Queen's Economics Department Working Paper No. 1494

Information Equivalence Among Transformations of Semiparametric Nonlinear Panel Data Models

Nicholas Brown
Queen's University

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

12-2022

Information Equivalence Among Transformations of Semiparametric Nonlinear Panel Data Models*

Nicholas Brown†
Department of Economics
Queen's University

Date of draft: December 1, 2022

Abstract

This paper considers transformations of nonlinear semiparametric mean functions that yield moment conditions for estimation. Such transformations are said to be information equivalent if they yield the same asymptotic efficiency bound. I derive a unified theory of algebraic equivalence for moment conditions created by a given linear transformation. The main equivalence result states that under standard regularity conditions, transformations that create conditional moment restrictions in a given empirical setting need only to have an equal rank to reach the same efficiency bound. Examples are included, where I compare feasible and infeasible transformations of both nonlinear models with multiplicative heterogeneity and linear models with arbitrary unobserved factor structures.

JEL Classification Codes: C14, C33, C36

Keywords: Semiparametric efficiency, nonlinear regression, generalized least squares, fixed-T

*I would like to thank Jeffrey Wooldridge and Peter Schmidt for their guidance and advice. I would also like to thank the participants of the AEASP seminar series, 2020 Red Cedar Conference, MSU Econometrics Reading Group, and 2021 MEA conference participants for their insightful questions and comments. All errors are my own.

†Correspondence to: Department of Economics, Queen's University, 94 University Ave, 209 Dunning Hall, Kingston, ON K7L 3N6

E-mail address: n.brown@queensu.ca

1 Introduction

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

This paper considers conditional mean models with unobserved heterogeneity. The general framework encompasses a large class of both linear and strictly nonlinear models, examples of which are given in Section 2.1. The models are referred to as "semiparametric" in the sense that nothing is assumed about the relationship between the heterogeneity and observables other than regularity conditions needed for asymptotic analysis. In place of assumptions on the conditional distribution of the heterogeneity, these models often require a transformation to eliminate or control for the unobservables.

I provide a unified framework for comparing such transformations in terms of the information they preserve. I show that transformations yielding conditional moment restrictions, given certain regularity assumptions, will provide the same $\sqrt{N}$-asymptotic efficiency bound if they have equal rank. This result is useful because once the researcher has multiple transformations that satisfy the conditions in my theory, they can choose one based solely on concerns of feasible inference, computational demand, and finite-sample bias reduction. I also demonstrate that for the examples in this paper, infeasible transformations that are functions of unobservables are often available to eliminate heterogeneity. An additional implication of my results is that once a researcher has a feasible transformation that is the same rank as the infeasible one, they do not need to search for additional transformations in order to reach the efficiency bound of the infeasible moment restrictions.

As mentioned above, the within and first-differencing transformations are the most common in the linear panel case for eliminating additive heterogeneity. When the covariates are strictly exogenous with respect to the idiosyncratic errors, these transformations provide conditional moment restrictions that can be exploited for estimation of the population parameters. For a given conditional variance matrix, Arellano and Bover (1995) suggest that Generalized Least Squares (GLS) on the demeaned equations is equivalent to the efficient 3SLS estimator. This claim was later proven in Im et al. (1999) along with a proof that the GLS estimators on the demeaned and first-differenced data are equivalent.

Their result shows that two commonly used methods of estimation preserve the same information in the linear case. However, they limit their investigation to a small number of estimators and only allow for a single time-invariant individual effect. My approach nests the results of Im et al. (1999), but also applies to more general interactive fixed effects models. Because some of the estimators for these models rely on nonlinear first-step estimation, it is beneficial to show that two transformations have the same information bound, so the empirical researcher can choose the one that is easier to compute and has better finite-sample properties.

One approach to estimation of nonlinear models with a multiplicative heterogeneity term is the fixed effects Poisson (FEP) estimator. Hausman et al. (1984) derive the FEP as the conditional maximum likelihood estimator of a multinomial distribution[1]. Wooldridge (1999) shows that the FEP is in fact consistent under a much weaker strict exogeneity assumption due the likelihood's implicit transformation of the data. Another approach is the generalized next-differencing transformation first studied by Chamberlain (1992) and Wooldridge (1997), which subtracts from a time period the next period outcome, weighted by the quotient of the mean functions. While generalized next-differencing was originally proposed for a sequential exogeneity setting, I study it here in the context of strict exogeneity. To the best of my knowledge, this paper is the first to show information equivalence between these two transformations.

In Section 2, I define information equivalence in a first order asymptotic sense. The efficiency bounds studied will apply to "small-T" settings where asymptotics are derived with T fixed as $N \to \infty$. I then derive sufficient conditions under which transformations of the data that yield moment restrictions for estimation preserve the same information. This result is general and can apply to a number of finite and asymptotic settings. In Section 3, I apply the main result from Section 2 to a nonlinear multiplicative model, a linear model with an unknown factor structure, and a linear random trend model. Section 4 discusses implementation of the efficiency bound associated with a given transformation. Section 5 provides concluding remarks along with potential directions for future research.

2 Information Equivalence

In what follows, $(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{c}_i)$ is assumed to be randomly sampled. The matrix $(\boldsymbol{y}_i, \boldsymbol{x}_i)$ is $T \times (1 + K)$ and observable whereas the random $p \times 1$ vector $\boldsymbol{c}_i$ is unobservable. All statements involving expressions of random variables hold with probability one. Finally, I assume regularity conditions suitable for asymptotic analysis such as bounds on the higher-order moments of the data.

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

2.1 Model

The following conditional mean assumption specifies the empirical setting:

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

$$E(y_{it}|\boldsymbol{x}_i, \boldsymbol{c}_i) = m_t(\boldsymbol{x}_{it}, \boldsymbol{\beta}_0, \boldsymbol{c}_i) \tag{1}$$

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

Equation (1) specifies a nonlinear semiparametric conditional mean function with strictly exogenous covariates where $\boldsymbol{\beta}_0$ is a $K \times 1$ vector of parameters[2]. The mean function itself is allowed to vary over time periods. Assuming the function is known up to its first and third arguments is equivalent to saying that it is correctly specified. That is, if $(\boldsymbol{x}_i, \boldsymbol{c}_i)$ were observed, estimation would be trivial. The heterogeneity is also allowed to enter the mean function in any arbitrary way. In the linear panel case, the simplest and most common specification is an individual-specific intercept. In nonlinear cases, the heterogeneity is often included as a multiplicative term.

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

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

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

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

This model is common among applied microeconometric researchers. Im et al. (1999) show that the 3SLS estimator of $\boldsymbol{\beta}_0$ using the differenced covariates as instruments is algebraically equivalent to GLS estimators based on both the within and differenced transformed residuals. This example is discussed in Section 3.3.

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

$$E(y_{it}|\boldsymbol{x}_i, \boldsymbol{c}_i) = \boldsymbol{c}_i'\boldsymbol{f}_t + \boldsymbol{x}_{it}\boldsymbol{\beta}_0$$

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

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

model.

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

The standard approach to estimation is to first-difference the outcomes to yield another linear model with only an additive individual effect. If strict exogeneity is assumed with respect to $\boldsymbol{x}_i$, we have the same empirical setting as above, and so the same analysis will apply. I discuss the general model in Section 3.2. ■

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

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

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

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

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

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

which I refer to as the generalized next-differencing transformation. Both of this transformations are studied in generality in Section 3.1.

In an analogy to the linear setting, I also discuss an exponential random trend model:

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

which can be motivated by the form $E(y_{it}|\boldsymbol{x}_i, c_i, a_i) = \exp(\gamma_i + \alpha_i t + \boldsymbol{x}_{it}\boldsymbol{\beta}_0)$. This model has received no attention in the econometric literature to the best of my knowledge. However, it may have practical applications to treatment effect analysis. Wooldridge (2022) considers estimation of treatment effect parameters in nonlinear mean models under generalized parallel trends assumption. Including a multiplicative random trend would weaken the parallel trends assumption needed for identification of average treatment effects. I discuss how the results of this paper could apply to such a model in Section 3.1. ■

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

production technology:

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

where (L, K) are labor and capital stock respectively. The heterogeneity can be written $\exp(-c_i)$. If $E(\epsilon_{it}|c_i, L_{it}, K_{it})$ is assumed constant[3], then the transformations studied in Section 3 can be used for estimation of the parameters and average partial effects under weak assumptions on the heterogeneity term. This example serves as an interesting bridge between the linear and nonlinear specifications as production theory can be stated in the above nonlinear fashion, but production function estimation is often carried out after log-linearization for which the results of Im et al. (1999) would apply. The specific form of the error is reminiscent of a stochastic frontier model with a time-invariant inefficiency term. See Section V of Amsler et al. (2009). ■

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

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

$$\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0) E(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{c}_i) = \boldsymbol{0} \tag{2}$$

and is differentiable in $\boldsymbol{\beta}$ over the interior of its parameter space for every $\boldsymbol{x} \in \mathcal{X}$. ■

$\boldsymbol{A}$ is a residual maker matrix that is zero at the true parameter value $\boldsymbol{\beta}_0$. I assume $L \leqslant T$, which corresponds to the examples studied in Section 3. While $L > T$ is theoretically possible and would rely on the same theory of g-inverses employed in this paper, I do not consider such a case. In fact, cases of the examples in Section 3 where $L > T$ often correspond to linearly dependent and hence redundant sets of moment conditions.

