Panel Treatment Effects Measurement - Factor or Linear Projection Modelling?*

Cheng Hsiao† Qiankun Zhou‡

This version October 04 2023

Abstract

We discuss methods of measuring the treatment effects of a unit through the use of other units in panel data by either the factor approach or the linear projection approach under different sample configuration of cross-sectional dimension N and time series dimension T. We derive the prediction intervals of factor approach or linear projection approach when either both N and T are large, or N fixed and $T \to \infty$, or T fixed and N large, or both N and T are finite. Monte Carlo simulation and empirical example are also conducted to consider their finite sample performances.

Keywords: Treatment effects, Linear projection, Factor approach, Panel data analysis, Interactive effects, Confidence interval.

JEL classification: C01, C21, C31

*We would like to thank the editor, Edward Vytlacil, two anonymous referees, Jing Kong for helpful comments. Partial research support of China NSF grant #72033008 to the first author is also gratefully acknowledged.

†Department of Economics, University of Southern California, University Park, Los Angeles, California 90089; Paula and Gregory Chow Institute for Studies in Economics, Xiamen University, China. Email: chsiao@usc.edu.

‡Department of Economics, Louisiana State University, Baton Rouge, LA 70803. Email: qzhou@lsu.edu.

1 Introduction

Factor models (e.g., Lawley and Maxwell (1971)) are popular analytic tools for high dimensional probability models. They are widely applied to macro and financial economics (e.g., Chamberlain and Rothschild (1983), Connor and Korajczk (1986)), Forni and Reichlin (1998), Ross (1976), Sargent and Sims (1977)), and are also used to generate parsimonious predictive models from high dimensional time series data (e.g., Stock and Watson, (1989, 2002)). However, factors are unobserved, they need to be estimated from the observed data. Moreover, future shocks at time $t+h$, is unknown, nor are the future outcomes $\mathbf{y}_{t+h}$ at time t. This raises the question that from predictive accuracy perspective, whether the factor model is likely to generate more accurate predictions than simply to use a subset of the observed data directly? To bypass the complicated issues of how to obtain future common factors from unknown future observed data, we consider the issues of measuring (or predicting) the treatment effects when there exist both the pre-treatment and post-treatment data to identify the underlying factor models or linear projection model. A frequently used measure of the accuracy of a predictor $\hat{\theta}$ is to consider the length of its $(1-\alpha)$ confidence interval, $\hat{\theta} \pm c_\alpha \sigma_{\hat{\theta}}$, where $\sigma_{\hat{\theta}}$ is the standard error of $\hat{\theta}$ and c_α denotes the critical value two-sided size α test. However, if $\hat{\theta}$ is a biased predictor, the length of $\hat{\theta} \pm c_\alpha \sigma_{\hat{\theta}}$ is grossly misleading. Therefore, in this paper, we consider the predictive accuracy issues using the criteria of the bias and prediction error variance.

The rest of the paper is organized as follows. Section 2 discusses modelling a unit in a panel data set by a factor model or a linear projection model. Section 3 discusses model estimation and prediction. Section 4 considers their application in the measurement of treatment effects. Section 5 discusses the pros and cons of factor based approach (FB) and linear projection approach (LP) in terms of bias and prediction error variance under different sample configurations, namely, when both the cross-sectional dimension (N) and time series dimension (T) are large and $\frac{N}{T} \to a \neq 0 < \infty$ as $(N,T) \to \infty$, N fixed and $T \to \infty$, T fixed and N large, or both N and T are finite. Generalization to model with exogenous regressors are discussed in Section 6. Monte Carlo are reported in Section 7. An empirical application to the impact of the Germany reunion on GDP per capita is provided in Section 8. Concluding remarks are in Section 9. All mathematical proofs are presented in the appendix.

2 Modelling Panel Data by a Factor Model or a Linear Projection Model

Suppose the N cross-sectional units at time t can be modelled by a factor model of the form

$$y_{it} = \boldsymbol{\lambda}_i'\mathbf{f}_t + u_{it}, \qquad i = 1, \cdots, N; t = 1, \ldots, T, \tag{2.1}$$

where $\mathbf{f}_t$ is r dimensional common factors and $\boldsymbol{\lambda}_i$ is r dimensional factor loadings for the i-th unit,[1] u_{it} is the random error term with mean zero.

Let $\Lambda = (\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots, \boldsymbol{\lambda}_N)' = (\boldsymbol{\lambda}_1, \tilde{\Lambda})'$, $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_T)'$, $\mathbf{u}_i = (u_{i1}, \ldots, u_{iT})'$, and $\mathbf{u}_t = (u_{1t}, u_{2t}, \ldots, u_{Nt})' = (u_{1t}, \tilde{\mathbf{u}}_t')'$. Stacking all N cross-sectional units one after another at time t, $\mathbf{y}_t = (y_{1t}, y_{2t}, \ldots, y_{Nt})' = (y_{1t}, \tilde{\mathbf{y}}_t')'$, we have

$$\mathbf{y}_t = \Lambda\mathbf{f}_t + \mathbf{u}_t, \qquad t = 1, \ldots, T. \tag{2.2}$$

Alternatively, we can stack ith individual's T time series observations as $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT})'$, then

$$\mathbf{y}_i = \mathbf{F}\boldsymbol{\lambda}_i + \mathbf{u}_i, \qquad i = 1, \ldots, N. \tag{2.3}$$

For model (2.1), we assume:

Assumption 1: The factor process satisfies $E\left\|\mathbf{f}_t\right\|^4 \leq M < \infty$ and $\frac{1}{T}\sum_{t=1}^{T}\mathbf{f}_t\mathbf{f}_t' \to \Sigma_f$, where Σ_f is an $r \times r$ nonsingular constant matrix.[2]

Assumption 2: The loading $\boldsymbol{\lambda}_i$ is either fixed constant or it is stochastic with $E\left\|\boldsymbol{\lambda}_i\right\|^4 \leq M < \infty$. In either case, $\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\lambda}_i\boldsymbol{\lambda}_i' \to \Sigma_\lambda$, where Σ_λ is an $r \times r$ nonsingular constant matrix.

and either

Assumption 3: The random error terms $\mathbf{u}_t$ is independently identically distributed over t with nonsingular covariance matrix

$$E(\mathbf{u}_t\mathbf{u}_t') = \tilde{\Omega} = \begin{pmatrix} \sigma_1^2 & \mathbf{c}' \\ \mathbf{c} & \Omega \end{pmatrix}, \tag{2.4}$$

where $\sigma_1^2 = E\left(u_{1t}^2\right)$, $\Omega = E(\tilde{\mathbf{u}}_t\tilde{\mathbf{u}}_t')$, and $\mathbf{c} = E(\tilde{\mathbf{u}}_t u_{1t})$. Moreover, all N nonzero eigenvalues of $\tilde{\Omega}$ are $O(1)$.

Assumptions 1 and 2 are standard assumptions for factor models (e.g., Lawley and Maxwell (1971), Bai (2003), Bai and Ng (2002), etc.). They allow $\boldsymbol{\lambda}_i$ and $\mathbf{f}_t$ to be either fixed or random. However, our analysis below is conditional on observed y_{it}. Instead of expressing $\boldsymbol{\lambda}_i$

[1] In the event r is unknown, one can use Bai and Ng's (2002) information criteria to identify r.

[2] We let $\|\mathbf{A}\| = \sqrt{tr\left(\mathbf{A}\mathbf{A}'\right)}$.

as $E(\boldsymbol{\lambda}_i|\mathbf{y}_t)$ and $\mathbf{f}_t$ as $E(\mathbf{f}_t|\mathbf{y}_t)$, for notational ease, we shall simply consider $\boldsymbol{\lambda}_i$ and $\mathbf{f}_t$ as fixed constants, and consider $\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\lambda}_i\boldsymbol{\lambda}_i'$ as Σ_λ and $\frac{1}{T}\sum_{t=1}^{T}\mathbf{f}_t\mathbf{f}_t'$ as Σ_f. Assumption 3 allows $\mathbf{u}_t$ to be heteroskedastic and weakly cross-correlated but independent over t. This assumption can be generalized to cover u_{it} both weakly cross-corrlated and weakly time-dependent.[3]

Model (2.1) can be alternatively written as a linear projection (LP) model where y_{it} is a linear function of $(y_{1t},\ldots,y_{i-1,t},y_{i+1,t},\ldots,y_{Nt})$ (e.g., Hsiao et. al (2012), Li and Bell (2017)). Since cross-sectional labelling is arbitrary, there is no loss of generality to consider the first unit y_{1t} as a linear function of the rest of y_{it},

$$\begin{aligned} y_{1t} &= E(y_{1t}|\tilde{\mathbf{y}}_t) + \eta_t \\ &= \mathbf{w}'\tilde{\mathbf{y}}_t + \eta_t, \end{aligned} \tag{2.5}$$

where $\tilde{\mathbf{y}}_t = (y_{2t},\ldots,y_{Nt})'$ and $E(y_{1t}|\tilde{\mathbf{y}}_t)$ denotes the conditional mean of y_{1t} if the conditional mean is linear in $\tilde{\mathbf{y}}_t$ (e.g., $\mathbf{y}_t$ is Guassian) or the linear projection if the conditional mean is nonlinear in $\tilde{\mathbf{y}}_t$. The coefficient $\mathbf{w}$ related to the underlying factor model (2.1) is

$$\begin{aligned} \mathbf{w} &= \left[E(\tilde{\mathbf{y}}_t\tilde{\mathbf{y}}_t')\right]^{-1} E(\tilde{\mathbf{y}}_t y_{1t}) \\ &= (\tilde{\Lambda}\Sigma_f\tilde{\Lambda}' + \Omega)^{-1}(\tilde{\Lambda}\Sigma_f\boldsymbol{\lambda}_1 + \mathbf{c}), \end{aligned} \tag{2.6}$$

with $\tilde{\Lambda}$ denoting the factor loading matrix for control units $\tilde{\mathbf{y}}_t$.

The error term, η_t, by construction is orthogonal to $\tilde{\mathbf{y}}_t$ with mean 0. Assumptions 1-3 implies that η_t is independent over t with variance σ_η^2, which expressed in the form of factor model (2.1) takes the form,

$$\begin{aligned} \sigma_\eta^2 &= E[y_{1t} - \mathbf{w}'\tilde{\mathbf{y}}_t]^2 \\ &= E[\boldsymbol{\lambda}_1'\mathbf{f}_t + u_{1t} - (\boldsymbol{\lambda}_1'\Sigma_f\tilde{\Lambda}' + \mathbf{c}')(\tilde{\Lambda}\Sigma_f\tilde{\Lambda}' + \Omega)^{-1}(\tilde{\Lambda}\mathbf{f}_t + \tilde{\mathbf{u}}_t)]^2 \\ &= E\left[\left(\boldsymbol{\lambda}_1' - (\boldsymbol{\lambda}_1'\Sigma_f\tilde{\Lambda}' + \mathbf{c}')(\tilde{\Lambda}\Sigma_f\tilde{\Lambda}' + \Omega)^{-1}\tilde{\Lambda}\right)\mathbf{f}_t + u_{1t} - (\boldsymbol{\lambda}_1'\Sigma_f\tilde{\Lambda}' + \mathbf{c}')(\tilde{\Lambda}\Sigma_f\tilde{\Lambda}' + \Omega)^{-1}\tilde{\mathbf{u}}_t\right)\right]^2 \\ &= \sigma_1^2 + \boldsymbol{\lambda}_1'\Sigma_f\boldsymbol{\lambda}_1 - (\boldsymbol{\lambda}_1'\Sigma_f\tilde{\Lambda}' + \mathbf{c}')(\tilde{\Lambda}\Sigma_f\tilde{\Lambda}' + \Omega)^{-1}(\tilde{\Lambda}\Sigma_f\boldsymbol{\lambda}_1 + \mathbf{c}). \end{aligned} \tag{2.7}$$

3 Model Estimation and Prediction

Neither Λ, $\mathbf{F}$, nor $\mathbf{w}$ are known, they have to be estimated. In the case of factor model (2.1), Λ and $\mathbf{F}$ are unidentified under Assumption 1 and 2. However, since we are concerned with

[3] Bai (2003 Assumption C) actually allows u_{it} to be both weakly cross-correlated and weakly time dependent. Although we can derive similar results under Bai (2003) Assumption C, they involve a much more complicated algebraic manipulation without shedding much insight than the results demonstrated under the more restrictive assumption 3 and 4 here. See also Remark 5.2.

predicting y_{1t}, not structural identification of Λ and $\mathbf{F}$, there is no loss of generality to follow Anderson and Rubin (1956), Bai (2009), etc., to assume either $\Sigma_f = \mathbf{I}_r$ and Σ_λ being a diagonal matrix or $\Sigma_\lambda = \mathbf{I}_r$ and Σ_f being a diagonal matrix. Under the normalization $\Sigma_f = \mathbf{I}_r$, when $N > T$, $\mathbf{F}$ can be estimated as $\sqrt{T}$ times the r eigenvectors corresponding to the r largest eigenvalues of the determinant equation

$$\left| \frac{1}{N} \sum_{i=1}^{N} \mathbf{y}_i \mathbf{y}_i' - \delta \mathbf{I}_T \right| = 0, \tag{3.1}$$

where δ denotes the eigenvalues of (3.1), and Λ can be estimated as

$$\hat{\boldsymbol{\lambda}}_i = \frac{1}{T} \hat{\mathbf{F}}' \mathbf{y}_i, \quad i = 1, \ldots, N. \tag{3.2}$$

Alternatively, one can use the normalization $\Sigma_\lambda = \mathbf{I}_r$ and estimate Λ as $\sqrt{N}$ times the r eigenvectors that corresponding to the r largest eigenvalues of the determinant equation

$$\left| \frac{1}{T} \sum_{t=1}^{T} \mathbf{y}_t \mathbf{y}_t' - \delta^* \mathbf{I}_N \right| = 0, \tag{3.3}$$

when $T > N$, where δ^* denotes the eigenvalues of (3.3),. Then $\mathbf{F}$ can be estimated as

$$\hat{\mathbf{F}}' = \frac{1}{N} \hat{\Lambda}' \left(\mathbf{y}_1, \ldots, \mathbf{y}_T \right). \tag{3.4}$$

Remark 3.1 *As long as both $T > r$ and $N > r$ hold, in principal, one can use either (3.1) or (3.3) to estimate either Λ or $\mathbf{F}$. However, as shown by Bai et. al (2016), the asymptotic variance of $\hat{\mathbf{F}}$ and $\hat{\Lambda}$ are smaller using (3.1) and (3.2) than using (3.3) and (3.4) if $N > T$. On the other hand, using (3.3) and (3.4) yields smaller asymptotic variance of $\hat{\Lambda}$ and $\hat{\mathbf{F}}$ than using (3.3) and (3.4) when $T > N$.*

Given $\hat{\boldsymbol{\lambda}}_1$ and $\hat{\mathbf{f}}_t$, y_{1t} can be predicted by (e.g., Bai and Ng (2021) and Li et. al (2023))

$$\tilde{y}_{1,t} = \hat{\boldsymbol{\lambda}}_1' \hat{\mathbf{f}}_t, \tag{3.5}$$

where under the normalization $\Sigma_f = \mathbf{I}_r$, $\hat{\boldsymbol{\lambda}}_1$ is obtained by (3.2) and $\hat{\mathbf{f}}_t$ is obtained from (3.1), while under the normalization $\Sigma_\lambda = \mathbf{I}_r$, $\hat{\boldsymbol{\lambda}}_1$ is the first row of $\hat{\Lambda}$ and $\hat{\mathbf{f}}_t$ is obtained by (3.4).

