young H. Lee, and Peter Schmidt. 2013. "Panel data models with multiple time-varying individual effects." *Journal of Econometrics* 174 1–14. 10.1016/j.jeconom.2012.12.002.

Breitung, Jörg, and Philipp Hansen. 2021. "Alternative estimation approaches for the factor augmented panel data model with small T." *Empirical Economics* 60 327–351. 10.1007/s00181-020-01948-7.

Breitung, Jörg, and Nazarii Salish. 2021. "Estimation of heterogeneous panels with systematic slope variations." *Journal of Econometrics* 220 399–415. 10.1016/j.jeconom.2020.04.007.

Breusch, Trevor, Hailong Qian, Peter Schmidt, and Donald J Wyhowski. 1997. "REDUNDANCY OF MOMENT CONDITIONS." *Journal of Econometrics* 91.

Brown, Nicholas. 2022. "Information equivalence among transformations of semiparametric nonlinear panel data models *."Technical report, https://www.researchgate.net/publication/344047637_Information-equivalence_among_transformations_of_semiparametric_nonlinear_panel_data_models.

Brown, Nicholas, and Kyle Butts. 2022. "A Unified Framework for Dynamic Treatment Effect Estimation in Interactive Fixed Effect Models."Technical report.

Brown, Nicholas, Peter Schmidt, and Jeffrey M Wooldridge. 2021. "Simple Alternatives to the Common Correlated Effects Model."Technical report. 10.13140/RG.2.2.12655.76969/1.

Campello, Murillo, Antonio F. Galvao, and Ted Juhl. 2019. "Testing for Slope Heterogeneity Bias in Panel Data Models." *Journal of Business and Economic Statistics* 37 749–760. 10.1080/07350015.2017.1421545.

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

Chudik, Alexander, and M. Hashem Pesaran. 2015. "Common correlated effects estimation of heterogeneous dynamic panel data models with weakly exogenous regressors." *Journal of Econometrics* 188 393–420. 10.1016/j.jeconom.2015.03.007.

Davidson, James. 1994. *Stochastic Limit Theory: An Introduction for Econometricians*. Oxford University Press, . 10.1093/0198774036.001.0001.

Dhrymes, Phoebus J. 2013. *Mathematics for Econometrics*. Springer Science and Business Media.

Hansen, Lars Peter. 1982. "Large Sample Properties of Generalized Method of Moments Estimators." *Econometrica* 50 1029–1054. 10.2307/1912775.

Harding, M, C Lamarche, and C Muris. 2022. "Estimation of a Factor-Augmented Linear Model with Applications Using Student Achievement Data."Technical report, https://arxiv.org/abs/2203.03051.

Hayakawa, Kazuhiko. 2016. "Identification problem of GMM estimators for short panel data models with interactive fixed effects." *Economics Letters* 139 22–26. 10.1016/j.econlet.2015.12.012.

Horn, Roger A, and Charles R Johnson. 2012. *Matrix Analysis*. Cambridge University Press.

Juhl, Ted, and Oleksandr Lugovskyy. 2014. "A Test for Slope Heterogeneity in Fixed Effects Models." *Econometric Reviews* 33 906–935. 10.1080/07474938.2013.806708.

Juodis, Artūras, and Vasilis Sarafidis. 2018. "Fixed T dynamic panel data estimators with multifactor errors." *Econometric Reviews* 37 893–929. 10.1080/00927872.2016.1178875.

Karabiyik, Hande, Simon Reese, and Joakim Westerlund. 2017. "On the role of the rank condition in CCE estimation of factor-augmented panel regressions." *Journal of Econometrics* 197 (1): 60–64.

Moon, Hyungsik Roger, and Martin Weidner. 2015. "Linear Regression for Panel With Unknown Number of Factors as Interactive Fixed Effects." *Econometrica* 83 1543–1579. 10.3982/ecta9382.

Murtazashvili, Irina, and Jeffrey M. Wooldridge. 2008. "Fixed effects instrumental variables estimation in correlated random coefficient panel data models." *Journal of Econometrics* 142 539–552. 10.1016/j.jeconom.2007.09.001.

Neal, Timothy. 2015. "Estimating Heterogeneous Coefficients in Panel Data Models with Endogenous Regressors and Common Factors."Technical report.

Norkutė, Milda, Vasilis Sarafidis, Takashi Yamagata, and Guowei Cui. 2021. "Instrumental variable estimation of dynamic linear panel data models with defactored regressors and a multifactor error structure." *Journal of Econometrics* 220 416–446. 10.1016/j.jeconom.2020.04.008.

Papke, Leslie E. 2005. "The effects of spending on test pass rates: Evidence from Michigan." *Journal of Public Economics* 89 821–839. 10.1016/j.jpubeco.2004.05.008.

Papke, Leslie E., and Jeffrey M. Wooldridge. 2008. "Panel data methods for fractional response variables with an application to test pass rates." *Journal of Econometrics* 145 121–133. 10.1016/j.jeconom.2008.05.009.

Pesaran, M Hashem. 2006. "Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure." *Econometrica* 74 967–1012.

Prokhorov, Artem, and Peter Schmidt. 2009. "GMM redundancy results for general missing data problems." *Journal of Econometrics* 151 (1): 47–55.

Vos, Ignace De, and Gerdie Everaert. 2021. "Bias-Corrected Common Correlated Effects Pooled Estimation in Dynamic Panels." *Journal of Business and Economic Statistics* 39 294–306. 10.1080/07350015.2019.1654879.

Vos, Ignace De, and Joakim Westerlund. 2019. "On CCE estimation of factor-augmented models when regressors are not linear in the factors." *Economics Letters* 178 5–7. 10.1016/j.econlet.2019.02.001.

Westerlund, Joakim. 2019. "On Estimation and Inference in Heterogeneous Panel Regressions with Interactive Effects." *Journal of Time Series Analysis* 40 852–857. 10.1111/jtsa.12432.

Westerlund, Joakim, and Yousef Kaddoura. 2022. "CCE in heterogenous fixed-T panels." *The Econometrics Journal*.

Westerlund, Joakim, Yana Petrova, and Milda Norkute. 2019. "CCE in fixed-T panels." *Journal of Applied Econometrics* 34 746–761. 10.1002/jae.2707.

Wooldridge, Jeffrey 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.