Under the previous assumptions,

$$E(\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)\boldsymbol{y}_i|\boldsymbol{x}_i) = E(\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0) E(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{c}_i)|\boldsymbol{x}_i) = \boldsymbol{0} \tag{3}$$

by iterated expectations. We can thus use equation (3) as the basis of a GMM estimator of $\boldsymbol{\beta}_0$, where any function of $\boldsymbol{x}_i$ can be used as an instruments for $\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)\boldsymbol{y}_i$ to improve efficiency. Note that $\boldsymbol{A}$ could contain external instrumental variables that do not appear in the mean function. This more general case is considered in Section 2.2.

I note that whenever the heterogeneity is additively or multiplicatively separable, there always exists an infeasible transformation that satisfies Assumption MAT. For example, suppose $E(y_{it}|\boldsymbol{x}_i, c_i) = c_i f_t(\boldsymbol{x}_i, \boldsymbol{\beta}_0)$ for some set of nonlinear functions $\{f_t\}_{t=1}^T$. Then the residual-maker matrix from regressing on the stacked vector

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

$c_i \boldsymbol{f}(\boldsymbol{x}_i, \boldsymbol{\beta}_0) = (c_i f_1(\boldsymbol{x}_i, \boldsymbol{\beta}_0), ..., c_i f_T(\boldsymbol{x}_i, \boldsymbol{\beta}_0))'$ produces the relevant conditional moment restrictions[4]. However, this estimator is generally difficult to work with because it is a function of unobservables and highly nonlinear in the parameters of interest. I prove in a later section that this matrix provides the same asymptotic information bound as the FEP estimator, where robust inference is possible through a variety of canned statistical packages. This fact demonstrates the importance of a general theory: once we find a feasible transformation with the same rank as the infeasible one, we can implement an efficient estimator that is at least asymptotically equivalent to the infeasible information bound.

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

Lemma 1. *Suppose $\boldsymbol{A}(\boldsymbol{x}, \boldsymbol{\beta})$ is an $L \times T$ matrix satisfying Assumption MAT. Then for any $(\boldsymbol{x}^0, \boldsymbol{c}^0) \in \mathcal{X} \times \mathcal{C}$ such that $|m_t(\boldsymbol{x}_t^0, \boldsymbol{\beta}_0, \boldsymbol{c}^0)| > 0$ for some t, Rank$(\boldsymbol{A}(\boldsymbol{x}^0, \boldsymbol{\beta}_0)) < T$.*

Proof. See Appendix for proof. □

The theory for choosing optimal instruments is well-known: when the conditional variance is nonsingular, the optimal GMM estimator uses instruments $(Var(\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)\boldsymbol{y}_i|\boldsymbol{x}_i)^{-1} E(\nabla_{\boldsymbol{\beta}} \boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)\boldsymbol{y}_i|\boldsymbol{x}_i))'$. However, in most nontrivial cases when $\boldsymbol{A}$ is $T \times T$, the conditional variance matrix of $\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)\boldsymbol{y}_i$ is singular even when $Var(\boldsymbol{y}_i|\boldsymbol{x}_i)$ is nonsingular; I consider such examples in Section 3. I make one additional assumption on the transformation studied that allows for such a generality. Assumption SYS specifies consistency of a particular linear system that is necessary for the definition of the asymptotic efficiency bound. It will allow us to use a certain class of generalized inverses when the conditional variance is singular.

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

Consistency of a linear system only requires the existence of a solution and not necessarily uniqueness. In fact, Section 3 considers relevant cases for which uniqueness does not hold. Assumption SYS is posed in Newey (2001) for studying censored and truncated regression. It holds trivially when the conditional variance is nonsingular, in which case case the unique solution is $Var(\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)\boldsymbol{y}_i|\boldsymbol{x}_i)^{-1} E(\nabla_{\boldsymbol{\beta}} \boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)\boldsymbol{y}_i|\boldsymbol{x}_i)$. The results in Chamberlain (1987) and Newey (2001) show that the semiparametric efficiency bound for estimating $\boldsymbol{\beta}_0$ using equation (3) and Assumptions CM, MAT, and SYS is

$$E\left(E(\nabla_{\boldsymbol{\beta}} \boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)\boldsymbol{y}_i|\boldsymbol{x}_i)' Var(\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)\boldsymbol{y}_i|\boldsymbol{x}_i)^{-} E(\nabla_{\boldsymbol{\beta}} \boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)\boldsymbol{y}_i|\boldsymbol{x}_i)\right)^{-1} \tag{4}$$

where "$-$" denotes a symmetric g-inverse[5]. That is, no $\sqrt{N}$-consistent estimator of $\boldsymbol{\beta}_0$ based on equation (3)

[4]The case of interactive fixed effects is similar and handled in Section 3.2.

[5]A g-inverse for matrix $\boldsymbol{\Omega}$ is a matrix $\boldsymbol{\Omega}^{-}$ such that $\boldsymbol{\Omega}\boldsymbol{\Omega}^{-}\boldsymbol{\Omega} = \boldsymbol{\Omega}$. This condition is weaker than the Moore-Penrose inverse

has a smaller asymptotic variance than (4).

Theorem 5.2 in Newey (2001) shows that the efficiency bound in (4) is invariant to the choice of symmetric g-inverse under Assumption SYS. If the conditional variance is nonsingular, then the g-inverse can be replaced by a proper inverse as in Chamberlain (1987). Otherwise, any g-inverse will work as long as the consistency assumption holds. The matrix in (4) is also equivalent to the asymptotic variance of the GMM estimator based on the moment conditions in (3), using the optimal instruments $(Var(\boldsymbol{A}(\boldsymbol{x}_i,\boldsymbol{\beta}_0)\boldsymbol{y}_i|\boldsymbol{x}_i)^{-}E(\nabla_{\boldsymbol{\beta}}\boldsymbol{A}(\boldsymbol{x}_i,\boldsymbol{\beta}_0)\boldsymbol{y}_i|\boldsymbol{x}_i))'$. The system is just identified and so no weight matrix is required for the asymptotic bound. Realizing this efficiency bound is the subject of Section 4.

The rest of the paper is concerned with studying transformations of the observed data that provide the same semiparametric efficiency bound as defined in (4). The following definition characterizes the types of transformations I consider:

Definition: Let Assumption CM hold, and let $\boldsymbol{A}(\boldsymbol{x}_i,\boldsymbol{\beta})$ and $\boldsymbol{B}(\boldsymbol{x}_i,\boldsymbol{\beta})$ be $L\times T$ and $M\times T$, respectively. Given $\boldsymbol{A}$ and $\boldsymbol{B}$ satisfy Assumptions MAT and SYS, the matrices are **information equivalent transformations** if their semiparametric efficiency bounds given by (4) are equal. ■

Information equivalence defined above is an equivalence relation on the set of $K\times K$ real-valued matrices since it is defined via matrix equivalences. This fact will be used in Section 3 to show information equivalence between general forms of applied transformations. Information equivalence is similar to the definition of redundancy of moment conditions as given by Breusch et al. (1999). However, the results in this paper are not direct consequences of their redundancy results, as I allow the moment conditions to have singular covariance matrices.

2.2 General Equivalence Result

I now prove a unifying theory of information equivalence. Consider the empirical setting proposed in Section 2.1 where Assumption CM holds. I suppose there is a $T\times T$ matrix $\boldsymbol{M}(\boldsymbol{z}_i,\boldsymbol{\beta})$ satisfying Assumptions MAT and SYS, where $\boldsymbol{z}_i$ is allowed to include any element of $\boldsymbol{x}_i$ and outside instruments. Dropping the arguments and writing $\boldsymbol{M}_i=\boldsymbol{M}(\boldsymbol{z}_i,\boldsymbol{\beta}_0)$ for simplicity, we have the following moment conditions:

$$E(\boldsymbol{M}_i\boldsymbol{y}_i|\boldsymbol{z}_i)=\boldsymbol{0} \tag{5}$$

Equation (5) includes the case of unconditional moment restrictions.

I denote $\boldsymbol{V}_i=E(\boldsymbol{y}_i\boldsymbol{y}_i'|\boldsymbol{z}_i)$ and let $\boldsymbol{B}_i=\boldsymbol{B}(\boldsymbol{z}_i,\boldsymbol{\beta}_0)$ be a $J\times T$ matrix such that $E(\boldsymbol{B}_i\boldsymbol{y}_i|\boldsymbol{z}_i)=\boldsymbol{0}$. The following assumptions are pivotal for the general result of this section, so I refer to them as Assumptions GR.1 and GR.2.

which requires three other non-redundant properties. It is worth noting that the Moore-Penrose inverse is unique, but a g-inverse is not necessarily; this fact will be used to prove the main results in Section 3. For a general treatment of g-inverses, see Rao and Mitra (1978).