Similarly, we can construct the post-treatment prediction of $y^0_{1,T+h}$ (e.g., the outcomes in the absence of the treatment at time $T + h$) as

$$\tilde{y}_{1,T+h} = \hat{\boldsymbol{\lambda}}_1' \hat{\mathbf{f}}_{T+h}, \quad h = 1, \ldots, m, \tag{3.6}$$

conditional on $\hat{\Lambda}$ and $\tilde{\mathbf{y}}_t = (y_{2t}, \ldots, y_{Nt})'$. Let $\widehat{\tilde{\Lambda}}$ denotes the rows of $\hat{\Lambda}$ from the second to the N-th, then[4]

$$\hat{\mathbf{f}}_{T+h} = \left(\widehat{\tilde{\Lambda}}'\widehat{\tilde{\Lambda}}\right)^{-1} \widehat{\tilde{\Lambda}}'\tilde{\mathbf{y}}_{T+h} \qquad \text{for} \quad h = 1, \ldots, m. \tag{3.7}$$

Let

$$\tilde{e}_{1,T+h} = y_{1,T+h} - \tilde{y}_{1,T+h} = u_{1,T+h} + \left(\boldsymbol{\lambda}_1'\mathbf{f}_{T+h} - \hat{\boldsymbol{\lambda}}_1'\hat{\mathbf{f}}_{T+h}\right). \tag{3.8}$$

The asymptotic variance of $\tilde{e}_{1,T+h}$ when $(N,T) \to \infty$ as shown by Bai (2003) takes the form,

$$Var\left(\tilde{e}_{1,T+h}\right) = \sigma_1^2 + \frac{1}{N}\boldsymbol{\lambda}_1'\Sigma_\lambda^{-1}\left(\frac{1}{N}\Lambda'\Omega\Lambda\right)\Sigma_\lambda^{-1}\boldsymbol{\lambda}_1 + \frac{\sigma_1^2}{T}\mathbf{f}_{T+h}'\Sigma_f^{-1}\mathbf{f}_{T+h} + o\left(1\right). \tag{3.9}$$

In the case of linear projection model (2.5), there are huge literature on how best to estimate $\mathbf{w}$ (e.g., Draper and Van Nostrand (1979), Kwon (2021), Smith and Campell (1980), Stein (1981), Vinod (1978), etc.,). However, we are not concerned with how best to estimate $\mathbf{w}$, but how best to obtain the predictor $\hat{y}_{1,t} = \hat{\mathbf{w}}'\tilde{\mathbf{y}}_t$ with minimum mean square error,

$$S_y = \sum_{t=1}^{T}\left(y_{1,t} - \mathbf{w}'\tilde{\mathbf{y}}_t\right)^2. \tag{3.10}$$

Taking partial derivative of S_y with respect to $\mathbf{w}$ yields $\hat{\mathbf{w}}$ as a least squares estimator,

$$\hat{\mathbf{w}} = \left(\sum_{t=1}^{T}\tilde{\mathbf{y}}_t\tilde{\mathbf{y}}_t'\right)^{-1}\left(\sum_{t=1}^{T}\tilde{\mathbf{y}}_t y_{1t}\right). \tag{3.11}$$

when $T > N > r$.[5] Conditional on $\mathbf{Y}_T = (\mathbf{y}_1, \ldots, \mathbf{y}_T)$, $\hat{\mathbf{w}}$ is unbiased.

Multiplying $\hat{\mathbf{w}}$ with $\tilde{\mathbf{y}}_{T+h}$ yields the unbiased predictor of $y_{1,T+h}$ conditional on $\mathbf{Y}_T$ and $\tilde{\mathbf{y}}_{T+h}$ as

$$\hat{y}_{1,T+h} = \hat{\mathbf{w}}'\tilde{\mathbf{y}}_{T+h}, \quad h = 1, \ldots, m. \tag{3.12}$$

Let

$$\begin{aligned} \hat{e}_{1,T+h} &= y_{1,T+h} - \hat{y}_{1,T+h} \\ &= \eta_{T+h} + \left(\mathbf{w}' - \hat{\mathbf{w}}'\right)\tilde{\mathbf{y}}_{T+h}. \end{aligned} \tag{3.13}$$

Conditional on $\mathbf{Y}_T$ and $\tilde{\mathbf{y}}_{T+h}$, the prediction error variance is equal to

$$\begin{aligned} Var\left(\hat{e}_{1,T+h}\right) &= \sigma_\eta^2 + \tilde{\mathbf{y}}_{T+h}'Var\left(\hat{\mathbf{w}}\right)\tilde{\mathbf{y}}_{T+h} \\ &= \sigma_\eta^2\left[1 + \tilde{\mathbf{y}}_{T+h}'\left(\sum_{t=1}^{T}\tilde{\mathbf{y}}_t\tilde{\mathbf{y}}_t'\right)^{-1}\tilde{\mathbf{y}}_{T+h}\right]. \end{aligned} \tag{3.14}$$

[4] A referee has suggested to consider James-Stein estimator or empirical Bayes estimator for $\mathbf{f}_t$ under the assumption that $\mathbf{f}_t \sim N(0, I)$ and $\tilde{\mathbf{y}}_t|\mathbf{f}_t \sim N\left(\tilde{\Lambda}\mathbf{f}_t, \Omega\right)$. The Stein estimator is biased but minimizes the risk. We hope to investigate the pros and cons of these estimators vs (3.7) in the future.

[5] We will discuss the LP approach when $N > T$ in subsection 5.3.

4 Treatment Effects Measurement

The treatment effects are typically measured as the difference between the outcomes under the treatment, y_{it}^1, and the outcomes in the absence of the treatment, y_{it}^0,

$$\Delta_{it} = y_{it}^1 - y_{it}^0. \tag{4.1}$$

Panel data, contrary to the cross-sectional data, provides the possibility to estimate the evolution of the treatment effects over time (e.g., Hsiao et. al (2012), Ke and Hsiao (2022)). However, the observed data for the ith individual at time t takes the form,

$$y_{it} = d_{it} y_{it}^1 + (1 - d_{it}) y_{it}^0, \quad i = 1, \cdots, N, t = 1, \cdots, T, \tag{4.2}$$

where d_{it} denotes the treatment status dummy with $d_{it} = 1$ if the ith individual at time t is under the treatment and 0 if not. Then to provide an estimate of the treatment effect, Δ_{it}, one needs to substitute the missing y_{it}^1 or y_{it}^0 by its predicted value.

Assuming y_{it}^1 is observed but not y_{it}^0, then

$$\hat{\Delta}_{it} = y_{it}^1 - \hat{y}_{it}^0, \tag{4.3}$$

where $\hat{y}_{it}^0$ denotes the predicted value (or the counterfactuals) of y_{it}^0. Conditional on observed y_{it}^1, the point estimates and confidence interval of the estimated $\hat{\Delta}_{it}$ depend on

$$\begin{aligned} E\left(\hat{\Delta}_{it}\right) &= E(y_{it}^1 - \hat{y}_{it}^0) = (y_{it}^1 - y_{it}^0) + E(y_{it}^0 - \hat{y}_{it}^0) \\ &= \Delta_{it} + E\left(y_{it}^0 - \hat{y}_{it}^0\right), \end{aligned} \tag{4.4}$$

and

$$Var(\hat{\Delta}_{it}) = E(y_{it}^0 - \hat{y}_{it}^0)^2. \tag{4.5}$$

In other words, the bias and variance of $\hat{\Delta}_{it}$ depend on the bias and error variance of $\hat{y}_{it}^0$ (or $\hat{y}_{it}^1$ if y_{it}^1 is observed), but not y_{it}^1. Econometricians and statisticians have suggested various methods to control the bias in the observed data due to selection on observables or unobservables under various assumptions (e.g., Heckman (1997), Heckman and Vytlacil (2001), Imbens and Angrist (1994), Imbens and Lemieux (2008), Rosenbaum and Rubin (1983, 1985)).

In this paper, we consider the panel data approach suggested by Hsiao et. al (2012). We consider the measurement of treatment effects on the first unit. We assume $y_{it} = y_{it}^0$ for $i = 1, \ldots, N$ and $t = 1, \ldots, T$. From periods $T + 1$ onwards, $y_{1t} = y_{1t}^1$ for $t = T + 1, \ldots, T + m$, while all other units remain untreated, and their outcomes are independent of d_{1t}, namely

$$f\left(y_{it} | d_{1t}\right) = f\left(y_{it}\right), \tag{4.6}$$

such that $y_{it} = y_{it}^0$ for $i = 2, \ldots, N$ and $t = 1, \ldots, T+m$ (e.g., Hsiao et. al (2012)). For ease of notation, we use y_{it} to denote y_{it}^0.

Even though the issues of bias and variance of treatment effects are the issues of prediction bias and variance of y_{it}^0, there is a fundamental difference between the prediction of treatment effects and the usual post-sample predictions of y_{it}^0. The usual post-sample prediction of y_{it}^0 assumes there is no post-sample information while the treatment effects measurement using panel data assume there are post-sample information of $y_{it}^0 = \boldsymbol{\lambda}_i'\mathbf{f}_t + u_{it}$ for $i = 2, \ldots, N$, which can be used to improve the accuracy of predicting $y_{1,T+h}^0$ for $h = 1, \ldots, m$. That is, we can use principal component approach to estimate the realized $\boldsymbol{\lambda}_i$ and $\mathbf{f}_t$ for model (2.1) or regression method to estimate $\mathbf{w}$ for model (2.5), then combine the post-treatment information of $y_{i,T+h}$ to generate post-treatment factor predictor (3.6) or post-treatment LP predictor (3.12).

Remark 4.1 *Our analysis is based on a single treated unit. When there are multiple treated units, in principle, one can use either FB or LP approach one by one, or first aggregate all treated units into one single unit, then use either FB or LP approach to generate the predictions for the aggregated unit. As long as one method is likely to generate more accurate predictions for any treated unit, aggregate more accurately predicted units is likely to generate more accurate aggregated predictions for whatever linear aggregation method is used (e.g., Hsiao (2021)). However, these could be computationally laborious if there are many treated units. An alternative approach could be to first aggregate multiple treated units into a single unit, then use either the LP or FB approach to generate the predictions for the aggregated units. Although aggregation can give rise to complicated issues (e.g., Hsiao et. al (2021b)), as long as Assumptions 1-3 hold for the factor modelling of the aggregated model, the analytical results in the following sections still hold.*

5 Treatment Effects Prediction Intervals

When $\boldsymbol{\lambda}_1$, $\mathbf{f}_{T+h}$, $\mathbf{w}$ and $\tilde{\mathbf{y}}_{T+h}$ are known, the post-treatment predictor[6]

$$\tilde{y}_{1,T+h} = \boldsymbol{\lambda}_1'\mathbf{f}_{T+h}, \quad h = 1, \ldots, m, \tag{5.1}$$

or

$$\hat{y}_{1,T+h} = \mathbf{w}'\tilde{\mathbf{y}}_{T+h}, \quad h = 1, \ldots, m, \tag{5.2}$$

are unbiased predictors with prediction error variance σ_1^2 and σ_η^2, respectively.

[6]The LP or FB prediction formula for $y_{1,T+h}$ is the same for any h under Assumption 3. In particular, one can just consider $h = 1$.