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

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

The notation for $\boldsymbol{M}_i$ in Assumption GR.1 is motivated by the standard notation for a residual maker matrix. In fact, one possible sufficient condition for Assumption GR.1 is that $Rank(\boldsymbol{V}_i) = J$ and that $\boldsymbol{V}_i$ shares a null space with $\boldsymbol{B}_i$. This assumption would also suffice for Assumption GR.2 since $\boldsymbol{B}_i'$ spans the column space of $\boldsymbol{V}_i$, and is relevant in linear panel models with additive heterogeneity. We can then let $\boldsymbol{M}_i$ be a residual maker matrix from regressing on a basis vector for the null space of $\boldsymbol{B}_i$. Another relevant setting to this paper is when $\boldsymbol{M}_i = \boldsymbol{I}_T - \boldsymbol{P}_i$ where $\boldsymbol{P}_i$ has rank $T - J$ and $\boldsymbol{B}_i\boldsymbol{P}_i = \boldsymbol{0}$. This setting characterizes the nonlinear models studied in Section 3 and is also sufficient for Assumptions GR.1 and GR.2.

I now provide a lemma that is essential to the proof of the general equivalence result.

Lemma 2. *$\boldsymbol{B}_i'(\boldsymbol{B}_i\boldsymbol{V}_i\boldsymbol{B}_i')^{-1}\boldsymbol{B}_i$ is a g-inverse of $\boldsymbol{M}_i\boldsymbol{V}_i\boldsymbol{M}_i'$.*

Proof. See Appendix for proof. □

Theorem 1. *The equality*

$$\boldsymbol{B}_i'(\boldsymbol{B}_i\boldsymbol{V}_i\boldsymbol{B}_i')^{-1}\boldsymbol{B}_i = \boldsymbol{M}_i'(\boldsymbol{M}_i\boldsymbol{V}_i\boldsymbol{M}_i')^{-}\boldsymbol{M}_i \tag{6}$$

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

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

$$\boldsymbol{M}_i'(\boldsymbol{M}_i\boldsymbol{V}_i\boldsymbol{M}_i')^{-}\boldsymbol{M}_i \tag{7}$$

is invariant to the choice of g-inverse as $Rank(\boldsymbol{M}_i\boldsymbol{V}_i\boldsymbol{M}_i') = Rank(\boldsymbol{M}_i)$ by Assumption 4. Since $\boldsymbol{B}_i'(\boldsymbol{B}_i\boldsymbol{V}_i\boldsymbol{B}_i')^{-1}\boldsymbol{B}_i$ is such a g-inverse by Lemma 2 and $\boldsymbol{B}_i\boldsymbol{M}_i = \boldsymbol{B}_i$ we have

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

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

The proof of Theorem 1 is included in the text because equation (7) provides the framework for evaluating information equivalence. To see how, I include an additional orthogonality assumption that simplifies the efficiency bound in (4).

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

Assumption ORTH is clearly sufficient for Assumption MAT. The transformations studied in the next section satisfy Assumption ORTH for all values of $\boldsymbol{\beta} \in \mathbb{R}^K$ for which the mean function is well-defined. However it only needs to be defined on a convex open set so that it applies with respect to differentiation. Note that ORTH does not say anything about point identification of $\boldsymbol{\beta}_0$. Assumption CM guarantees $E(\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)\boldsymbol{y}_i|\boldsymbol{x}_i) = \mathbf{0}$ only at $\boldsymbol{\beta}_0$ because $E(\boldsymbol{y}_{it}|\boldsymbol{x}_i, \boldsymbol{c}_i) = m_t(\boldsymbol{x}_{it}, \boldsymbol{\beta}_0, \boldsymbol{c}_i)$. I also note that every transformation considered in Section 3 satisfies Assumption ORTH.

The following lemma is a consequence of Assumption ORTH, and greatly simplifies the bound in (4).

Lemma 3. *Let $\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta})$ satisfy Assumption ORTH. Then under regularity conditions which allow us to pass the gradient operator through the conditional expectation,*

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

Proof. See Appendix for proof. □

Note that the right-hand side of Lemma 3 is allowed to depend on unobserved heterogeneity. This fact will be demonstrated in Section 3.1. It also allows us to say something about finite sample equivalence among certain types of transformations. I summarize these results here:

Corollary 1. *Let $\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta})$ be a $L \times T$ matrix satisfying Assumptions MAT, SYS, and ORTH. Then $\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)$ has the following efficiency bound:*

$$E\left(\nabla_{\boldsymbol{\beta}}\boldsymbol{m}_i(\boldsymbol{\beta}_0)'\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)'(\,\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)E(\boldsymbol{y}_i\boldsymbol{y}_i'|\boldsymbol{x}_i)\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)'\,)^{-}\,\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)\nabla_{\boldsymbol{\beta}}\boldsymbol{m}_i(\boldsymbol{\beta}_0)\right)^{-1} \tag{8}$$

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

$$\nabla_{\boldsymbol{\beta}}\boldsymbol{m}_i'\boldsymbol{A}_i'(\boldsymbol{A}_i\boldsymbol{V}_i\boldsymbol{A}_i')^{-1}\boldsymbol{A}_i\boldsymbol{m}_i(\boldsymbol{\beta}) = \nabla_{\boldsymbol{\beta}}\boldsymbol{m}_i'\boldsymbol{B}_i'(\boldsymbol{B}_i\boldsymbol{V}_i\boldsymbol{B}_i')^{-1}\boldsymbol{B}_i\boldsymbol{m}_i(\boldsymbol{\beta}) \tag{9}$$

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

Corollary 1 allows us to directly apply the result from Theorem 1 to the relevant cases in Section 3. For information equivalence, it will suffice to show that the relevant transformations satisfying Assumptions MAT, SYS, and ORTH only need to satisfy a rank assumption to be information equivalent. The choice of $\boldsymbol{M}$ will become apparent based on the empirical setting.

Corollary 2 gives an even more powerful result than equivalence of efficiency bounds. For example, if the moment conditions in (5) are conditional on $\boldsymbol{x}_i$, the efficient GMM estimator of $\boldsymbol{\beta}_0$, say $\widehat{\boldsymbol{\beta}}$, solves

$$\sum_{i=1}^{N} \nabla_{\boldsymbol{\beta}} \boldsymbol{m}_i' \boldsymbol{M}_i' (\boldsymbol{M}_i \boldsymbol{V}_i \boldsymbol{M}_i')^{-} \boldsymbol{M}_i \boldsymbol{m}_i(\widehat{\boldsymbol{\beta}}) = \boldsymbol{0} \tag{10}$$

Corollary 2 tells us that the efficient estimator based on $E(\boldsymbol{A}_i \boldsymbol{y}_i | \boldsymbol{x}_i)$ and $E(\boldsymbol{B}_i \boldsymbol{y}_i | \boldsymbol{x}_i)$ are algebraically equivalent. When the transformations are themselves functions of the parameters, or the mean function is nonlinear in the parameters, implementation of the efficient instruments generally depends on first-stage estimators and this finite-sample equivalence result breaks down. The proof of Theorem 4.2 in Im et al. (1999) uses a specific form of the argument in the proof above.

This result does not generally apply to the nonlinear estimation problems considered in this paper because it will require the score of the moment functions to not be functions of the underlying parameters. However, it will demonstrate why the results on linear random trend estimators hold in finite samples. It will generally apply to linear estimators where heterogeneity loads onto known common macro variables. This fact suggests further applications to panel data transformations with strictly exogenous covariates, which I explore in the next section.

3 Examples of Information Equivalence

This section considers the application of Theorem 1 to a variety of interesting empirical settings.

3.1 Multiplicative Heterogeneity

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

$$E(y_{it} | \boldsymbol{x}_i, c_i) = c_i m_t(\boldsymbol{x}_{it}, \boldsymbol{\beta}_0) \tag{11}$$

This specification has grown in popularity in recent years. For example, see McCabe and Snyder (2014, 2015), Schlenker and Walker (2016), Krapf et al. (2017), Fischer et al. (2018), Castillo et al. (2020), and Williams et al. (2020). The most common specification of equation (11) is the exponential mean function, as demonstrated in Example 2 of Section 2.1. Often, the data generating process is a count variable with a mass point at zero, but the model can apply to any nonnegative outcome. This assumption typically means $m_t(\boldsymbol{x}, \boldsymbol{\beta}_0) > 0$ for all $\boldsymbol{x} \in \mathcal{X}$, which the rank assumptions made in this section will also imply.

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

$$u_{it}(\boldsymbol{\beta}) = y_{it} - \left(\sum_{s=1}^{T} y_{is}\right) p_{it}(\boldsymbol{\beta}) \tag{12}$$

$$r_{i,t,s}(\boldsymbol{\beta}) = y_{it} - y_{is}\frac{m_t(\boldsymbol{x}_{it},\boldsymbol{\beta})}{m_s(\boldsymbol{x}_{is},\boldsymbol{\beta})} \tag{13}$$

where $p_{it}(\boldsymbol{\beta}) = m_t(\boldsymbol{x}_{it},\boldsymbol{\beta})\left(\sum_{s=1}^{T} m_s(\boldsymbol{x}_{is},\boldsymbol{\beta})\right)^{-1}$. Equation (12) is reminiscent of the linear within transformation. However, the transformation in the linear case demeans using the time averages, whereas the generalized within transformation weights by the pseudo-probability $p_{it}(\boldsymbol{\beta})$. The generalized differencing residual in equation (13) allows a large number of differencing procedures, including next- and first-differencing, as well as fixing t and allowing s to vary. Any generalized differencing procedure is allowed so long as it produces a full rank transformation.

In contrast to the linear model with an additive effect, the transformations in equations (12) and (13) will not eliminate the heterogeneity, but still create valid moment conditions. For example, taking the mean of equation (13) conditional on $(\boldsymbol{x}_i, c_i)$ gives

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

which still yields conditional moment restrictions by iterated expectations.

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

$$\boldsymbol{u}_i(\boldsymbol{\beta}) = (\boldsymbol{I}_T - \boldsymbol{p}_i(\boldsymbol{\beta})\mathbf{1}')\boldsymbol{y}_i \tag{14}$$

$$\boldsymbol{r}_i(\boldsymbol{\beta}) = \boldsymbol{D}_i(\boldsymbol{\beta})\boldsymbol{y}_i \tag{15}$$

where $\mathbf{1}$ is a $T \times 1$ vector of ones and $\boldsymbol{D}_i(\boldsymbol{\beta})$ is the $T-1 \times T$ weighted generalized differencing matrix that yields the desired residuals as in (13). I refer to transformations in equations (14) and (15) as the **generalized within** and **generalized differencing** transformations respectively. Then an iterated expectations argument shows $E(\boldsymbol{u}_i(\boldsymbol{\beta}_0)|\boldsymbol{x}_i) = \mathbf{0}$ and $E(\boldsymbol{r}_i(\boldsymbol{\beta}_0)|\boldsymbol{x}_i) = \mathbf{0}$. Thus equations (14) and (15) satisfy Assumption MAT and suggest moment conditions for efficient GMM estimation that could reach their respective efficiency bounds in (4).

As discussed in the Introduction, equation (14) is the foundation of the FEP estimator. The FEP is defined in Hausman et al. (1984) as the MLE of a conditional Multinomial distribution with probability and count

parameters $\boldsymbol{p}_i(\boldsymbol{\beta}_0) = (p_{i1}(\boldsymbol{\beta}_0), ..., p_{iT}(\boldsymbol{\beta}_0))'$ and n_i. Wooldridge (1999) shows that the FEP is consistent under Assumption CM using the fact that equation (14) has a zero conditional mean at $\boldsymbol{\beta}_0$ regardless of the true distribution of $\boldsymbol{y}_i|(\boldsymbol{x}_i, c_i)$. This robustness result helped lead to its proliferation in empirical research. As for efficiency, Hahn (1997) shows that the FEP is asymptotically efficient under the full set of Multinomial distributional assumptions. Verdier (2018) strengthens this result substantially by showing efficiency under just zero conditional correlation and conditional mean-variance equality. Brown and Wooldridge (2022) extend this result to allow arbitrary constant conditional mean-variance dispersion.

Equation (15) was first studied by Chamberlain (1992) and Wooldridge (1997) in the context of next-differencing for nonlinear models. It can also allow for estimation of $\boldsymbol{\beta}_0$ under weaker forms of exogeneity, like sequential exogeneity in the next-differencing case of $s = t + 1$, rather than the strict exogeneity implied by Assumption CM. However, remarkably less is known about efficient estimation based on equation (15) when compared to equation (14) in the context of strict exogeneity as studied here.

The transformations defined in (14) and (15) are clearly not the only transformations that satisfy Assumption MAT. Consider the residual maker matrix from regressing on the mean function defined by equation (11): $(\boldsymbol{I}_T - \boldsymbol{m}_i(\boldsymbol{\beta})(\boldsymbol{m}_i(\boldsymbol{\beta})'\boldsymbol{m}_i(\boldsymbol{\beta}))^{-1}\boldsymbol{m}_i(\boldsymbol{\beta})')$. This matrix satisfies Assumption ORTH and thus Assumption MAT since it is algebraically orthogonal to the mean function by construction. It is also well-known that the matrix is symmetric, idempotent, and has rank $T - 1$. I will refer to this matrix as the **residual maker** transformation. It is also important to note that this transformation is equal to the the infeasible residual maker matrix from regressing on $c_i\boldsymbol{m}_i(\boldsymbol{\beta})$, thus making it equivalent to an infeasible transformation.

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

Lemma 4 will show that the efficiency bounds of the within and residual maker transformations are well-defined. First I will assume that $E(\boldsymbol{y}_i\boldsymbol{y}_i'|\boldsymbol{x}_i)$ is strictly positive definite, a weaker assumption than the conditional variance of $\boldsymbol{y}_i$ itself being positive definite. Under this assumption, the conditional variance of the generalized differencing transformation is nonsingular. Before I can verify Assumption SYS, I will need an additional rank assumption for each respective transformation.

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

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

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

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

$$\boldsymbol{V}_i^- \left[Var((\boldsymbol{I}_T - \boldsymbol{p}_i(\boldsymbol{\beta}_0)\boldsymbol{1}')\boldsymbol{y}_i|\boldsymbol{x}_i)\right] \boldsymbol{V}_i^- = \boldsymbol{V}_i^- \tag{16}$$

as shown in Lemma 2 of Verdier (2018) so that it is a reflexive inverse and is also clearly symmetric. Assumption RK.2 suffices for the bound in (4) existing, as I show in the next lemma that $\boldsymbol{V}_i^-\nabla_{\boldsymbol{\beta}}\boldsymbol{m}_i(\boldsymbol{\beta}_0)$ is a solution to the system in Assumption SYS. This fact, along with the fact that $\boldsymbol{V}_i^-\boldsymbol{m}_i(\boldsymbol{\beta}_0) = \boldsymbol{0}$ and Lemma 3, gives the bound in (4) as the expectation above. The following lemma shows that all transformations studied satisfy Assumption SYS and so any symmetric g-inverse will suffice.

Lemma 4. *Suppose Assumptions CM, RK.1, and RK.2 hold and that $E(\boldsymbol{y}_i\boldsymbol{y}_i'|\boldsymbol{x}_i)$ is positive definite. Then the generalized differencing, generalized within, and residual maker transformations satisfy Assumption SYS. Further, either of the $T \times T$ transformations with any arbitrary row deleted also satisfy Assumption SYS.*

Proof. See Appendix for proof. □

The main consequence of Lemma 4 is that the asymptotic efficiency bound is well-defined and invariant to symmetric g-inverse for all of the transformations studied in this section. Now I can formally state the application of the main equivalence theorem to the transformations studied in this section. First note that Assumptions CM, RK.1, RK.2, and the positive definiteness of $E(\boldsymbol{y}_i\boldsymbol{y}_i'|\boldsymbol{x}_i)$ are sufficient for each of the transformations studied to satisfy Assumptions SYS and ORTH (and thus MAT) so that their asymptotic efficiency bounds are well-defined and given by (8). Also, the equivalence among estimators based on these moment conditions is only guaranteed to hold asymptotically because $\boldsymbol{\beta}_0$ enters the transformations in a highly nonlinear way.

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

Proof. See Appendix for proof. □

The proof of Theorem 2 is independent of which row is deleted in choosing $\boldsymbol{Q}$ and the type of differencing chosen in $\boldsymbol{D}$ satisfying Assumption RK.1, reinforcing the importance of the rank assumptions. As in Theorem 1, transformations with rank $L < T$ can be shown to be information equivalent via a similar argument, but this fact is not directly relevant to the current results. It's also important to note that the list of information equivalent transformations is not necessarily exhaustive, as any $T \times T$ or $(T-1) \times T$ matrix with rank $T-1$ and respective orthogonality condition will be information equivalent to the transformations in Theorem 2 by Theorem 1. Finally, I point out that the generalized within and differencing transformations are information equivalent to the infeasible residual-maker matrix, thus demonstrating that these feasible transformations give the same information bound as a transformation based off of unobserved heterogeneity.

Similar to the discussion after Theorem 1, the results in Theorem 2 could also apply to mean functions that have already been transformed. Consider the multiplicative random trend from Example 2, $y_{it} = c_i a_i^t m_t(\boldsymbol{x}_{it}, \boldsymbol{\beta}_0) u_{it}$, where u_{it} is an idiosyncratic error. If we assume the outcomes are bounded away from zero, we could first divide each outcome by the previous period. We now have the multiplicative model $y_{it}^* = a_i \frac{m_t(\boldsymbol{x}_{it}, \boldsymbol{\beta}_0)}{m_{t-1}(\boldsymbol{x}_{i,t-1}, \boldsymbol{\beta}_0)} \frac{u_{it}}{u_{i,t-1}}$. If $\frac{u_{it}}{u_{i,t-1}}$ is independent of $\boldsymbol{x}_i$ and a_i with mean 1, we have the model from equation (11). Then all of the transformations studied here are information equivalent on the $(T-1) \times 1$ vector of transformed outcomes $\boldsymbol{y}_i^*$.