Under Assumption 1 and 2, Λ and $\mathbf{F}$ can be either random or fixed. Since our comparison of the length of confidence interval is conditional on observed $\mathbf{Y}_T$ and $\tilde{\mathbf{y}}_{T+h}$, there is no loss of generality to treat Λ and $\mathbf{F}$ as fixed constants instead of expressing it as realized $\boldsymbol{\lambda}_i$ and $\mathbf{f}_t$ conditional on y_{it}. Substituting unknown $\boldsymbol{\lambda}_1$, $\mathbf{f}_t$ or $\mathbf{w}$ by $\hat{\boldsymbol{\lambda}}_1$, $\hat{\mathbf{f}}_t$ or $\hat{\mathbf{w}}$ into (5.1) or (5.2) in lieu of $\boldsymbol{\lambda}_1$, $\mathbf{f}_t$ or $\mathbf{w}$ may or may not yield unbiased predictors. The LP predictor (3.12) is unbiased conditional on $\mathbf{Y}_T$ and $\tilde{\mathbf{y}}_{T+h}$, because the estimator (3.11) is unbiased and converges to $\mathbf{w}$ as $T \to \infty$. The prediction error variance conditional on $\mathbf{Y}_T$ and $\tilde{\mathbf{y}}_{T+h}$ is given by (3.14). On the other hand, the FB predictor (3.6) is the product of $\hat{\boldsymbol{\lambda}}_1$ and $\hat{\mathbf{f}}_{T+h}$. To get a good estimator of $\hat{\boldsymbol{\lambda}}_1$, we need $T \to \infty$. To get a good estimator of $\hat{\mathbf{f}}_{T+h}$, we need $N \to \infty$. Thus, whether (3.6) is asymptotically unbiased or its prediction error variance depend on the configuration of N and T.

5.1 Case 1: $(N, T) \to \infty$

Proposition 5.1 *Under Assumptions 1-3, when $(N, T) \to \infty$ and $\frac{N}{T} \to a \neq 0 < \infty$, both (3.6) and (3.12) are unbiased predictors of y_{1t} with the same asymptotic variance.*

The equivalence between (3.6) and (3.12) can be seen as follows. When $(N, T) \to \infty$, $\hat{\Lambda}$, $\hat{\mathbf{F}}$ and $\hat{\mathbf{w}}$ are consistent. The prediction error variance for FB approach (3.6) under the normalization $\Sigma_f = \mathbf{I}_r$ is (e.g., Bai (2003))

$$Var\left(\tilde{e}_{1,T+h}\right) = \sigma_1^2 + \boldsymbol{\lambda}_1' \left(\tilde{\Lambda}'\tilde{\Lambda}\right)^{-1} \left(\tilde{\Lambda}'\Omega\tilde{\Lambda}\right) \left(\tilde{\Lambda}'\tilde{\Lambda}\right)^{-1} \boldsymbol{\lambda}_1 + \frac{1}{T}\sigma_1^2 \mathbf{f}_{T+h}' \mathbf{f}_{T+h} + O\left(\frac{1}{\min\left(T, N\right)}\right). \quad (5.3)$$

Under the assumption that u_{it} is independent over i and t, $\mathbf{c} = E(\tilde{\mathbf{u}}_t u_{1t}) = \mathbf{0}$, the LP prediction error expressed in the factor form is[7]

$$\begin{aligned} \hat{e}_{1,T+h} &= \boldsymbol{\lambda}_1' \mathbf{f}_{T+h} + u_{1,T+h} - \left(\tilde{\Lambda}\boldsymbol{\lambda}_1\right)' \left(\tilde{\Lambda}\tilde{\Lambda}' + \Omega\right)^{-1} \left(\tilde{\Lambda}\mathbf{f}_{T+h} + \tilde{\mathbf{u}}_{T+h}\right) + O_p\left(\frac{1}{\sqrt{T}}\right) \\ &= \boldsymbol{\lambda}_1' \left(\mathbf{I}_r - \tilde{\Lambda}' \left(\tilde{\Lambda}\tilde{\Lambda}' + \Omega\right)^{-1} \tilde{\Lambda}\right) \mathbf{f}_{T+h} + u_{1,T+h} - \left(\tilde{\Lambda}\boldsymbol{\lambda}_1\right)' \left(\tilde{\Lambda}\tilde{\Lambda}' + \Omega\right)^{-1} \tilde{\mathbf{u}}_{T+h} + O_p\left(\frac{1}{\sqrt{T}}\right), \end{aligned} \quad (5.4)$$

[7] The derivation of prediction error variance remains the same under the normalization $\Sigma_\lambda = \mathbf{I}_r$. All it is needed is to replace $\tilde{\Lambda}$ in the formula with $\Lambda^* = \tilde{\Lambda}\Sigma_f^{1/2}$.

with error variance of prediction

$$\begin{aligned} Var\left(\hat{e}_{1,T+h}\right) &= \left[\boldsymbol{\lambda}_1'\left(\mathbf{I}_r - \tilde{\Lambda}'\left(\tilde{\Lambda}\tilde{\Lambda}' + \Omega\right)^{-1}\tilde{\Lambda}\right)\right] E\left(\mathbf{f}_{T+h}\mathbf{f}_{T+h}'\right)\left[\left(\mathbf{I}_r - \tilde{\Lambda}'\left(\tilde{\Lambda}\tilde{\Lambda}' + \Omega\right)^{-1}\tilde{\Lambda}\right)\boldsymbol{\lambda}_1\right] \\ &\quad +\sigma_1^2 + \boldsymbol{\lambda}_1'\tilde{\Lambda}'\left(\tilde{\Lambda}\tilde{\Lambda}' + \Omega\right)^{-1}\Omega\left(\tilde{\Lambda}\tilde{\Lambda}' + \Omega\right)^{-1}\left(\tilde{\Lambda}\boldsymbol{\lambda}_1\right) + O\left(\frac{1}{T}\right) \\ &= \sigma_1^2 + \boldsymbol{\lambda}_1'\left(\tilde{\Lambda}'\Omega^{-1}\tilde{\Lambda} + \mathbf{I}_r\right)^{-1}\boldsymbol{\lambda}_1 + O\left(\frac{1}{T}\right). \end{aligned} \tag{5.5}$$

However, under Assumptions 1-3, the eigenvalues of both $\tilde{\Lambda}'\Omega^{-1}\tilde{\Lambda} + \mathbf{I}_r$ and $\tilde{\Lambda}'\tilde{\Lambda}$ are $O\left(N\right)$, hence as $(N,T) \to \infty$, $\boldsymbol{\lambda}_1'\left(\tilde{\Lambda}'\tilde{\Lambda}\right)^{-1}\left(\tilde{\Lambda}'\Omega\tilde{\Lambda}\right)\left(\tilde{\Lambda}'\tilde{\Lambda}\right)^{-1}\boldsymbol{\lambda}_1 \to 0$ and $\left(\tilde{\Lambda}'\tilde{\Lambda}\right)^{-1} \to 0$. Thus,

$$Var\left(\hat{e}_{1t}\right) \to \sigma_1^2 \;\text{ and }\; Var\left(\tilde{e}_{1t}\right) \to \sigma_1^2. \tag{5.6}$$

That is, when $(N,T) \to \infty$, $\boldsymbol{\lambda}_i$ and $\mathbf{f}_{T+h}$ or $\mathbf{w}$ can be estimated precisely. There is no difference whether one predicts the outcomes by factor approach or linear projection method.

When $\mathbf{c} \neq \mathbf{0}$, we can similarly obtain that LP is as efficient as the FB. See Appendix for details.

5.2 Case 2: N fixed, $T \to \infty$

Proposition 5.2 *Under Assumptions 1-3, when N is fixed and $T \to \infty$, the LP approach yields smaller mean square prediction error than the FB approach.*

When N is fixed and $T \to \infty$, $\hat{\mathbf{w}} \to_p \mathbf{w}$ for the LP projector (3.12), while the FB predictor (3.6) is based on the product of $\hat{\boldsymbol{\lambda}}_1'\hat{\mathbf{f}}_{T+h}$. Although $\hat{\boldsymbol{\lambda}}_1$ can be consistently estimated by (3.3) (Bai (2003)), $\hat{\mathbf{f}}_{T+h}$ is estimated with finite number of cross-sectional observations $\mathbf{y}_{T+h}$, i.e.,

$$\hat{\mathbf{f}}_{T+h} = \left(\widehat{\tilde{\Lambda}}\widehat{\tilde{\Lambda}}'\right)^{-1}\widehat{\tilde{\Lambda}}'\mathbf{y}_{T+h}, \tag{5.7}$$

which although unbiased, will have variance of order $\frac{1}{N}$. Thus, the asymptotic variance of $\tilde{e}_{1,T+h} = u_{1,T+h} + \boldsymbol{\lambda}_1'\left(\mathbf{f}_{T+h} - \hat{\mathbf{f}}_{T+h}\right)$ takes the form[8]

$$Var\left(\tilde{e}_{1,T+h}\right) = \sigma_1^2 + \boldsymbol{\lambda}_1'\left(\tilde{\Lambda}'\tilde{\Lambda}\right)^{-1}\left(\tilde{\Lambda}'\Omega\tilde{\Lambda}\right)\left(\tilde{\Lambda}'\tilde{\Lambda}\right)^{-1}\boldsymbol{\lambda}_1. \tag{5.8}$$

On the other hand, the LP prediction error variance for (3.12) under the assumption that u_{1t} is uncorrelated with u_{it} for $i \geq 2$ (i.e., $\mathbf{c} = \mathbf{0}$ for (2.3)) takes the form (e.g., (2.7) with $\mathbf{c} = \mathbf{0}$)

$$\sigma_\eta^2 = \sigma_1^2 + \boldsymbol{\lambda}_1'\left(\tilde{\Lambda}'\Omega^{-1}\tilde{\Lambda} + \Sigma_f^{-1}\right)^{-1}\boldsymbol{\lambda}_1. \tag{5.9}$$

[8] There is a one to one correspondence between the estimated $\hat{\Lambda}$ and $\hat{\mathbf{F}}$ under (3.1) or (3.3). Let $\hat{\Lambda}^*$ and $\hat{\mathbf{F}}^*$ denote the estimates under the normalization $\frac{1}{T}\hat{\mathbf{F}}'\hat{\mathbf{F}} = \mathbf{I}_r$, and $\hat{\Lambda}$ and $\hat{\mathbf{F}}$ denote the estimates under the normalization $\frac{1}{N}\hat{\Lambda}'\hat{\Lambda} = \mathbf{I}_r$, then $\hat{\mathbf{F}}^* = \hat{\mathbf{F}}\Sigma_f^{1/2}$ and $\hat{\Lambda}^* = \hat{\Lambda}\Sigma_f^{-1/2}$.

We note that $\left(\tilde{\Lambda}'\tilde{\Lambda}\right)^{-1}\left(\tilde{\Lambda}'\Omega\tilde{\Lambda}\right)\left(\tilde{\Lambda}'\tilde{\Lambda}\right)^{-1} \geq \left(\tilde{\Lambda}'\Omega\tilde{\Lambda}\right)^{-1} \geq \left(\tilde{\Lambda}'\Omega\tilde{\Lambda}+\mathbf{I}_r\right)^{-1}$ because $\left(\tilde{\Lambda}'\tilde{\Lambda}\right)^{-1}\left(\tilde{\Lambda}'\Omega\tilde{\Lambda}\right)\left(\tilde{\Lambda}'\tilde{\Lambda}\right)^{-1}$ and $\left(\tilde{\Lambda}'\Omega\tilde{\Lambda}\right)^{-1}$ are the covariance matrix of the least squares and the generalized least squares estimator of $\mathbf{f}_t$ of

$$\tilde{\mathbf{y}}_t = \tilde{\Lambda}\mathbf{f}_t + \tilde{\mathbf{u}}_t,$$

and Σ_f is positive definite.

Similar results hold when $\mathbf{c} \neq \mathbf{0}$. For details, see Appendix for a discussion.

Remark 5.1 *Our analysis for the case when N is fixed and $T \to \infty$ is based on the MLE estimates of $\mathbf{f}_t$ and $\mathbf{w}$. A referee has suggested to replace the estimates for $\hat{\mathbf{f}}_{T+h}$ (5.7) by the James-Stein estimator (1961) estimator, $\hat{\mathbf{f}}_{T+h,JS}$. The James-Stein estimator brings concentrated attention on the total square error loss function. It is particularly relevant when there are multiple treated units. However, the James-Stein estimator is a biased estimator, which leads to a biased predictor of $y_{1,T+h}$. Moreover, our simulation results have shown that the risk is not independent of the dimension of regressors. We hope to investigate in using the Stein or empirical Bayes estimators in lieu of the MLE in the future.*

5.3 Case 3: T fixed, $N \to \infty$

When T is fixed and N is large, we use the normalization $\Sigma_f = \mathbf{I}_r$ and obtain the estimated $\mathbf{F}$, $\hat{\mathbf{F}}^*$, as $\sqrt{T}$ times the eigenvectors corresponding to the r largest eigenvalues of (3.1). Conditional on $\hat{\mathbf{F}}^*$, we can estimate $\boldsymbol{\lambda}_i^*$ by

$$\hat{\boldsymbol{\lambda}}_i^* = \frac{1}{T}\hat{\mathbf{F}}^{*\prime}\mathbf{y}_i, \quad i = 1, \ldots, N. \tag{5.10}$$

The FB predictor now takes the form

$$\tilde{y}_{1,T+h} = \hat{\boldsymbol{\lambda}}_1^{*\prime}\hat{\mathbf{f}}_{T+h}^*. \tag{5.11}$$

It is not feasible to directly implement the LP approach when $N > T$ because $\frac{1}{T}\sum_{t=1}^{T}\tilde{\mathbf{y}}_t\tilde{\mathbf{y}}_t'$ is a singular matrix. We suggest the following two approaches to implement the LP approach.

The first approach is to use a subset of $(N-1)$ observed $\tilde{\mathbf{y}}_t$ to generate predicted value of y_{1t}, $\hat{y}_{1,t}$. Hsiao et al (2012) suggest to use a model selection criterion, Li and Bell (2017) suggest to use Lasso (e.g., Tibshirani (1996)), and Chen (2023) suggest to use online convex optimization to select a subset of $(N-1)$ y_{it} as control units, say $\tilde{\mathbf{y}}_t^*$, where the dimension of $\tilde{\mathbf{y}}_t^*$ is less than T. When T is greater than the dimension of $\tilde{\mathbf{y}}_t^*$, the LP predictor,

$$\hat{y}_{1,T+h} = \hat{\mathbf{w}}^{*\prime}\tilde{\mathbf{y}}_{T+h}^*, \tag{5.12}$$

is feasible. The predictor is asymptotically unbiased with prediction error variance (3.14).

The second approach is a prediction averaging approach. We suggest to randomly break up the $(N-1)$ control units into G subgroups, each consists of N_g cross-sectional units subject to N_g less than T, then use LP to generate predicted value of $y_{1,T+h}$ by $\hat{y}_{1,T+h}^g = \hat{\mathbf{w}}_g'\tilde{\mathbf{y}}_{g,T+h}$ for each subgroup $g = 1, \ldots, G$, where

$$\hat{\mathbf{w}}_g = \left(\sum_{t=1}^T \tilde{\mathbf{y}}_{gt}\tilde{\mathbf{y}}_{gt}'\right)^{-1} \sum_{t=1}^T \tilde{\mathbf{y}}_{gt} y_{1t}. \tag{5.13}$$

where $\tilde{\mathbf{y}}_{gt}$ is a $N_g \times 1$ vector consists of N_g cross-sectional units that belongs to the g-th subgroup, i.e., $\tilde{\mathbf{y}}_{gt} = \left(1_{(i\in g)} y_{it}\right)$, $g = 1, \ldots, G$. Then generate the predicted value of $y_{1,T+h}$ by

$$\widehat{\hat{y}}_{1,T+h} = \frac{1}{G}\sum_{g=1}^G \hat{y}_{1,T+h}^g. \tag{5.14}$$

Proposition 5.3 *Under Assumptions 1-3, when T is fixed and $N \to \infty$, predictors (5.11), (5.12) and (5.14) are all asymptotically unbiased. However*

(i) The predictor (5.11) is likely to have smaller prediction error variance than the predictor (5.12).

(ii) The predictor (5.14) is likely to have the smallest prediction error variance than the predictor (5.11) and (5.12) if $\boldsymbol{\lambda}_i$ is independently distributed over i and $\mathbf{u}_i$ is independent over i with $E(\mathbf{u}_i\mathbf{u}_i') = \breve{\Omega}$.

We note that since $\hat{\mathbf{f}}_t^* \to_p \mathbf{f}_t^*$ as $N \to \infty$ and $\hat{\boldsymbol{\lambda}}_1^*$ in (5.10) is unbiased, so is (5.11). The covariance matrix of $\hat{\boldsymbol{\lambda}}_1^*$ under the normalization condition $\Sigma_{f^*} = \mathbf{I}_r$ is

$$Var(\hat{\boldsymbol{\lambda}}_i^*) = \sigma_i^2 \left(\hat{\mathbf{F}}^{*\prime}\hat{\mathbf{F}}^*\right)^{-1} = \frac{1}{T}\sigma_i^2\mathbf{I}_r. \tag{5.15}$$

Then the variance of the prediction error for FB approach $\tilde{e}_{1,T+h} = u_{1,T+h} + \mathbf{f}_{T+h}^{*\prime}\left(\boldsymbol{\lambda}_1^* - \hat{\boldsymbol{\lambda}}_1^*\right)$ takes the form

$$Var\left(\tilde{e}_{1,T+h}\right) = \sigma_1^2\left(1 + \frac{1}{T}\mathbf{f}_{T+h}^{*\prime}\mathbf{f}_{T+h}^*\right). \tag{5.16}$$

The prediction error of (5.12) takes the form

$$\widehat{e}_{1,T+h}^* = \eta_{1,T+h}^* + \left(\mathbf{w}^* - \hat{\mathbf{w}}^*\right)'\tilde{\mathbf{y}}_{T+h}^*, \tag{5.17}$$

with

$$Var\left(\widehat{e}^{*}_{1,T+h}\right)=\sigma^2_{\eta^*}\left(1+\tilde{\mathbf{y}}^{*\prime}_{T+h}Var\left(\hat{\mathbf{w}}^{*}\right)\tilde{\mathbf{y}}^{*}_{T+h}\right), \tag{5.18}$$

where * denotes the sample selected by Lasso.

Under the assumption that $\boldsymbol{\lambda}_i$ are independently distributed over i with $E(\boldsymbol{\lambda}_i)=0$,[9] the prediction error of $\widehat{\widehat{y}}_{1,T+h}$ in (5.14) takes the form

$$\begin{aligned}
\widehat{e}_{1,T+h} &= y_{1,T+h}-\frac{1}{G}\sum_{g=1}^{G}\hat{y}^{g}_{1,T+h} \\
&= u_{1,T+h}+\boldsymbol{\lambda}_1'\mathbf{f}_{T+h}-\frac{1}{G}\sum_{g=1}^{G}\mathbf{w}_g'\tilde{\mathbf{y}}_{g,T+h}+\frac{1}{G}\sum_{g=1}^{G}\left(\mathbf{w}_g-\hat{\mathbf{w}}_g\right)'\tilde{\mathbf{y}}_{g,T+h} \\
&= u_{1,T+h}+\frac{1}{G}\sum_{g=1}^{G}\left\{\boldsymbol{\lambda}_1'\left[\mathbf{I}_r-\left(\Sigma_f\tilde{\Lambda}_g'+\mathbf{c}_g'\right)\left(\tilde{\Lambda}_g\Sigma_f\tilde{\Lambda}_g'+\Omega_g\right)^{-1}\tilde{\Lambda}_g\Sigma_f\right]\mathbf{f}_{T+h}\right. \\
&\quad \left.-\left(\boldsymbol{\lambda}_1'\Sigma_f\tilde{\Lambda}_g'+\mathbf{c}_g'\right)\left(\tilde{\Lambda}_g\Sigma_f\tilde{\Lambda}_g'+\Omega_g\right)^{-1}\tilde{\mathbf{u}}_{g,T+h}\right\}+\frac{1}{G}\sum_{g=1}^{G}\left(\mathbf{w}_g-\hat{\mathbf{w}}_g\right)'\tilde{\mathbf{y}}_{g,T+h}
\end{aligned} \tag{5.19}$$

under the normalization $\Sigma_f=\mathbf{I}_r$, where, for $g=1,\ldots,G$, $\tilde{\Lambda}_g$ denotes the $N_g\times r$ submatrix of the $(N-1)\times r$ matrix $\tilde{\Lambda}$ that associated with $1_{(i\in g)}y_{it}$, $\mathbf{c}_g=E(\tilde{\mathbf{u}}_{g,t}u_{1t})$ and $\Omega_g=E\left(\tilde{\mathbf{u}}_{g,t}\tilde{\mathbf{u}}_{g,t}'\right)$ with $\tilde{\mathbf{u}}_{g,t}$ consists of the N_g elements of $1_{(i\in g)}u_{it}$.

Since y_{it}, $\boldsymbol{\lambda}_i$ and $\mathbf{f}_t$ are stationary under Assumption 1, 2 and 4, when $N\to\infty$, so is G with fixed $N_g<T$, it can be shown that

$$\begin{aligned}
Var\left(\widehat{e}_{1,T+h}\right) &\to \sigma_1^2+\frac{1}{G}E\left(\eta^2_{g,T+h}+\sigma^2_{\eta g}\tilde{\mathbf{y}}_{g,T+h}'\left(\frac{1}{T}\sum_{t=1}^{T}\tilde{\mathbf{y}}_{gt}\tilde{\mathbf{y}}_{gt}'\right)^{-1}\tilde{\mathbf{y}}_{g,T+h}\right) \\
&\to \sigma_1^2,
\end{aligned} \tag{5.20}$$

where $E\left(\eta^2_{g,T+h}\right)=\sigma^2_{\eta g}=\sigma_1^2+\boldsymbol{\lambda}_1'\Sigma_f\boldsymbol{\lambda}_1-\left(\boldsymbol{\lambda}_1'\Sigma_f\tilde{\Lambda}_g'+\mathbf{c}_g'\right)\left(\tilde{\Lambda}_g\Sigma_f\tilde{\Lambda}_g'+\Omega_g\right)^{-1}\left(\tilde{\Lambda}_g\Sigma_f\boldsymbol{\lambda}_1+\mathbf{c}_g\right)$.

In other words, when $N\to\infty$ and T is fixed, the average method of LP approach can yield smaller asymptotic variance than the FB approach.

5.4 Case 4: Both N and T are Finite

Proposition 5.4 *Conditional on $\mathbf{Y}_T$ and $\tilde{\mathbf{y}}_{T+h}$, the factor predictor is biased. The bias depends on the realized value of $\boldsymbol{\lambda}_1$ and $\mathbf{f}_{T+h}$ and is of order $O\left(\frac{1}{N}\right)+O\left(\frac{1}{T}\right)$ (e.g., Bai (2003)). However, the LP predictor remains unbiased. Although prediction error distributions of LP and*

[9] If $E(\boldsymbol{\lambda}_i)=\boldsymbol{\lambda}\neq 0$, mode (2.1) can be written as $(y_{it}-\bar{y}_t)=(\boldsymbol{\lambda}_i-\boldsymbol{\lambda})'\mathbf{f}_t+(u_{it}-\bar{u}_t)$, where $\bar{y}_t=\frac{1}{N}\sum_{i=1}^{N}y_{it}$ and $\bar{u}_t=\frac{1}{N}\sum_{i=1}^{N}u_{it}$. The term $\bar{u}_t$ will converge to zero as $N\to\infty$ under Assumption 3.

FB predictors with finite N and T depend on the realized $\tilde{\mathbf{y}}_{T+h}$ and $\mathbf{f}_{T+h}$, from the expected mean square prediction error perspective, the LP approach is likely to yield smaller prediction error variance.

When both N and T are finite, the LP predictor $\hat{y}_{1,T+h} = \hat{\mathbf{w}}'\tilde{\mathbf{y}}_{T+h}$ remains unbiased because $E(\hat{\mathbf{w}}) = \mathbf{w}$. Its prediction error variance is equal to (3.14). On the other hand, the FB predictor $\hat{\boldsymbol{\lambda}}_1'\hat{\mathbf{f}}_{T+h}$ is a biased predictor when N and T are finite (Bai (2003)). The bias is of order $O\left(\frac{1}{T}\right) + O\left(\frac{1}{N}\right)$. The FB prediction error

$$\begin{aligned} \tilde{e}_{1,T+h} &= u_{1,T+h} + \boldsymbol{\lambda}_1'\mathbf{f}_{T+h} - \hat{\boldsymbol{\lambda}}_1'\hat{\mathbf{f}}_{T+h} \\ &= u_{1,T+h} + \boldsymbol{\lambda}_1'\left(\mathbf{f}_{T+h} - \hat{\mathbf{f}}_{T+h}\right) + \mathbf{f}_{T+h}'\left(\boldsymbol{\lambda}_1 - \hat{\boldsymbol{\lambda}}_1\right) - \left(\boldsymbol{\lambda}_1 - \hat{\boldsymbol{\lambda}}_1\right)'\left(\mathbf{f}_{T+h} - \hat{\mathbf{f}}_{T+h}\right), \end{aligned} \tag{5.21}$$

where $\boldsymbol{\lambda}_1$ is the first row of $\hat{\Lambda}$ obtained as $\sqrt{N}$ times the r eigenvectors corresponding to the r largest eigenvalues of $\frac{1}{T}\sum_{t=1}^{T}\mathbf{y}_t\mathbf{y}_t'$, arranged in decreasing order, and

$$\hat{\mathbf{f}}_{T+h} = \left(\hat{\tilde{\Lambda}}'\hat{\tilde{\Lambda}}\right)^{-1}\hat{\tilde{\Lambda}}'\tilde{\mathbf{y}}_{T+h}, \tag{5.22}$$

where $\hat{\tilde{\Lambda}}$ denotes the rows of $\hat{\Lambda}$ from the second to the N-th.

It is shown in the Appendix that the variance of FB prediction error is

$$Var\left(\tilde{e}_{1,T+h}\right) = \sigma_1^2 + \boldsymbol{\lambda}_1' Var\left(\hat{\mathbf{f}}_{T+h}\right)\boldsymbol{\lambda}_1 + \frac{\sigma_1^2}{T}\mathbf{f}_{T+h}'\Sigma_f^{-1}\mathbf{f}_{T+h} + O\left(\frac{1}{NT}\right). \tag{5.23}$$

Subsection 5.2 shows that $\sigma_\eta^2 \leq \sigma_1^2 + \boldsymbol{\lambda}_1' Var\left(\hat{\mathbf{f}}_{T+h}\right)\boldsymbol{\lambda}_1$. Whether $\frac{\sigma_1^2}{T}\mathbf{f}_{T+h}'\Sigma_f^{-1}\mathbf{f}_{T+h}$ is greater or smaller than $\sigma_\eta^2\tilde{\mathbf{y}}_{T+h}'\left(\sum_{t=1}^{T}\tilde{\mathbf{y}}_t\tilde{\mathbf{y}}_t'\right)^{-1}\tilde{\mathbf{y}}_{T+h}$ depends on $\tilde{\mathbf{y}}_{T+h}$ or $\mathbf{f}_{T+h}$. However, if $\tilde{\mathbf{y}}_{T+h}\tilde{\mathbf{y}}_{T+h}'$ is close to $E\left(\tilde{\mathbf{y}}_{T+h}\tilde{\mathbf{y}}_{T+h}'\right) = E\left(\frac{1}{T}\sum_{t=1}^{T}\tilde{\mathbf{y}}_t\tilde{\mathbf{y}}_t'\right) = \left(\tilde{\Lambda}\Sigma_f\tilde{\Lambda}' + \Omega\right)$, then[10]

$$\sigma_\eta^2\tilde{\mathbf{y}}_{T+h}'\left(\sum_{t=1}^{T}\tilde{\mathbf{y}}_t\tilde{\mathbf{y}}_t'\right)^{-1}\tilde{\mathbf{y}}_{T+h} \simeq \sigma_\eta^2\left(\frac{N}{T}\right) = \sigma_1^2\left(\frac{N}{T}\right) + \boldsymbol{\lambda}_1'\left(\tilde{\Lambda}'\Omega\tilde{\Lambda} + \mathbf{I}_r\right)^{-1}\boldsymbol{\lambda}_1\left(\frac{N}{T}\right), \tag{5.24}$$

under the assumption that $T > N$.