As mentioned earlier, a multiplicative random trend model would weaken the generalized parallel trends assumption of Wooldridge (2022) and allow for more robust estimation of treatment effect parameters. Wooldridge assumes that pre-treatment outcome paths are parallel after applying a monotonic transformation to the mean function. Under an exponential mean assumption, we have

$$\begin{aligned} \log\left(E(y_{it}|c_i, a_i, \boldsymbol{x}_i)\right) &= \log(\exp(c_i + a_i t + \boldsymbol{x}_{it}\boldsymbol{\beta}_0)) \\ &= c_i + a_i t + \boldsymbol{x}_{it}\boldsymbol{\beta}_0 \end{aligned}$$

Then the generalized parallel trends assumption of Wooldridge (2022) can hold after controlling for a unit-specific linear trend via dividing contemporary outcomes by prior outcomes. Because we have the information equivalence result in Theorem 2, it will be easier to derive the semiparametric efficiency bound for the nonlinear imputation estimators in Wooldridge (2022) by comparing them to known transformations.

3.2 Linear Factor Model

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

$$E(y_{it}|\boldsymbol{x}_i, \boldsymbol{\gamma}_i) = \boldsymbol{x}_{it}\boldsymbol{\beta}_0 + \boldsymbol{f}_t'\boldsymbol{\gamma}_i \tag{17}$$

where $\boldsymbol{f}_t$ is a $p \times 1$ vector of common factors and $\boldsymbol{\gamma}_i$ is a $p \times 1$ vector of heterogeneous factor loadings. I follow Ahn et al. (2013) in assuming the factors are deterministic. I stack the factors into the $T \times p$ matrix $\boldsymbol{F} = (\boldsymbol{f}_1, ..., \boldsymbol{f}_T)'$. If $\boldsymbol{f}_t$ was known, efficient estimation of $\boldsymbol{\beta}_0$ is possible by considering a GLS estimator based on the residuals $\boldsymbol{M}_{\boldsymbol{F}}\boldsymbol{y}_i = (\boldsymbol{I}_T - \boldsymbol{F}(\boldsymbol{F}'\boldsymbol{F})^{-1}\boldsymbol{F}')\boldsymbol{y}_i$. However, the challenge for practitioners comes from the fact that $\boldsymbol{F}$ is unobserved and must be estimated[6].

Pesaran (2006) adds the additional reduced form equation

$$\boldsymbol{x}_i = \boldsymbol{F}\boldsymbol{\Gamma}_i + \boldsymbol{v}_i \tag{18}$$

where $\boldsymbol{\Gamma}_i$ is a $p \times K$ matrix of factor loadings and $\boldsymbol{v}_i$ is a $T \times K$ matrix of mean zero idiosyncratic errors. I write $\boldsymbol{z}_i = (\boldsymbol{y}_i, \boldsymbol{x}_i)$; under the assumptions in Pesaran (2006), equations (17) and (18) imply

$$E(\boldsymbol{z}_i) = \boldsymbol{F}\boldsymbol{C}\boldsymbol{Q} \tag{19}$$

where $\boldsymbol{C}$ is the $p \times K+1$ mean matrix of factor loadings and $\boldsymbol{Q}$ is a a full rank $K+1 \times K+1$ matrix[7]. Assuming $p \leqslant K+1$, $\boldsymbol{C}\boldsymbol{Q}$ is full rank, which suggests that $E(\boldsymbol{z}_i)$ can control for the space spanned by $\boldsymbol{F}$. The pooled common correlated effects estimator (CCEP) is defined as

$$\widehat{\boldsymbol{\beta}}_{CCEP} = \left(\sum_{i=1}^{N} \boldsymbol{x}_i' \boldsymbol{M}_{\widehat{\boldsymbol{F}}} \boldsymbol{x}_i \right)^{-1} \sum_{i=1}^{N} \boldsymbol{x}_i' \boldsymbol{M}_{\widehat{\boldsymbol{F}}} \boldsymbol{y}_i \tag{20}$$

where $\widehat{\boldsymbol{F}} = \overline{\boldsymbol{Z}} = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{y}_i, \boldsymbol{x}_i)$.

Westerlund et al. (2019) shows that when T is fixed and $N \to \infty$, $\boldsymbol{M}_{\widehat{\boldsymbol{F}}} \xrightarrow{p} \boldsymbol{M}_{\boldsymbol{F}} - \boldsymbol{P}_{-p}$ where $\boldsymbol{P}_{-p}$ is a nonlinear function of the model's errors. When $p = K+1$ and the number of cross-sectional averages equals the number of factors, $\boldsymbol{P}_{-p} = \boldsymbol{0}$, and so the CCE transformation is asymptotically equivalent to the infeasible factor residual-maker matrix $\boldsymbol{M}_{\boldsymbol{F}}$. I first start with the case $p = K+1$ so that the rank of the cross-sectional averages equal the rank of their limit, and standard asymptotic theory will apply. I discuss the $p < K+1$ case after the main theorem of this section.

Another fixed-T approach comes from Ahn et al. (2013). They do not make the reduced form assumption in equation (18). Instead, they introduce new parameters that allow them to eliminate $\boldsymbol{F}$. They impose the following p^2 normalizations on the factor matrix:

$$\boldsymbol{F} = (\boldsymbol{\Theta}', -\boldsymbol{I}_p)' \tag{21}$$

[6]Estimation of $\boldsymbol{F}$ is generally impossible because both $\boldsymbol{F}$ and $\boldsymbol{\gamma}_i$ are unobserved. However, it is often possible to estimate $\boldsymbol{F}$ up to an unobserved rotation that allows one to remove the factor structure completely.

[7]I assume $\boldsymbol{C}$ has full row rank for the purposes of this paper. I also assume $\overline{\boldsymbol{C}} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{C}_i$ is full rank for all N with probability one. For a discussion about this important CCE rank condition, see Westerlund and Urbain (2013)

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

$$\boldsymbol{H}(\boldsymbol{\theta})=\begin{pmatrix}\boldsymbol{I}_{T-p}\\ \boldsymbol{\Theta}'\end{pmatrix} \tag{22}$$

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

The Ahn et al. (2013) technique involves jointly estimating $(\boldsymbol{\beta}_0',\boldsymbol{\theta}')'$ with the use of many instruments. However, they do not estimate the optimal instruments in their paper. Rather than considering joint efficient estimation, I focus on the QLD transformation outside of their estimation problem and consider its asymptotic efficiency bound. Brown (2022) shows that the QLD parameters are identified under the CCE model and proposes a linear pooled QLD estimator to compare to the pooled CCE estimator. It is therefore fair to compare the QLD and CCE transformation because they can both be identified using the same set of moments.

Suppose $\boldsymbol{\Omega}_i=E(\boldsymbol{u}_i\boldsymbol{u}_i'|\boldsymbol{x}_i)$ is known and has full rank. Define the CCE GLS and QLD GLS estimators as

$$\widehat{\boldsymbol{\beta}}_{CCEGLS}=\left(\sum_{i=1}^N\boldsymbol{x}_i'\boldsymbol{M}_{\boldsymbol{F}}(\boldsymbol{M}_{\boldsymbol{F}}\boldsymbol{\Omega}_i\boldsymbol{M}_{\boldsymbol{F}})^-\boldsymbol{M}_{\boldsymbol{F}}\boldsymbol{x}_i\right)^{-1}\sum_{i=1}^N\boldsymbol{x}_i'\boldsymbol{M}_{\boldsymbol{F}}(\boldsymbol{M}_{\boldsymbol{F}}\boldsymbol{\Omega}_i\boldsymbol{M}_{\boldsymbol{F}})^-\boldsymbol{M}_{\boldsymbol{F}}\boldsymbol{y}_i \tag{23}$$

$$\widehat{\boldsymbol{\beta}}_{QLDGLS}=\left(\sum_{i=1}^N\boldsymbol{x}_i'\boldsymbol{H}(\boldsymbol{\theta})(\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{\Omega}_i\boldsymbol{H}(\boldsymbol{\theta}))^{-1}\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{x}_i\right)^{-1}\sum_{i=1}^N\boldsymbol{x}_i'\boldsymbol{H}(\boldsymbol{\theta})(\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{\Omega}_i\boldsymbol{H}(\boldsymbol{\theta}))^{-1}\boldsymbol{H}(\boldsymbol{\theta})'\boldsymbol{y}_i \tag{24}$$

Note that the form of $\widehat{\boldsymbol{\beta}}_{CCEGLS}$ is equivalent to the GLS estimator based on the infeasible transformation $\boldsymbol{M}_{\boldsymbol{F}}$ that takes the factors as given. As such, knowing that CCE asymptotically spans the same space as $\boldsymbol{F}$ tells us that the CCE GLS estimator considered below is asymptotically the same as the infeasible GLS estimator that assumes the factors are known.