Similarly, if $\mathbf{f}_{T+h}\mathbf{f}_{T+h}' \simeq \Sigma_f$, then $\frac{\sigma_1^2}{T}\mathbf{f}_{T+h}'\Sigma_f^{-1}\mathbf{f}_{T+h} \simeq \sigma_1^2\left(\frac{r}{T}\right)$. Then

$$\begin{aligned} & Var\left(\tilde{e}_{1,T+h}\right) - Var\left(\hat{e}_{1,T+h}\right) \\ = \; & \sigma_1^2\left(1 + \frac{r}{T}\right) + \boldsymbol{\lambda}_1'\left(\tilde{\Lambda}'\tilde{\Lambda}\right)^{-1}\left(\tilde{\Lambda}'\Omega\tilde{\Lambda}\right)\left(\tilde{\Lambda}'\tilde{\Lambda}\right)^{-1}\boldsymbol{\lambda}_1 - \left[\sigma_1^2 + \boldsymbol{\lambda}_1'\left(\tilde{\Lambda}'\Omega\tilde{\Lambda} + \mathbf{I}_r\right)^{-1}\boldsymbol{\lambda}_1\right]\left(1 + \frac{N}{T}\right) \\ = \; & \sigma_1^2\left(\frac{r-N}{T}\right) + \boldsymbol{\lambda}_1'\left[\left(\tilde{\Lambda}'\tilde{\Lambda}\right)^{-1}\left(\tilde{\Lambda}'\Omega\tilde{\Lambda}\right)\left(\tilde{\Lambda}'\tilde{\Lambda}\right)^{-1} - \boldsymbol{\lambda}_1'\left(\tilde{\Lambda}'\Omega\tilde{\Lambda} + \mathbf{I}_r\right)^{-1}\right]\boldsymbol{\lambda}_1 \\ & -\frac{N}{T}\boldsymbol{\lambda}_1'\left(\tilde{\Lambda}'\Omega\tilde{\Lambda} + \mathbf{I}_r\right)^{-1}\boldsymbol{\lambda}_1, \end{aligned} \tag{5.25}$$

[10]The linear projection coefficient are invariant to the normalization we set up for Σ_λ or Σ_f. Here, for notational ease, we use $\Sigma_f = \mathbf{I}_r$.

where the first term is negative under the assumption that $N > r$, the sign of the second term is positive as shown in subsection 5.2 and the sign of the last term is negative. Thus, the sign of (5.25) cannot be determined a prior given that Λ are unknown. Either way, the difference is likely to be order $O\left(\frac{1}{T}\right)$. Since the bias is of order $O\left(\frac{1}{T}\right) + O\left(\frac{1}{N}\right)$. It is likely that with finite N and T, the LP method could perform equally well or better than the FB method.

Remark 5.2 *Although the analytic results discussed in this section are derived under Assumptions 1-3. Similar results hold when u_{it} are both weakly cross-correlated and weakly time dependent as defined in Bai (2009, Assumption C) or Jiang et. al (2021, Assumption 2) because if u_{it} are weakly cross-correlated and weakly time-dependent, it implies that only a finite number of cross-sectional units, u_{jt} are significantly correlated with u_{it} over i and only a finite number of time series observations $u_{i,t\pm s}$ are significantly correlated with u_{it} over t. They do not change the probability limit over i and t when divided by NT.*

6 Generalization to Factor Augmented Linear Regression Models

Consider

$$
\begin{aligned}
y_{it} &= \mathbf{x}_{it}'\boldsymbol{\beta} + \boldsymbol{\lambda}_i'\mathbf{f}_t + v_{it}, && (6.1)\\
v_{it} &= \boldsymbol{\lambda}_i'\mathbf{f}_t + u_{it}, \quad i = 1, \ldots, T; i = 1, \ldots, N; && (6.2)
\end{aligned}
$$

where u_{it} satisfies Assumption 1, 2, and 3 or 4, and $\mathbf{x}_{it}$ are independent of u_{it} (e.g., Bai (2009), Hsiao and Zhou (2019), Pesaran (2006), Xu (2017)).[11] We consider a predictive model of the form

$$\hat{y}_{it} = \mathbf{x}_{it}'\hat{\boldsymbol{\beta}} + \hat{v}_{it}. \tag{6.3}$$

The error of predicting y_{it} then takes the form

$$e_{it} = y_{it} - \hat{y}_{it} = \mathbf{x}_{it}'\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right) + \left(v_{it} - \hat{v}_{it}\right), \tag{6.4}$$

i.e., the error consists of two parts, the part due to the error of estimating $\boldsymbol{\beta}$ and the part due to $\left(v_{it} - \hat{v}_{it}\right)$. If the same estimation method is applied to estimate $\boldsymbol{\beta}$, the part due to $\mathbf{x}_{it}'\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)$ is identical in (6.4) no matter which method is utilized to construct $\hat{v}_{it}$. Thus, the analysis of the relative merits between FB and LP methods continue to hold for models of (6.2) and (6.3).

[11] One may view model (6.1) and (6.2) as a predictive model of y_{it} observing everything else where measurement of treatment effcts is just a special case.

To see this, let $\mathbf{X}_t = (\mathbf{x}_{1t}, \ldots, \mathbf{x}_{Nt})'$, $\mathbf{X}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})'$, $\tilde{\mathbf{X}} = (\mathbf{X}_1, \ldots, \mathbf{X}_T)$, and $\mathbf{X}^* = (\mathbf{X}_1, \ldots, \mathbf{X}_N)$. Let

$$\mathbf{M}_X = \mathbf{I}_T - \tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-}\tilde{\mathbf{X}}', \tag{6.5}$$

and

$$\mathbf{M}^* = \mathbf{I}_N - \mathbf{X}^*(\mathbf{X}^{*'}\mathbf{X}^*)^{-}\mathbf{X}^{*\prime}, \tag{6.6}$$

where $\mathbf{A}^-$ denotes the generalized inversion of a square matrix $\mathbf{A}$. By construction, we have

$$\mathbf{M}_X\mathbf{X}_t = \mathbf{0}, \qquad \text{for } t = 1, \ldots, T, \tag{6.7}$$

and

$$\mathbf{M}^*\mathbf{X}_i = \mathbf{0}, \qquad \text{for } i = 1, \ldots, N. \tag{6.8}$$

Consequently, multiplying $\mathbf{M}_X$ to $\mathbf{y}_t$ equations

$$\mathbf{y}_t = \mathbf{X}_t\boldsymbol{\beta} + \Lambda\mathbf{f}_t + \mathbf{u}_t, \quad t = 1, \ldots, T, \tag{6.9}$$

yields

$$\mathbf{y}_t^* = \Lambda^*\mathbf{f}_t + \mathbf{u}_t^*, \qquad t = 1, \ldots, T, \tag{6.10}$$

where $\mathbf{y}_t$, $\mathbf{X}_t$ and $\mathbf{u}_t$ are defined as before, and $\mathbf{y}_t^* = \mathbf{M}_X\mathbf{y}_t$, $\Lambda^* = \mathbf{M}_X\Lambda$, $\mathbf{u}_t^* = \mathbf{M}_X\mathbf{u}_t$.

Similarly, multiplying $\mathbf{M}^*$ to $\mathbf{y}_i$ equations

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{F}\boldsymbol{\lambda}_i + \mathbf{u}_i, \qquad i = 1, \ldots, N, \tag{6.11}$$

yields

$$\mathbf{y}_i^* = \mathbf{F}^*\boldsymbol{\lambda}_i + \mathbf{u}_i^*,$$

where $\mathbf{y}_i^* = \mathbf{M}^*\mathbf{y}_i$, $\mathbf{F}^* = \mathbf{M}^*\mathbf{F}$ and $\mathbf{u}_i^* = \mathbf{M}^*\mathbf{u}_i$.[12]

Both $\mathbf{y}_t^*$ or $\mathbf{y}_i^*$ are now in the factor model form. Hence the analysis of predictor error interval for[13]

$$\hat{y}_{1t}^* = \hat{\mathbf{w}}^{*\prime}\tilde{\mathbf{y}}_t^*, \tag{6.12}$$

or[14]

$$\tilde{y}_{1t}^* = \hat{\boldsymbol{\lambda}}_1^*\hat{\mathbf{f}}_t, \tag{6.13}$$

[12]There exists possibility that $\mathbf{M}^*$ could be a null matrix (e.g., Hsiao et al (2021)). In this situation, one could either use Bai's (2009) least squares method to simultaneously estimate $\boldsymbol{\beta}$, Λ, and $\mathbf{F}$ to construct a FB predictor or use Hsiao and Zhou's (2019) method to construct the LP prediction.

[13]Predictor (6.12) is in the form of the semiparametric estimator of Hsiao and Zhou (2019).

[14]One could also consider the factor based approach of predicting y_{1t} using the estimation of $\boldsymbol{\beta}, \boldsymbol{\lambda}_1$ and $\mathbf{f}_t$ as in Bai and Ng (2021), Xu (2017), etc.,.

for the cases of $(N, T) \to \infty$, N fixed, $T \to \infty$, T fixed and $N \to \infty$ or both N and T are identical to Case (1)-(4).

Bai (2009), Pesaran (2006), Hsiao, et al (2022), etc., have suggested different methods to estimate $\boldsymbol{\beta}$. Whatever method is adopted, the prediction error variance of

$$\hat{y}_{1t} = \mathbf{x}_{1t}'\hat{\boldsymbol{\beta}} + \hat{\mathbf{w}}^{*\prime}\left(\tilde{\mathbf{y}}_t - \tilde{\mathbf{X}}_t\hat{\boldsymbol{\beta}}\right), \tag{6.14}$$

or

$$\tilde{y}_{1t} = \mathbf{x}_{1t}'\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\lambda}}_1^*\hat{\mathbf{f}}_t, \tag{6.15}$$

is equal to the prediction error variance of (6.12) or (6.13) plus $\mathbf{x}_{1t}'Var\left(\hat{\boldsymbol{\beta}}\right)\mathbf{x}_{1t}$. Hence similar results to those of Case (1)-(4) can be drawn.

7 Monte Carlo Simulations

We use Monte Carlo to investigate if our analytic results with regard to FB vs LP discussed in Sections 5 and 6 continue to hold with finite N and T. We set the number of factors to $r = [N^{1/3}]$ and the individual specific effects, α_i, are $iid\ Uniform(0, 2)$. We consider the following DGPs that includes

DGP1: Pure factor model

$$y_{it} = \alpha_i + \sum_{j=1}^{r} \lambda_{j,i} f_{jt} + u_{it}, \tag{7.1}$$

where the factor loadings $\lambda_{1,i}, \ldots, \lambda_{r,i}$ are $iid\ N(0, 1)$.

DGP2. Model with exogenous variables and unobserved factors

$$y_{it} = \alpha_i + x_{1,it}\beta_1 + x_{2,it}\beta_2 + \sum_{j=1}^{r} \lambda_{j,i} f_{jt} + u_{it}. \tag{7.2}$$

The covariates $x_{k,it}$ $(k = 1, 2)$ are (positively) correlated with the factors and extra factors as follows

$$x_{k,it} = 1 + \rho_{ki} x_{k,it-1} + \sum_{j=1}^{r} c_j f_{jt} + \eta_{k,it}, \quad k = 1, 2,$$

where $\rho_{k,i} \sim iid\ U(0.1, 0.9)$, $c_j \sim iid\ U(1, 2)$ and the error term $\eta_{k,it}$ is $iid\ \left(\chi^2(1) - 1\right)$. We let $\beta_1 = 1$ and $\beta_2 = 2$.

For these two DGPs, we consider five different generations of the errors:

Case 1: Factors $f_{1t}, \ldots, f_{rt}$ are $\chi^2(1)$ and u_{it} are non-normally distributed errors with

$$u_{it} \sim iid\ \left(\chi^2(1) - 1\right). \tag{7.3}$$

Case 2: Factors $f_{1t}, \ldots, f_{rt}$ are $\chi^2(1)$ and u_{it} are serially-correlated

$$u_{it} = \rho_{u,i} u_{it-1} + v_{it}, \tag{7.4}$$

where the idiosyncratic errors v_{it} are $IID\mathcal{N}\left(0, \sigma_{v,i}^2\right)$ with $\sigma_{v,i}^2$ being random draws from $\left(1 + 0.5\chi^2(2)\right)$ and $\rho_{u,i}$ are i.i.d draws from $U(0.2, 0.8)$ for $i = 1, \ldots, N$.

Case 3: Factors $f_{1t}, \ldots, f_{rt}$ are $\chi^2(1)$ and u_{it} are weakly cross-correlated

$$u_{it} = \epsilon_{it} + 0.3\epsilon_{i+1,t} + 0.3\epsilon_{i-1,t}, \tag{7.5}$$

where ϵ_{it} is $iid\ N(0, \sigma_i^2)$ with $\sigma_i^2 \sim (iid\ \chi^2(1) + 1)/2$ for $i = 1, \ldots, N + 1$.