I consider the asymptotic forms of the estimator with the limits replacing their feasible counterparts. While a formal proof of consistency is left for future work, I note that $\boldsymbol{\Omega}_i$ is assumed positive definite with probability one. As the factor loadings are assumed to have full rank for all N and as $N\to\infty$, and $p=K+1$, we would expect convergence of $(\boldsymbol{M}_{\widehat{\boldsymbol{F}}}\boldsymbol{\Omega}_i\boldsymbol{M}_{\widehat{\boldsymbol{F}}})^-$ to $(\boldsymbol{M}_{\boldsymbol{F}}\boldsymbol{\Omega}_i\boldsymbol{M}_{\boldsymbol{F}})^-$ by the argument in Karabiyik et al. (2017). The QLD GLS consistency argument is much simpler because $\boldsymbol{H}(\boldsymbol{\theta})$ is positive definite by construction for any realization of $\boldsymbol{\theta}$.

I now show that the GLS estimators are asymptotically equivalent:

Theorem 3. *Suppose Assumption CM holds, $E(\boldsymbol{y}_i\boldsymbol{y}_i'|\boldsymbol{x}_i)$ is positive definite, and $Rank(\boldsymbol{F})=p<T$. Then $\widehat{\boldsymbol{\beta}}_{CCEGLS}=\widehat{\boldsymbol{\beta}}_{QLDGLS}$.*

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

Because $\boldsymbol{H}(\boldsymbol{\theta})$ and $\boldsymbol{M_F}$ are only available asymptotically, the best we can achieve is an asymptotic equivalence result. However, there are still important finite-sample considerations that come from this result. As stated earlier, the CCE estimator of $\boldsymbol{M_F}$ also includes an additional term when $p < K + 1$ and is then not asymptotically equal to the infeasible GLS estimator. Brown and Westerlund (2022) provide a simple test for the validity of available cross-sectional averages in CCE regression. Along with the Ahn et al. (2013) tests for p using the CCE model as in Brown (2022), this test allows researchers to choose the relevant factors in estimation so that only p cross-sectional averages are used. They can then asymptotically achieve the efficiency bound studied in Theorem 3 while also removing the finite-sample variability from estimating irrelevant factors[8].

CCE also has clear benefits over QLD in terms of inference. Brown et al. (2022) show that the asymptotic variance of the CCE estimator can sometimes depend on uncertainty from estimating the factors. Thus, a valid bootstrap procedure for estimating CCE standard errors requires re-estimating the cross-sectional averages with each new bootstrap sample. The same result will naturally hold for a two-step QLD GLS estimator, but instead one needs to re-run the optimization procedure that estimates $\boldsymbol{\theta}$. Because estimation of $\boldsymbol{\theta}$ comes from a nonlinear and overidentified problem, computational costs of bootstrapping CCE are significantly lower than for QLD. It is thus easier to perform inference on a CCE GLS estimator than one based on the QLD transformation, even though both achieve the same information bound. One should care about such computational ease because the analytic standard errors for such estimators can be difficult to compute, especially when accounting for the GLS transformation.

3.3 Random Trend

I now consider a particular factor specification that is common in applied settings:

$$E(y_{it}|\boldsymbol{x}_i, c_i, a_i) = c_i + a_i t + \boldsymbol{x}_{it}\boldsymbol{\beta}_0 \tag{25}$$

Equation (25) is often called a random trend model because the outcome variable has an unobserved heterogeneous response to the observable time trend[9]. A standard technique in dealing with the heterogeneous trend is to first-difference. Define $\Delta y_{it} = y_{it} - y_{i,t-1}$ with similar definitions for $\Delta \boldsymbol{x}_{it}$ and Δu_{it}. Then

$$\Delta y_{it} = a_i + \Delta \boldsymbol{x}_{it}\boldsymbol{\beta}_0 + \Delta u_{it} \tag{26}$$

Under the strict exogeneity assumption of Assumption CM, we have $E(\Delta u_{it}|\boldsymbol{x}_i) = \mathbf{0}$ for each $t \geqslant 2$. Thus we have strictly exogenous covariates with an additive heterogeneity term. The most popular technique for

[8] A CCE estimator that uses irrelevant factor proxies will generally be less efficient than one that drops such proxies. This fact holds because CCE comes from a linear regression that estimates unit-specific slopes on the cross-sectional averages. Imposing the linear restriction that some of these slopes are zero necessarily decreases the variance of CCE.

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

estimating $\boldsymbol{\beta}_0$ in a linear model with additive heterogeneity is fixed effects estimation, which applies the within transformation, $\boldsymbol{I}_{T-1} - \frac{1}{T-1}\mathbf{1}_{T-1}\mathbf{1}'_{T-1}$ where here $\mathbf{1}_{T-1}$ is a $T-1 \times 1$ vector of ones, to the first differenced residuals $\Delta y_{it} - \Delta \boldsymbol{x}_{it}\boldsymbol{\beta}_0$.

Another way to eliminate the heterogeneity in equation (25) is to apply the first-differencing transformation again on equation (26). This technique is often referred to as second-differencing. The regression is then run for $\Delta y_{it} - \Delta y_{i,t-1}$ on $\Delta \boldsymbol{x}_{it} - \Delta \boldsymbol{x}_{i,t-1}$. Since the heterogeneous terms correspond to a known intercept and time trend, we can also run a full fixed regression on equation (25), which treats $(c_1, ..., c_N, a_1, ..., a_N)$ as parameters.

One final transformation to consider is the forward orthogonal deviations (FOD) operator in Arellano and Bover (1995). This matrix applies the following transformation to the errors u_{it} in equation (26):

$$\frac{(T-t)}{(T-t+1)}\left(u_{it} - \frac{1}{(T-t)}(u_{i,t+1} + ... + u_{it})\right) \tag{27}$$

The transformation can be written in matrix form as

$$\text{diag}(\frac{T-1}{T}, ..., \frac{1}{2})^{1/2} \times \begin{pmatrix} 1 & -(T-1)^{-1} & -(T-1)^{-1} & \dots & -(T-1)^{-1} & -(T-1)^{-1} & -(T-1)^{-1} \\ 0 & 1 & -(T-2)^{-1} & \dots & -(T-2)^{-1} & -(T-2)^{-1} & -(T-2)^{-1} \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & 0 & 0 & \dots & 0 & 1 & -1 \end{pmatrix} \tag{28}$$

I denote this FOD transformation as the matrix $\boldsymbol{F}$. For each of the first $T-1$ observations, $\boldsymbol{F}$ subtracts off a weighted mean of the rest of the independent variables. While initially studied in the context of sequential exogeneity and predetermined systems, I study it here in the context of strict exogeneity to determine information equivalence. Since I am assuming first-differencing has already occurred, I consider the $(T-2) \times (T-1)$ matrix $\boldsymbol{F}$ which corresponds to the definition in equation (28) but only assumes $T-1$ dependent variables instead of T. Regardless of the number of time periods considered, $\boldsymbol{F}$ has full row rank.

To show information equivalence of the transformations described above, let $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ be the respective $(T-1) \times T$ and $(T-2) \times (T-1)$ full rank first-differencing matrices, $\boldsymbol{W} = \boldsymbol{I}_{T-1} - \frac{1}{T-1}\mathbf{1}_{T-1}\mathbf{1}'_{T-1}$ be the $(T-1) \times (T-1)$ within transformation that has rank $T-2$, $\boldsymbol{F}$ be the $(T-2) \times (T-1)$ full rank matrix defined

similarly to equation (28), and $\boldsymbol{M}$ be the $T \times T$ residual maker matrix from regressing on $(1, t)$. Then

$$\boldsymbol{D}_2\boldsymbol{D}_1 E((\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta}_0)|\boldsymbol{x}_i) = E(\boldsymbol{D}_2\boldsymbol{D}_1(\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta}_0)|\boldsymbol{x}_i) = \boldsymbol{0} \tag{29}$$

$$\boldsymbol{W}\boldsymbol{D}_1 E((\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta}_0)|\boldsymbol{x}_i) = E(\boldsymbol{W}\boldsymbol{D}_1(\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta}_0)|\boldsymbol{x}_i) = \boldsymbol{0} \tag{30}$$

$$\boldsymbol{M} E((\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta}_0)|\boldsymbol{x}_i) = E(\boldsymbol{M}(\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta}_0)|\boldsymbol{x}_i) = \boldsymbol{0} \tag{31}$$

$$\boldsymbol{F}\boldsymbol{D}_1 E((\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta}_0)|\boldsymbol{x}_i) = E(\boldsymbol{F}\boldsymbol{D}_1(\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta}_0)|\boldsymbol{x}_i) = \boldsymbol{0} \tag{32}$$

where equations (29)-(32) correspond to the residuals from the second-differencing, first-differencing then within, first-differencing then forward orthogonal deviations, and full fixed effects transformations respectively. Thus each of the transformations satisfy Assumption MAT and so we can apply the general theory from Section 2.2.

Theorem 4. *Suppose Assumption CM holds and $E(\boldsymbol{y}_i\boldsymbol{y}_i'|\boldsymbol{x}_i)$ is positive definite. Then $\boldsymbol{D}_2\boldsymbol{D}_1$, $\boldsymbol{W}\boldsymbol{D}_1$, $\boldsymbol{F}\boldsymbol{D}_1$ and $\boldsymbol{M}$ are information equivalent.*

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

The simplicity of the proof follows from the general nature of the unified theory proved in Section 2 and thus demonstrates its usefulness. In the language of Im et al. (1999), the GLS estimators based on the residuals in equations (29)-(32) are algebraically equivalent for a given covariance matrix. Theorem 3 can thus be seen as a generalization of Theorem 4.3 of Im et al. (1999).

Finally, Phillips (2020) demonstrates that matrix inversion for estimators based on first-differencing can involve significantly more computational resources than those based on forward orthogonal deviations. He demonstrates with simulation evidence that computational time increases quickly with T even for relatively small values of N. While instruments need to satisfy two conditions given in Phillips (2020), which are not necessarily assumed here, the results in Section 2 are purely algebraic and also apply to moment functions that contain outside instruments.

4 Implementing Efficient Estimators

I now consider implementation of the efficiency bounds discussed in the paper. Given a transformation $\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta})$ satisfying Assumptions SYS and ORTH (and thus MAT) the estimator $\widehat{\boldsymbol{\beta}}_A$ that solves

$$\sum_{i=1}^{N} \nabla_{\boldsymbol{\beta}} \boldsymbol{m}_i(\boldsymbol{\beta}_0)' \boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)' (\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0) E(\boldsymbol{y}_i\boldsymbol{y}_i'|\boldsymbol{x}_i) \boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)')^{-} \boldsymbol{A}(\boldsymbol{x}_i, \widehat{\boldsymbol{\beta}}_A) \boldsymbol{y}_i = \boldsymbol{0} \tag{33}$$

is $\sqrt{N}$-asymptotically normal with asymptotic variance equal to the efficiency bound given by equation (4).

First-stage estimation of $\boldsymbol{\beta}_0$ can come from a GMM estimator with an arbitrary weight matrix. Second, one needs to consistently estimate $E(\boldsymbol{y}_i\boldsymbol{y}_i'|\boldsymbol{x}_i)$. A nonparametric regression estimator can be used in principle, but in practice this estimator may give highly imprecise estimates when T and K are relatively large. In the multiplicative heterogeneity setting, Brown and Wooldridge (2022) provides a simple and attractive parametric framework for the FEP setting. They assume $Var(y_{it}|\boldsymbol{x}_i, c_i) = \alpha E(y_{it}|\boldsymbol{x}_i, c_i)$ where $\alpha > 0$ is an identified coefficient along with a constant conditional correlation matrix.

Asymptotically justified standard errors can be derived using the familiar sample analog to the efficiency bound in (4). The researcher can then test the validity of parts of Assumption CM. For strict exogeneity, Wooldridge (2010, Chapter 18) suggests including functions of lead values of independent variables and running a joint test of significance. This method's most attractive feature is the weakness of its alternative hypothesis. The null maintains strict exogeneity while the alternative is merely that strict exogeneity fails. It is also easy to implement and can be tested in most standard statistical packages. However, there is no guidance on how to choose which regressors to include or their functional forms.

Another possible way to examine strict exogeneity is via a Hausman test. The researcher could choose a competing estimator based on the desired alternative hypothesis. In the nonlinear multiplicative example of Section 3.1, suppose the researcher believes that $E(y_{it}|\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{it}, c_i) = m_t(\boldsymbol{x}_{it}, \boldsymbol{\beta}_0)$. Then the generalized next-differencing transformation $\boldsymbol{D}_i(\boldsymbol{\beta}) = (\boldsymbol{r}_{i,1,2}(\boldsymbol{\beta}), ..., \boldsymbol{r}_{i,T-1,T}(\boldsymbol{\beta}))'$ still provides valid moment conditions. However, instruments designed to reach the efficiency bound in (4) will not be valid under sequential exogeneity alone. Chamberlain (1992b) derives the asymptotic efficiency bound for moment conditions under sequential exogeneity, and provides an implementable estimator that reaches said bound. Under the null hypothesis, both estimators are consistent, with the generalized next-differencing estimator as in (15) being asymptotically efficient. Under the alternative, only Chamberlain's instruments are valid (and in fact asymptotically efficient among $\sqrt{N}$-asymptotically normal estimators).

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

Another possible application of the results involve finite-sample and computational concerns. Suppose the researcher studying the random trend model in equation (25) has an unbalanced panel where selection into the sample is independent of $(\boldsymbol{x}_i, \boldsymbol{u}_i, c_i, a_i)$ and wants to implement the efficient estimator that makes no assumptions on the relationship between the covariates and heterogeneity[10]. Despite the fact that the second-differencing

[10]Independent selection is not necessary to apply these results. However, the efficient instruments would need to be calculated differently according to the setting. As such, I consider the simplest case for exposition.

and the full within transformations yield the same efficient estimator asymptotically, missing data causes the algebraic equivalence to break down. When an observation is missing at one time period, differencing forces the researcher to drop the adjacent period as well. The within transformation can easily be modified to only drops missing observations and demean with whatever is left, which leads to better finite sample performance.

Some transformations lead to much simpler inference than others. For example, I described in Section 3 how bootstrapping CCE standard errors is generally much less computationally intensive than QLD standard errors because of the comparative first-stage estimation problems. When it comes to strictly estimation, a similar problem holds. For example, the generalized within transformation that defines the FEP estimator contains an inverse of the sums of the means in every moment condition, while the generalized first-differencing transformation only contains the inverse of one function in each moment condition. Because the sum is the same for every moment for the FEP, it may make calculating standard errors and bootstrap quantities simpler.

5 Conclusion

This paper considers linear transformations of nonlinear panel data models with unobserved heterogeneity. When covariates are strictly exogenous in the zero conditional mean sense, such transformations provide uncountable moment conditions exploitable for estimation. I consider specifically the asymptotic efficiency bound for estimating the model's parameters. This matrix specifies a lower bound for the asymptotic variance of $\sqrt{N}$-consistent estimators.

Transformations of the data are said to be information equivalent if they yield the same asymptotic efficiency bound. The main result of Section 2 is a unified framework for comparing the efficiency bounds of such transformations. It shows that, besides regularity conditions, transformations that yield conditional moment restrictions have the same information bound as long as they have the same rank. I also simplify the form of the efficiency bound under a general and easily verifiable algebraic orthogonality property, which helps in determining other interesting relationships between instrumental variable estimators.

The theoretical framework is applied to show that the generalized within transformation, which provides the basis of the FEP estimator, is information equivalent to a number of other transformations. These transformations include generalizations of varying differencing techniques used in the linear panel data context such as next-, first-, and long-differencing, as well as the infeasible residual maker matrix from regression on the outcome variable's mean function. It is also shown that deleting any arbitrary row from the generalized within transformation does not lead to information loss.

I also generalize a result of Im et al. (1999) on linear models with additive heterogeneity to a factor-augmented error structure as studied in Pesaran (2006), Ahn et al. (2013), Westerlund et al. (2019), and Brown (2022). I show that any $T-p$ rank transformation of the data that eliminates the factors is information

equivalent to the infeasible transformation that treats the factors as known. Specifically, the QLD transformation of Ahn et al. (2013) and the CCE transformation of Pesaran (2006) are information equivalent to the infeasible fixed effects GLS estimator that treats the unobserved effects as known. I also show that in the case of a random heterogeneous trend model, first-differencing twice, first-differencing and then using a within transformation, and the true fixed effects estimator are information equivalent.

The work in this paper provides a basic framework for comparison of estimators for a broad class of nonlinear models. I primarily consider strictly exogenous covariates so that I could compare estimators using theoretically efficient instruments. However, the finite sample algebraic results hold regardless of validity of the instruments. As such, the main theorem in Section 2 can apply to any comparison of efficiency for instrumental variable estimators. There is also further work to be done considering sequential exogeneity and dynamic models. For instance, one may hope to compare various differencing techniques in estimation of dynamic linear models in the presence of additive heterogeneity.