Case 4: Factors $f_{1t}, \ldots, f_{rt}$ are $\chi^2(1)$ and u_{it} are both weakly cross-correlated and serially correlated

$$u_{it} = \epsilon_{it} + 0.3\epsilon_{i+1,t} + 0.3\epsilon_{i-1,t}, \tag{7.6}$$

where ϵ_{it} is serially correlated as

$$\epsilon_{it} = \rho_{\epsilon,i}\epsilon_{it-1} + v_{it}, \tag{7.7}$$

where the idiosyncratic errors v_{it} are $IID\mathcal{N}\left(0, \sigma_{v,i}^2\right)$ with $\sigma_{v,i}^2$ being random draws from $\left(1 + 0.5\chi^2(2)\right)$ and $\rho_{\epsilon,i}$ are i.i.d draws from $U(0.2, 0.8)$ for $i = 1, \ldots, N$. for $i = 1, \ldots, N + 1$.

Case 5: Serially-correlated factor: For DGP (7.1) and (7.2), we assume the factors are generated by

$$f_{jt} = \rho_{f,j} f_{j,t-1} + \epsilon_t, \quad j = 1, \ldots, r, \tag{7.8}$$

where $\rho_{f,j} \sim iid\ U(0.2, 0.8)$ and ϵ_t is $iid\ N(0, 1)$. The errors are generated by (7.3).

The treatment and control groups consist of 1 and $N - 1$ units, respectively. The treatment starts to affect the treated units at time $T + 1$. For these designs, we assume that the control units to be $N - 1 = 10, 30, 50, 100$ and the pretreatment time $T = 10, 30, 60$, and post-treatment periods $m = 5$. The number of replication is set at $R = 1000$.

To compare the accuracy of LP vs FB in estimating the treatment effects,[15] we consider the LP (3.12) (labelled as LP[16] and the associated prediction averaging (5.14), labelled as LP_ave[17]), and the FB (3.6) (labelled as FB) as well as the Box-Jenkins approach to generate univariate model (labelled as B-J)[18].

[15]For model with $\mathbf{x}_{it}$, we only consider the semiparametric approach as in Hsiao and Zhou (2019), where we first use the Bai (2009) least squares method to obtain $\hat{\boldsymbol{\beta}}$ and use the residuals to estimate the part that contribute to the counterfactuals.

[16]In the implementation of the LP approach when N is large and T is finite, we use the LASSO to select the subset of control units $\tilde{\mathbf{y}}_t$, where the upper limit of control units selected is set to be $\frac{T}{2}$.

[17]We consider prediction averaging (5.14) to generate predicted $y_{1,T+h}$ only when $N > T$, and let $G = 20$ if $T = 10$ and $G = 5$ if $T = 30$ and 60.

[18]Time series modeling of y_{1t} using Box-Jenkins procedure assuming y_{1t} follows an AR(1) model $y_{1t} = \phi y_{1t-1} + v_t$ and $y_{1,T+h}$ is predicted as $\hat{y}_{1,T+h} = \hat{\phi}^h y_{1T}$, where $\hat{\phi}$ denotes the OLS estimator of ϕ.

We consider two criteria for comparison: MAB and MSE. The MAB is the mean of absolute bias for the true outcome and the counterfactuals at each post-treatment date point.[19] The MSE is the mean of sum of squared bias for the true observation and the counterfactuals at each post-treatment date point.[20]

The simulation results for estimating treatment effects are summarized in Tables 1-10 for DGP1-2 with errors following (7.3)-(7.6) or serial correlated factors in (7.8) and for different combinations of N and T. The results confirm our analytic results that with finite N and T, the LP method yields more accurate predictions than the factor approach in terms of estimation bias and MSE, and both approaches are more accurate than using a univariate time series model to construct the counterfactuals, because they make use of the cross-sectional information to get more accurate predictions for post-treatment prediction of counterfactuals.

8 Empirical Application: German Reunification Revisited

In this section, we apply the LP and FB approaches to re-estimate the impact of the 1990 German reunification, one of the most significant political events in postwar Europe. The German Democratic Republic and the Federal Republic of Germany officially reunified on October 3, 1990. This event has been extensively investigated in the literature, to name a few, for instance, Abadie et. al (2015) and Ferman et al (2020) evaluate the impact of the German Reunification in 1991 on GDP per capita using the synthetic control approach.[21] Here we re-estimate the impact using the LP and FB approaches.

We use the Abadie et. al's (2015) data (which is an annual country-level panel data consisting 17 countries over the period 1960–2003) to compare the difference between LP and FB methods to predict the counterfactuals after German reunification in 1990. We transform the real GDP per capita into logarithmic figures. The German reunification occurred in 1990 (i.e., the intervention period), so the pre- and post-treatment periods are 1960 through 1990 and

[19] MAB is measured as

$$MAB = \frac{1}{Rm}\sum_{r=1}^{R}\sum_{t=T+1}^{T+m}\left|y_{1t}\left(r\right)-\hat{y}_{1t}\left(r\right)\right|,$$

with $y_{1t}(r)$ and $\hat{y}_{1t}(r)$ being the actual and counterfactual of y_{1t} in r-th replication. It represents the average distance between true treatment effect and estimated treatment effect by my method. Thus, the smaller the MAB is, the better performance the method is.

[20] MSE is calculated by

$$MSE = \frac{1}{Rm}\sum_{r=1}^{R}\sum_{t=T+1}^{T+m}\left(y_{1t}\left(r\right)-\hat{y}_{1t}\left(r\right)\right)^2$$

which is similar to MAB. The smaller it is, the better performance the method is.

[21] For a discussion of LP vs synthetic control method, see Wan et al (2018).

1991 through 2003, respectively. We follow the same strategy in the simulation section for implementation of the LP and FB approaches[22,23]. The estimation results are provided in Figure 1 and Figure 2 below. Figure 1 plots both the pre-treatment within sample and post treatment counterfactual paths of FB and LP approaches, and Figure 2 plots the estimated treatment effects based on the difference between the actual and FB or LP constructed counterfactuals together with the confidence intervals for the treatment effects estimated by both the LP and FB approaches.[24]

From these two graphs, we observe that: (1) The LP within sample fit dominates that of

[22]For the factor based approach, we use Bai and Ng's (2002) information criteria to select the number of factors and set the maximum number of unobserved factors as 5.

[23]Another specification is to consider a panel model with both covariates and interactive effects as in Abadie et al (2015), however, in the original dataset of Abadie et al (2015), there are lots of missing observations for these included variables (namely, trade openness, inflation rate, industry share, schooling levels, and investment rate). For illustration purpose, we just consider a pure factor specification for GDP per capita here.

[24]The confidence intervals of the treatment effects estimated by LP and FB are constructed as follows:

(1). 95% CI for LP approach:

$$\left[\hat{\Delta}_{1t} - 1.96 \times SE_{t,LP}, \hat{\Delta}_{1t} + 1.96 \times SE_{t,LP}\right],$$

for $t = 1, \ldots, m$, where

$$SE_{t,LP}^2 = \hat{\sigma}_\eta^2 + \tilde{\mathbf{y}}_t' \widehat{Var\left(\hat{\mathbf{w}}\right)} \tilde{\mathbf{y}}_t,$$

with

$$\begin{aligned} \hat{\sigma}_\eta^2 &= \frac{1}{T}\sum_{t=1}^{T}\left(y_{1t} - \hat{\mathbf{w}}'\tilde{\mathbf{y}}_t\right)^2, \\ \widehat{Var\left(\hat{\mathbf{w}}\right)} &= \hat{\sigma}_\eta^2 \left(\sum_{t=1}^{T} \tilde{\mathbf{y}}_t \tilde{\mathbf{y}}_t'\right)^{-1}. \end{aligned}$$

(2). 95% CI for FB approach:

$$\left[\hat{\Delta}_{1t} - 1.96 \times SE_{t,FB}, \hat{\Delta}_{1t} + 1.96 \times SE_{t,FB}\right],$$

for $t = 1, \ldots, m$, where

$$SE_{t,FB}^2 = \hat{\sigma}_1^2 + \hat{\boldsymbol{\lambda}}_1' \left(\widehat{\tilde{\Lambda}}'\widehat{\tilde{\Lambda}}\right)^{-1} \left(\widehat{\tilde{\Lambda}}'\hat{\Omega}\widehat{\tilde{\Lambda}}\right) \left(\widehat{\tilde{\Lambda}}'\widehat{\tilde{\Lambda}}\right)^{-1} \hat{\boldsymbol{\lambda}}_1 + \frac{1}{T}\hat{\sigma}_1^2 \hat{\mathbf{f}}_t'\hat{\mathbf{f}}_t,$$

where

$$\hat{\sigma}_1^2 = \frac{1}{T}\sum_{t=1}^{T}\left(y_{1t} - \hat{\boldsymbol{\lambda}}_1'\mathbf{f}_t\right)^2,$$

$\hat{\boldsymbol{\lambda}}_1'$ and $\widehat{\tilde{\Lambda}}$ are the PCA estimators for $\boldsymbol{\lambda}_1$ and $\tilde{\Lambda}$, respectively, and $\hat{\Omega}$ can be estimated by

$$\hat{\Omega} = \frac{1}{T}\sum_{t=1}^{T} \widehat{\tilde{\mathbf{u}}}_t \widehat{\tilde{\mathbf{u}}}_t' \text{ with } \widehat{\tilde{\mathbf{u}}}_t = \tilde{\mathbf{y}}_t - \widehat{\tilde{\Lambda}}\hat{\mathbf{f}}_t.$$

FB approach; (2) The FB estimates are very unstable due the instability of sample estimates of $\mathbf{f}_t$ with finite N; (3) The treatment effects estimated by LP approach are smooth and gently trending downward over time. The estimates are negative and statistically different from no treatment effects (=0) except for the first few years after reunification.[25] On the other hand, the estimated treatment effects using FB approach fluctuate widely over time. The absolute value of treatment effects estimated by FB approach are much larger than those obtained from LP approach. However, the estimated standard errors based on the FB approach is much larger than those based on LP approach, as a consequence, the confidence intervals for FB approach covers zero over time, while the intervals for LP approach do not cover zero, implying a small negative effects.

[25]The results based on synthetic control method are similar (Abadie et. al (2015, Fig. 2 and 3)). For a discussion of LP vs synthetic control method, see Wan et. al (2018).

Figure 1: Actual and predicted GDP per capita in both pre- and post-treatment periods

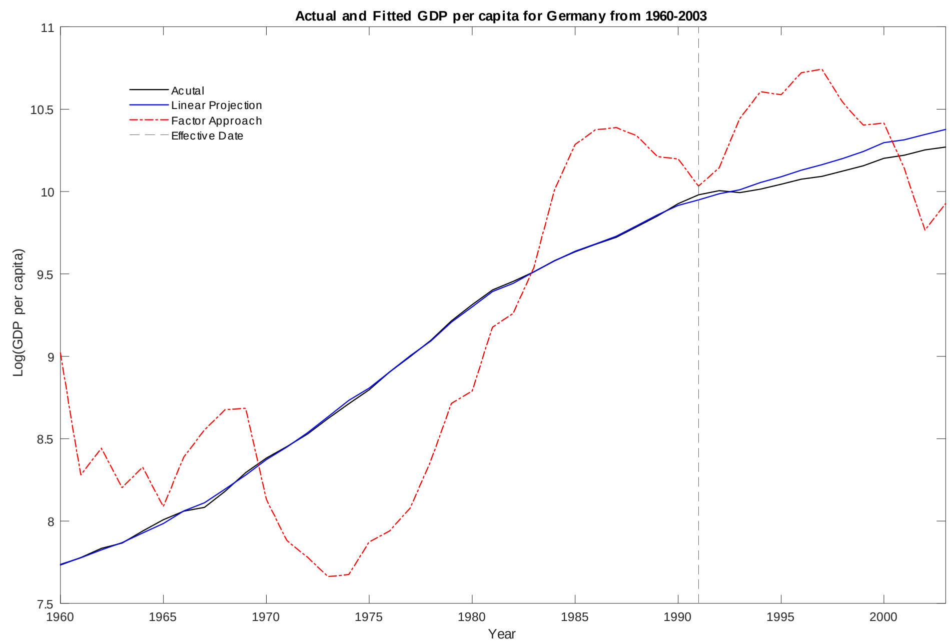
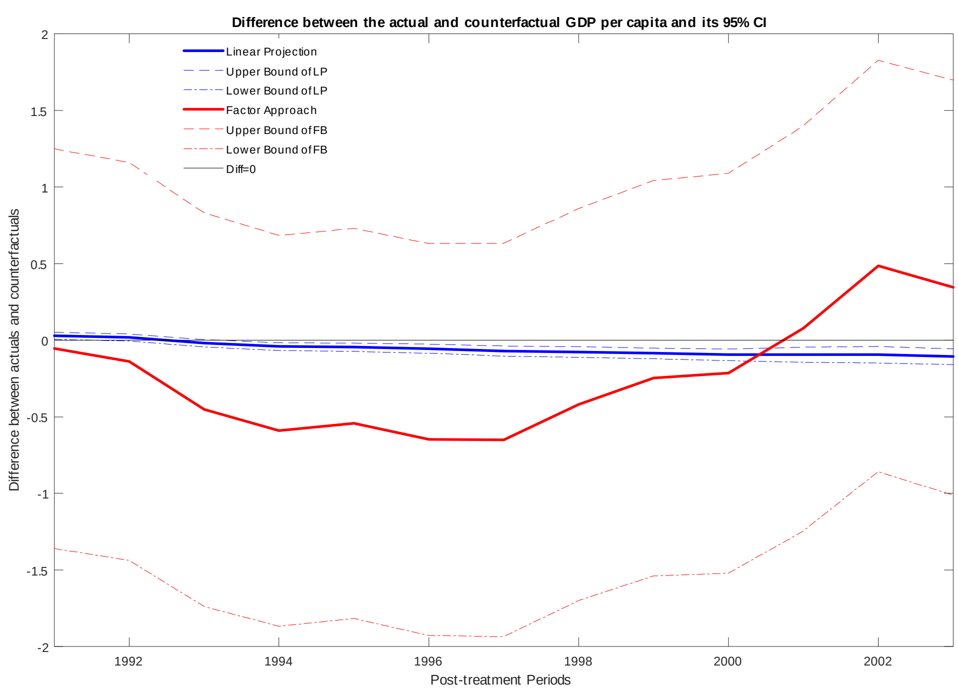


Figure 2: Difference in CIs between the actual and counterfactual GDP per capita in post-treatment periods



9 Concluding Remarks

Under the assumption that the observed data is generated by a factor model with known factor dimension, we consider the factor approach and linear projection approach to construct counterfactuals when either both N and T are large, or N finite, T large or both N and T are finite. We show that when post-treatment information on control units are available, they have the same asymptotic efficiency when both N and T are large. When either N is finite and T is large or both N and T are finite, the linear projection approach dominates the factor approach in terms of the bias and prediction error variance. When T is fixed and N is large, we propose a generalized LP predictor through prediction averaging and show that the generalized LP is more accurate than the FB predictor. Our simulations and an empirical analysis show that in the case of using panel data to construct counterfactuals, the LP approach indeed dominates the factor approach even the dimension of a factor model is known a priori. As a matter of fact, the LP approach can be applied to a variety of data generating processes. The equation[26]

$$y_{it} = E\left(y_{it}|\mathbf{X}_t\right) + \eta_{it}, \tag{9.1}$$

in a sense is a tautology, where $\mathbf{X}_t$ can include $\tilde{\mathbf{y}}_t$, lagged $\tilde{\mathbf{y}}_t$ or any covariates that satisfy $f\left(\mathbf{X}_t|d_{it}\right) = f\left(\mathbf{X}_t\right)$ (e.g., eq(4.6)). Chen (2023) has shown that the LP is an attractive choice against a wide class of matching or difference-in-difference estimators.

However, if post sample information on the control units are not available, it is not clear which method will be more preferable. For post-sample prediction, the analogous FB or LP predictions of $y_{1,T+h}$ for post sample prediction of $\tilde{y}_{1,T+h}$ depends on the predicted $\hat{\mathbf{f}}_{T+h}$,

$$\tilde{y}_{1,T+h} = \hat{\boldsymbol{\lambda}}_1'\hat{\mathbf{f}}_{T+h}, \quad h = 1, 2, \ldots m, \tag{9.2}$$

or predicted $\hat{y}_{1,T+h}$

$$\hat{y}_{1,T+h} = \hat{\mathbf{w}}'\widehat{\tilde{\mathbf{y}}}_{T+h}, \quad h = 1, 2, \ldots m, \tag{9.3}$$

respectively, where $\hat{\boldsymbol{\lambda}}_1$ or $\hat{\mathbf{w}}$ are estimated as the first row of $\hat{\Lambda}$ (which is estimated by (3.3)) or (3.11) using within sample observations $\mathbf{y}_t$, $t = 1, \ldots, T$. However, when no $\mathbf{y}_{T+h}$ are available, there is no consensus about what is the BLUE predictor of $\hat{\mathbf{f}}_{T+h}$ or $\hat{\mathbf{y}}_{T+h}$. For instance, $\hat{\mathbf{f}}_{T+h}$ or $\widehat{\tilde{\mathbf{y}}}_{T+h}$ could be generated from time series modelling of $\mathbf{f}_t$ or $\tilde{\mathbf{y}}_t$ from estimated $\hat{\mathbf{f}}_t$ or observed $\tilde{\mathbf{y}}_t$ for $t = 1, \ldots, T$ (e.g., Box and Jenkins (1976)) or some other methods (e.g., Goncalves and Perron (2014), Goncalves et al (2017)). Under the assumption that

$$\hat{\boldsymbol{\lambda}}_1' Cov\left(\hat{\mathbf{f}}_{T+h}\right)\hat{\boldsymbol{\lambda}}_1 = \hat{\mathbf{w}}' Cov\left(\widehat{\tilde{\mathbf{y}}}_{T+h}\right)\hat{\mathbf{w}}, \tag{9.4}$$

[26]If u_{it} is serially correlated, one may wish to expand the conditional set $\mathbf{y}_t$ to include lagged $\mathbf{y}_t$ to ensure that η_{it} is serially uncorrelated.

the analysis of LP vs FB when post-sample information of $\tilde{\mathbf{y}}_{T+h}$ are available remains applicable for post-sample prediction. However, (9.4) is a strong assumption. Since the prediction error variance of (9.2) or (9.3) depends on $Var\left(\hat{\mathbf{f}}_{T+h}\right)$ or $Var\left(\widehat{\tilde{\mathbf{y}}}_{T+h}\right)$ as well as Λ, $\mathbf{c}$ and Ω. Without such knowledge, it could be difficult to gauge which predictor is likely to yield more accurate prediction.

References

[1] Abadie, A., Diamond, A., & Hainmueller, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science*, 59, 495–510.

[2] Anderson, T.W., 1963, Asymptotic Theory for Principal Component Analysis, *Ann. Math. Statist.* 34(1): 122-148.

[3] Anderson, T.W. and Rubin, H., 1956, *Statistical Inference in Factor Analysis.* In: Neyman, J., Ed., Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability, Vol. 5, Berkeley, 111-150.

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

[5] Bai, J., 2009, Panel Data Models With Interactive Fixed Effects, *Econometrica* 77, 1229–1279.

[6] Bai, J., Li, K., and Lu, L. (2016). Estimation and Inference of FAVAR Models. *Journal of Business & Economic Statistics* 34 (4), 620–641.

[7] Bai, J. and Ng, S. (2021). Matrix completion, counterfactuals, and factor analysis of missing data. *Journal of the American Statistical Association*, 116(536):1746–1763.

[8] Box, G., and G.M. Jenkins, 1976, *Time Series Analysis: Forecasting and Control*, revised Edition, Holden-Day Publication.

[9] Connor, G., and R. Korajczyk, 1986, Performance measurement with the arbitrage pricing theory: A new framework for analysis, *Journal of Financial Economics,* 15, 373-394.

[10] Chamberlain, G., and M. Rothschild, 1983, Arbitrage, Factor Structure, and Mean-Variance Analysis on Large Asset Markets, *Econometrica* 51, 1281-1304.

[11] Chen, J., 2023, Synthetic Control as Online Linear Regression, *Econometrica* 9, 465-491.

[12] Draper, N.R., R.C. Van Nostrand, 1979, Ridge regression and James—Stein estimation, review and comments, *Technometrics*, 21, 451-466.

[13] Ferman, B., C. Pinto, and V. Possebom, 2020, Cherry Picking with Synthetic Controls, *Journal of Policy Analysis and Management* 39, 510–532.

[14] Forni, M., M. Hallin, M. Lippi and L. Reichlin, 1998, The Generalized Dynamic-Factor Model: Identification and Estimation, *The Review of Economics and Statistics* 82, 540-554

[15] Goncalves, S. and Perron, B. (2014). Bootstrapping factor-augmented regression models. *Journal of Econometrics*, 182(1):156–173.

[16] Goncalves, S., Perron, B., and Djogbenou, A., 2017, Bootstrap prediction intervals for factor models. *Journal of Business & Economic Statistics*, 35(1):53–69.

[17] Heckman, J.J., 1997, Constructing Counterfactuals Under Different Assumptions, mimeo, University of Chicago.

[18] Heckman, J.J., and E. Vytlacil, 2001, Local Instrumental Variables, in *Nonlinear Statistical Inference*, edited by C. Hsiao, K. Morimune, and J.L Powell, New York: Cambridge University Press, 1–46.

[19] Hsiao, C., 2021, Some thoughts on prediction in the presence of big data, *China Journal of Econometrics* 1, 1-16.

[20] Hsiao, C., 2022, *Analysis of Panel Data*, 4th edition, Cambridge University Press.

[21] Hsiao C, S. Ching, and S.K. Wan, 2012, A panel data approach for program evaluation: Measuring the benefits of political and economic integration of Hong Kong with mainland China. *Journal of Applied Econometrics* 27, 705–740.

[22] Hsiao, C., Shi, Z. & Zhou, Q., 2022a, Transformed Estimation for Panel Interactive Effects Models, *Journal of Business & Economic Statistics* 40, 1831-1848.

[23] Hsiao, C., Shen, Y. & Zhou, Q., 2022b, Multiple Treatment Effects in Panel-Heterogeneity and Aggregation, *Advances in Econometrics* Volume 43B, 81-101, Essays in honor of Hashem Pesaran.

[24] Hsiao, C., Xie, Y. & Zhou, Q., 2021, Factor dimension determination for panel interactive effects models: an orthogonal projection approach. *Computional Statistics* 36, 1481–1497.

[25] Hsiao, C., and Q. Zhou, 2019, Panel parametric, semiparametric, and nonparametric construction of counterfactuals, *Journal of Applied Econometrics*, 34(4):463–481.

[26] Imbens, G.W., and J.D. Angrist, 1994. Identification and Estimation of Local Average Treatment Effects, *Econometrica*, 62, 467–475.

[27] Imbens, G.W., and T. Lemieux, 2008, Regression Discontinuity Designs: A Guide to Practice, *Journal of Econometrics*, 142, 615–635.

[28] James, W., Stein, C. (1961), "Estimation with quadratic loss", *Proc. Fourth Berkeley Symp. Math. Statist. Prob.*, vol. 1, pp. 361–379

[29] Jiang, H., Li, X., Shen, Y., Wang, X., and Q. Zhou, 2023, Confidence Intervals of Treatment Effects in the Panel Data Approach for Programme Evaluation, working paper.

[30] Jiang, B., Yang, Y., Gao, J., and Hsiao, C. (2021), "Recursive Estimation in Large Panel Data Models: Theory and Practice," *Journal of Econometrics* 224, 439-465.

[31] Ke, X., and C. Hsiao (2022). Economic impact of the most drastic lockdown during COVID-19 pandemic—The experience of Hubei, China, *Journal of Applied Econometrics* 37, 187-209.

[32] Kwon, S. (2021). Optimal shrinkage estimation of fixed effects in linear panel data models. EliScholar–A Digital Platform for Scholarly Publishing at Yal e, p. 1. 2

[33] Lawley, D.N. and Maxwell, A.E., 1971, *Factor Analysis as a Statistical Method*. 2nd Edition, Butterworths, London.

[34] Li, K.T, and D. Bell, 2017, Estimation of average treatment effects with panel data: Asymptotic theory and implementation. *Journal of Econometrics*, 197, 65-75.

[35] Li, X., Y. Shen and Q. Zhou, 2023, Confidence Intervals of Treatment Effects in Panel Data Models with Interactive Fixed Effects, working paper.

[36] Pesaran, M.H., 2006, Estimation and Inference in Large Heterogeneous Panels with Cross Section Dependence, *Econometrica*, 74, 967-1012.

[37] Pesaran, M.H., and E. Tosetti, 2011, Large panels with common factors and spatial correlation, *Journal of Econometrics*, 161, 182-202.

[38] Rosenbaum, P.R., and D.B. Rubin, 1983, The Central Role of the Propensity Score in Observational Studies for Causal Effects, *Biometrika*, 70, 41–55.

[39] Rosenbaum, P.R., and D.B. Rubin, 1985, Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score, *The American Statistician*, 39, 33–38

[40] Ross, S., 1976, The arbitrage theory of capital asset pricing, *Journal of Economic Theory* 13, 341-360.

[41] Sargent, T.J., and Sims, C.A., 1977, Business Cycle Modeling without Pretending to Have Too Much a Priori Economic Theory, Minneapolis: Federal Reserve Bank of Minneapolis.

[42] Smith, G., & Campbell, F. (1980). A critique of some ridge regression methods. *Journal of the American Statistical Association,* 75(369), 74-81.

[43] Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 1135-1151.

[44] Stock, J.H., and M.W. Watson, 1989, New Indexes of Coincident and Leading Economic Indicators, NBER Chapters, in: *NBER Macroeconomics Annual* Volume 4, 351-409, National Bureau of Economic Research, Inc.

[45] Stock, J.H., and M.W. Watson, 2002, Forecasting using principal components from a large number of predictors, *Journal of the American Statistical Association*, 97, 1167–1179.

[46] Tibshirani, R., 1996, Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society, Series B*, 58, 267-288.

[47] Vinod, H. D. (1978). A survey of ridge regression and related techniques for improvements over ordinary least squares. *The Review of Economics and Statistics,* 121-131.

[48] Wan, S.K., Y. Xie and C. Hsiao, 2018, Panel data approach vs synthetic control method, *Economics Letters* 164, 121-123.

[49] Xu, Y., 2017, Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Model, *Political Analysis*, Vol. 25, Iss. 1, January 2017, pp. 57-76.

Table 1: Simulation results of treatment effects using LP and Factor approaches for DGP1 with Case 1

	LP		LP_ave		FB		B-J	
	MAB	MSE	MAB	MSE	MAB	MSE	MAB	MSE
T				$N = 10$				
10	1.756	6.979	-	-	2.032	10.165	2.054	67.239
30	1.375	4.005	-	-	1.598	6.269	1.673	6.061
60	1.246	3.213	-	-	1.479	4.697	1.705	6.624
T				$N = 30$				
10	1.785	7.005	1.577	5.223	1.9149	8.754	2.238	27.625
30	1.436	4.154	-	-	1.4291	5.317	2.074	9.065
60	1.230	2.997	-	-	1.291	4.036	1.975	8.128
T				$N = 50$				
10	1.924	7.825	1.682	5.643	2.252	11.714	2.411	12.296
30	1.515	4.600	1.270	3.266	1.7801	8.232	2.329	11.629
60	1.316	3.363	-	-	1.464	5.752	2.231	10.042
T				$N = 100$				
10	1.986	8.562	1.789	6.654	2.578	15.065	3.16	74.89
30	1.476	4.228	1.551	4.662	1.889	9.151	2.4176	11.574
60	1.317	3.399	1.119	2.445	1.682	7.977	2.444	12.613

Notes: 1. Refer to simulation section for definitions of ”LP”-”BJ” approaches to construct counterfactuals. 2: Refer to simulation section for the calculation of MAB and MSE. 3: ”-” implies the prediction averaging is not necessary for this N and T.