Appendix: Proofs

Proof of Lemma 1. $\boldsymbol{A}(\boldsymbol{x}^0, \boldsymbol{\beta}_0)\boldsymbol{m}_i(\boldsymbol{\beta}_0) = \mathbf{0}$ over the supports of $\boldsymbol{x}_i$ and $\boldsymbol{c}_i$ by Assumption MAT. As $|m_t(\boldsymbol{x}_t^0, \boldsymbol{\beta}_0, \boldsymbol{c}^0)| > 0$, $\boldsymbol{A}(\boldsymbol{x}^0, \boldsymbol{\beta}_0)$ has a nontrivial null space, and hence its rank is less than T. □

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

Proof of Lemma 3. Let $p_{it}(\boldsymbol{\beta}) = m_t(\boldsymbol{x}_{it}, \boldsymbol{\beta})\left(\sum_{s=1}^T m_s(\boldsymbol{x}_{is}, \boldsymbol{\beta})\right)^{-1}$, $\boldsymbol{p}_i(\boldsymbol{\beta}) = (p_{i1}(\boldsymbol{\beta}), ..., p_{it}(\boldsymbol{\beta}))'$, and $n_i = \sum_{s=1}^T y_{is}$. Let $\mathbf{1}$ be a $T \times 1$ vector of ones. First I directly show the conclusion holds for $\boldsymbol{I}_T - \boldsymbol{p}(\boldsymbol{\beta})\mathbf{1}'$ which satisfies the lemma's assumption. It also satisfies Assumption MAT, which is made clear in Section 3. I need the following derivation:

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

Stacking the T equations gives

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

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

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

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

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

Proof of Lemma 4. Write $E(\boldsymbol{y}_i\boldsymbol{y}_i'|\boldsymbol{x}_i) = \boldsymbol{\Sigma}_i$. Then for any $T-1 \times T$ transformation $\boldsymbol{A}(\boldsymbol{x}_i, \boldsymbol{\beta}_0)$ with rank $T-1$,

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

as $\boldsymbol{\Sigma}_i^{1/2}$ is $T \times T$ and full rank. Thus the conditional variance is nonsingular and (4) holds with a proper inverse. Any generalized differencing residual with transformation satisfying Assumption RK.1 has a nonsingular conditional variance. This result goes for $\boldsymbol{Q}(\boldsymbol{I}_T - \boldsymbol{p}_i(\boldsymbol{\beta}_0)\boldsymbol{1}')$ and $\boldsymbol{Q}(\boldsymbol{I}_T - \boldsymbol{m}_i(\boldsymbol{\beta}_0)(\boldsymbol{m}_i(\boldsymbol{\beta}_0)'\boldsymbol{m}_i(\boldsymbol{\beta}_0))^{-1}\boldsymbol{m}_i(\boldsymbol{\beta}_0)$ since their full transformations have rank $T-1$. Lemma 1 of Verdier (2018) shows $Rank((\boldsymbol{I}_T - \boldsymbol{p}_i(\boldsymbol{\beta}_0)\boldsymbol{1}')) = T-1$; the rank of the residual maker transformation is a well-known result.

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

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

the conditional gradients are given as

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

by Lemma 3. Then the systems defined by Assumption SYS for both transformations are consistent with $\boldsymbol{F}(\boldsymbol{x}_i) = \boldsymbol{V}_i^{-}\nabla_{\boldsymbol{\beta}}\boldsymbol{m}_i(\boldsymbol{\beta}_0)$ and the singularity assumption in Assumption RK.2 guarantees both efficiency bounds exist. □

Proof of Theorem 2. As mentioned in the text, Assumptions CM, RK.1, RK.2, and the positive definiteness of $E(\boldsymbol{y}_i\boldsymbol{y}_i'|\boldsymbol{x}_i)$ are sufficient for each of the transformations studied to satisfy Assumptions SYS and ORTH (and thus MAT) so that their asymptotic efficiency bounds are well-defined and given by (8). Let $\boldsymbol{B}_i$ be one of the full rank $T-1 \times T$ transformation (evaluated at $\boldsymbol{x}_i$ and $\boldsymbol{\beta}_0$) studied. $\boldsymbol{B}_i$ could be the generalized within transformation, or either the generalized within or residual maker transformation with any arbitrary row deleted. I will prove the theorem by showing each of these transformations are information equivalent to the full generalized within transformation via Theorem 1, and noting that a similar proof holds for the full residual maker transformation. Write $\boldsymbol{\Sigma}_i = E(\boldsymbol{y}_i\boldsymbol{y}_i'|\boldsymbol{x}_i)$. Since each of the potential $\boldsymbol{B}_i$ matrices satisfy Assumption

ORTH, its efficiency bound is given by (8):

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

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

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

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

This expression is equal to the expression in (16) by Theorem 1, so the generalized within transformation is information equivalent to $\boldsymbol{B}_i$. The proof for the residual maker transformation is similar with $\boldsymbol{M}_i = (\boldsymbol{I}_T - \boldsymbol{m}_i(\boldsymbol{\beta}_0)(\boldsymbol{m}_i(\boldsymbol{\beta}_0)'\boldsymbol{m}_i(\boldsymbol{\beta}_0))^{-1}\boldsymbol{m}_i(\boldsymbol{\beta}_0)')$ and $\boldsymbol{V}_i$ being the respective conditional covariance matrix. □

References

Ahn, S. C., Lee, Y. H., & Schmidt, P. (2013). Panel data models with multiple time-varying individual effects. *Journal of econometrics*, *174*(1), 1–14.

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

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

Breusch, T., Qian, H., Schmidt, P., & Wyhowski, D. (1999). Redundancy of moment conditions. *Journal of Econometrics*, *91*(1), 89–111.

Brown, N. L. (2022). *Moment-based estimation of linear panel data models with factor-augmented errors* (tech. rep.).

Brown, N. L., & Westerlund, J. (2022). *Testing factors in cce* (tech. rep.).

Brown, N. L., & Wooldridge, J. M. (2022). *More efficient estimation of multiplicative panel data models in the presence of serial correlation* (tech. rep.).

Brown, N. L., Wooldridge, J. M., & Schmidt, P. (2022). *Simple alternatives to the common correlated effects model* (tech. rep.).

Castillo, J. C., Mejía, D., & Restrepo, P. (2020). Scarcity without leviathan: The violent effects of cocaine supply shortages in the mexican drug war. *Review of Economics and Statistics*, (0).

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

Chamberlain, G. (1992). Efficiency bounds for semiparametric regression. *Econometrica: Journal of the Econometric Society*, *60*(3), 567–596.

Fischer, S., Royer, H., & White, C. (2018). The impacts of reduced access to abortion and family planning services on abortions, births, and contraceptive purchases. *Journal of Public Economics*, *167*, 43–68.

Hahn, J. (1997). A note on the efficient semiparametric estimation of some exponential panel models. *Econometric Theory*, *13*(4), 583–588.

Hausman, J., Hall, B. H., & Griliches, Z. (1984). Econometric models for count data with an application to the patents-r&d relationship. *Econometrica: Journal of the Econometric Society*, *52*(4), 909–938.

Im, K. S., Ahn, S. C., Schmidt, P., & Wooldridge, J. M. (1999). Efficient estimation of panel data models with strictly exogenous explanatory variables. *Journal of Econometrics*, *93*(1), 177–201.

Karabiyik, H., Reese, S., & Westerlund, J. (2017). On the role of the rank condition in cce estimation of factor-augmented panel regressions. *Journal of Econometrics*, *197*(1), 60–64.

Krapf, M., Ursprung, H. W., & Zimmermann, C. (2017). Parenthood and productivity of highly skilled labor: Evidence from the groves of academe. *Journal of Economic Behavior & Organization*, *140*, 147–175.

McCabe, M. J., & Snyder, C. M. (2014). Identifying the effect of open access on citations using a panel of science journals. *Economic Inquiry*, *52*(4), 1284–1300.

McCabe, M. J., & Snyder, C. M. (2015). Does online availability increase citations? theory and evidence from a panel of economics and business journals. *Review of Economics and Statistics*, *97*(1), 144–165.

Newey, W. K. (2001). Conditional moment restrictions in censored and truncated regression models. *Econometric Theory*, *17*(5), 863–888.

Pesaran, M. H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica*, *74*(4), 967–1012.

Phillips, R. F. (2020). Quantifying the advantages of forward orthogonal deviations for long time series. *Computational Economics*, *55*(2), 653–672.

Rao, C. R., & Mitra, S. K. (1972). Generalized inverse of a matrix and its applications. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Theory of Statistics*.

Schlenker, W., & Walker, W. R. (2016). Airports, air pollution, and contemporaneous health. *The Review of Economic Studies*, *83*(2), 768–809.

Verdier, V. (2018). Local semi-parametric efficiency of the poisson fixed effects estimator. *Journal of Econometric Methods*, *7*(1).

Westerlund, J. (2020). A cross-section average-based principal components approach for fixed-t panels. *Journal of Applied Econometrics*, *35*(6), 776–785.

Westerlund, J., Petrova, Y., & Norkutė, M. (2019). Cce in fixed-t panels. *Journal of Applied Econometrics*, *34*(5), 746–761.

Westerlund, J., & Urbain, J.-P. (2013). On the estimation and inference in factor-augmented panel regressions with correlated loadings. *Economics Letters*, *119*(3), 247–250.

Williams, M. L., Burnap, P., Javed, A., Liu, H., & Ozalp, S. (2020). Hate in the machine: Anti-black and anti-muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology*, *60*(1), 93–117.

Wooldridge, J. M. (1997). Multiplicative panel data models without the strict exogeneity assumption. *Econometric Theory*, *13*(5), 667–678.

Wooldridge, J. M. (1999). Distribution-free estimation of some nonlinear panel data models. *Journal of Econometrics*, *90*(1), 77–97.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data* (2nd ed., Vol. 1). MIT press.

Wooldridge, J. M. (2022). *Simple approaches to nonlinear difference-in-differences with panel data* (tech. rep.).