Table 2: Simulation results of treatment effects using LP and Factor approaches for DGP1 with Case 2

	LP		LP_ave		FB		B-J	
	MAB	MSE	MAB	MSE	MAB	MSE	MAB	MSE
T				$N = 10$				
10	1.590	4.752	-	-	1.910	7.508	2.313	305.1
30	1.283	2.827	-	-	1.514	4.223	1.699	5.733
60	1.150	2.196	-	-	1.397	4.022	1.641	5.613
T				$N = 30$				
10	1.579	4.489	1.694	5.551	2.003	8.740	2.130	8.813
30	1.358	3.190	-	-	1.458	4.612	1.991	8.086
60	1.221	2.497	-	-	1.333	4.336	1.949	8.069
T				$N = 50$				
10	1.663	5.099	1.713	5.665	2.450	14.461	2.422	11.374
30	1.373	3.218	1.129	2.143	1.629	6.207	2.238	9.859
60	1.271	2.695	-	-	1.393	4.280	2.165	9.546
T				$N = 100$				
10	1.790	6.262	1.592	4.919	2.666	16.166	2.732	16.773
30	1.327	2.932	1.312	2.898	1.921	9.037	2.435	11.460
60	1.243	2.645	1.038	1.829	1.582	6.184	2.472	12.072

See notes of Table 1.

Table 3: Simulation results of treatment effects using LP and Factor approaches for DGP1 with Case 3

	LP		LP_ave		FB		B-J	
	MAB	MSE	MAB	MSE	MAB	MSE	MAB	MSE
T	$N = 10$							
10	1.401	4.156	-	-	1.718	6.632	2.051	167.37
30	1.009	1.699	-	-	1.273	3.518	1.502	4.806
60	0.909	1.382	-	-	1.152	2.875	1.531	5.131
T	$N = 30$							
10	1.475	4.156	1.246	3.168	1.763	7.620	2.199	26.922
30	1.092	1.991	-	-	1.284	4.082	1.850	7.039
60	0.960	1.509	-	-	1.137	3.281	1.808	6.888
T	$N = 50$							
10	1.611	5.138	1.391	3.955	2.184	11.312	2.395	12.299
30	1.100	2.070	0.963	1.547	1.441	5.703	2.106	8.988
60	0.998	1.654	-	-	1.266	4.423	2.053	8.724
T	$N = 100$							
10	1.711	6.134	1.428	4.015	2.536	15.773	2.847	74.714
30	1.164	2.307	1.062	1.918	1.713	8.086	2.341	10.915
60	1.002	1.671	0.866	1.248	1.457	6.077	2.361	11.330

See notes of Table 1.

Table 4: Simulation results of treatment effects using LP and Factor approaches for DGP1 with Case 4

	LP1		LP_ave		FB		B-J	
	MAB	MSE	MAB	MSE	MAB	MSE	MAB	MSE
T				$N = 10$				
10	1.476	4.277	-	-	1.768	6.764	1.979	252.84
30	1.082	1.987	-	-	1.336	3.824	1.524	5.025
60	0.958	1.545	-	-	1.188	3.075	1.561	5.272
T				$N = 30$				
10	1.576	4.607	1.540	4.627	1.896	7.666	2.044	9.179
30	1.153	2.215	-	-	1.288	3.653	1.809	7.032
60	0.982	1.596	-	-	1.102	2.551	1.836	7.305
T				$N = 50$				
10	1.663	5.480	1.605	5.586	2.279	12.486	2.488	51.758
30	1.202	2.505	1.009	1.699	1.526	6.595	2.130	9.580
60	1.032	1.739	-	-	1.288	4.440	2.131	9.319
T				$N = 100$				
10	1.724	6.027	1.464	4.206	2.531	14.640	2.578	13.806
30	1.198	2.481	1.187	2.326	1.816	8.887	2.405	11.948
60	1.066	1.867	0.902	1.329	1.462	5.922	2.334	11.097

See notes of Table 1.

Table 5: Simulation results of treatment effects using LP and Factor approaches for DGP1 with Case 5

	LP1		LP_ave		FB		B-J	
	MAB	MSE	MAB	MSE	MAB	MSE	MAB	MSE
T				$N = 10$				
10	1.786	6.875	-	-	1.993	8.366	1.985	15.542
30	1.347	3.653	-	-	1.584	5.247	1.664	5.207
60	1.241	3.056	-	-	1.422	4.072	1.618	4.926
T				$N = 30$				
10	1.760	6.152	1.501	4.299	1.844	6.752	2.153	14.002
30	1.413	3.773	-	-	1.363	3.605	1.887	6.357
60	1.281	3.276	-	-	1.270	3.530	1.908	6.772
T				$N = 50$				
10	1.980	7.488	1.689	5.349	2.231	9.652	2.483	14.138
30	1.479	4.193	1.325	3.406	1.643	5.756	2.165	8.057
60	1.311	3.470	-	-	1.352	3.880	2.160	8.255
T				$N = 100$				
10	1.981	7.653	1.695	5.063	2.578	12.573	2.609	11.873
30	1.489	4.291	1.503	4.139	1.890	7.177	2.404	10.059
60	1.339	3.403	1.158	2.749	1.516	4.964	2.336	9.497

See notes of Table 1.

Table 6: Simulation results of treatment effects using LP and Factor approaches for DGP2 with Case 1

	LP		LP_ave		FB		B-J	
	MAB	MSE	MAB	MSE	MAB	MSE	MAB	MSE
T				$N=10$				
10	1.807	7.464	-	-	2.092	10.103	5.453	76.04
30	1.359	3.830	-	-	1.606	5.530	5.062	60.087
60	1.230	3.072	-	-	1.481	5.040	4.966	61.192
T				$N=30$				
10	1.810	7.407	1.556	5.323	1.9613	9.599	6.559	104.01
30	1.456	4.204	-	-	1.4863	5.255	6.093	85.636
60	1.284	3.168	-	-	1.320	4.181	6.153	89.765
T				$N=50$				
10	1.892	7.561	1.706	5.302	2.209	10.521	7.468	136.39
30	1.478	4.400	1.275	3.106	1.717	7.436	7.265	127.1
60	1.329	3.539	-	-	1.447	5.274	7.105	115.4
T				$N=100$				
10	2.109	9.629	3.604	67.617	2.746	17.616	8.806	186.54
30	1.494	4.424	1.557	4.664	1.947	9.424	8.169	150.92
60	1.338	3.475	1.155	2.731	1.690	7.433	7.951	135.01

See notes of Table 1.

Table 7: Simulation results of treatment effects using LP and Factor approaches for DGP2 with Case 2

	LP		LP_ave		FB		B-J	
	MAB	MSE	MAB	MSE	MAB	MSE	MAB	MSE
T				$N = 10$				
10	1.672	5.238	-	-	1.988	7.985	5.272	82.415
30	1.293	2.905	-	-	1.503	4.334	5.147	66.691
60	1.150	2.217	-	-	1.383	3.747	4.977	63.807
T				$N = 30$				
10	1.654	5.151	1.750	6.144	2.163	10.793	6.317	92.148
30	1.354	3.114	-	-	1.423	4.495	6.119	87.472
60	1.203	2.440	-	-	1.268	3.403	6.038	86.754
T				$N = 50$				
10	1.671	5.265	1.782	6.619	2.341	12.638	7.284	127.47
30	1.328	2.997	1.157	2.261	1.676	6.957	7.405	125.50
60	1.263	2.653	-	-	1.424	4.887	6.934	108.17
T				$N = 100$				
10	1.831	6.745	1.589	4.878	2.754	17.152	8.793	169.43
30	1.330	3.013	1.319	2.911	1.866	8.489	7.831	132.02
60	1.257	2.634	1.090	1.987	1.628	6.603	8.050	136.90

See notes of Table 1.

Table 8: Simulation results of treatment effects using LP and Factor approaches for DGP2 with Case 3

	LP		LP_ave		FB		B-J	
	MAB	MSE	MAB	MSE	MAB	MSE	MAB	MSE
T				$N = 10$				
10	1.434	4.099	-	-	1.763	6.799	5.277	71.171
30	1.041	1.828	-	-	1.293	3.506	4.793	56.884
60	0.933	1.443	-	-	1.176	2.876	4.873	60.302
T				$N = 30$				
10	1.453	4.140	1.285	3.441	1.797	8.080	6.514	100.56
30	1.071	1.940	-	-	1.259	3.871	6.262	96.885
60	0.937	1.446	-	-	1.071	2.515	6.200	94.571
T				$N = 50$				
10	1.564	4.950	1.400	3.904	2.139	10.779	7.748	137.77
30	1.121	2.120	0.973	1.608	1.540	6.433	6.933	109.35
60	0.980	1.591	-	-	1.306	4.686	6.997	115.08
T				$N = 100$				
10	1.685	5.851	1.446	4.312	2.563	16.14	8.422	160.36
30	1.140	2.187	1.108	2.191	1.729	7.711	8.038	136.93
60	0.996	1.651	0.883	1.286	1.495	6.342	8.130	149.42

See notes of Table 1.

Table 9: Simulation results of treatment effects using LP and Factor approaches for DGP2 with Case 4

	LP1		LP_ave		FB		B-J	
	MAB	MSE	MAB	MSE	MAB	MSE	MAB	MSE
T				$N = 10$				
10	1.538	4.926	-	-	1.907	7.567	5.093	63.818
30	1.088	2.035	-	-	1.323	3.641	4.824	58.395
60	0.966	1.564	-	-	1.182	2.885	4.857	63.187
T				$N = 30$				
10	1.556	4.975	1.596	5.087	1.951	8.730	6.293	92.523
30	1.162	2.317	-	-	1.318	4.680	6.126	89.728
60	0.995	1.603	-	-	1.195	3.870	6.115	85.686
T				$N = 50$				
10	1.663	5.430	1.601	5.041	2.209	10.775	7.196	122.57
30	1.174	2.314	1.018	1.774	1.466	5.308	7.119	114.34
60	1.029	1.782	-	-	1.284	5.409	7.001	112.98
T				$N = 100$				
10	1.767	6.375	1.398	4.105	2.524	14.575	8.392	160.55
30	1.198	2.479	1.173	2.320	1.842	9.219	8.342	149.73
60	1.064	1.855	0.900	1.366	1.457	6.282	7.947	137.41

See notes of Table 1.

Table 10: Simulation results of treatment effects using LP and Factor approaches for DGP2 with Case 5

	LP		LP_ave		FB		B-J	
	MAB	MSE	MAB	MSE	MAB	MSE	MAB	MSE
T	$N = 10$							
10	1.820	7.041	-	-	2.072	8.693	4.848	45.832
30	1.357	3.704	-	-	1.585	5.121	4.357	34.929
60	1.198	2.891	-	-	1.456	4.454	4.329	36.21
T	$N = 30$							
10	1.798	6.309	1.530	4.354	1.904	7.292	5.468	53.377
30	1.426	3.951	-	-	1.385	3.910	5.348	50.234
60	1.204	2.892	-	-	1.198	3.161	5.235	48.394
T	$N = 50$							
10	1.915	7.232	1.611	4.741	2.233	9.516	6.447	80.298
30	1.469	4.383	1.303	3.152	1.587	5.207	5.730	59.734
60	1.300	3.361	-	-	1.350	4.014	5.739	57.782
T	$N = 100$							
10	2.058	7.894	1.794	6.024	2.573	12.544	6.976	89.892
30	1.479	4.188	1.516	4.162	1.836	6.899	6.604	75.857
60	1.306	3.235	1.129	2.542	1.508	4.522	6.581	75.945

See notes of Table 1.

Appendix: Mathematical Proofs

This appendix provides the proofs that are omitted in the paper.

A.1 Equivalence of LP and FB when $\mathbf{c} \neq 0$ and $(N,T) \to \infty$

When $\mathbf{c} \neq 0$ (i.e., u_{it} is weakly cross-correlated), instead of directly comparing the prediction error variance of (3.12) and (3.6), we consider the transformed model

$$\begin{aligned} y_{1t}^* &= y_{1t} - \mathbf{c}'\Omega^{-1}\tilde{\mathbf{y}}_t \\ &= \boldsymbol{\lambda}_1'\mathbf{f}_t + u_{1t} - \mathbf{c}'\Omega^{-1}(\tilde{\Lambda}\mathbf{f}_t + \tilde{\mathbf{u}}_t) \\ &= \boldsymbol{\lambda}_1^{*\prime}\mathbf{f}_t + u_{1t}^*. \end{aligned} \tag{A.1}$$

where $\boldsymbol{\lambda}_1^* = \boldsymbol{\lambda}_1 - \tilde{\Lambda}'\Omega^{-1}\mathbf{c}$ and $u_{1t}^* = u_{1t} - \mathbf{c}'\Omega^{-1}\tilde{\mathbf{u}}_t$. Then

$$E\left[\begin{pmatrix} u_{1t}^* \\ \tilde{\mathbf{u}}_t \end{pmatrix} (u_{1t}^*, \tilde{\mathbf{u}}_t')\right] = \begin{pmatrix} \sigma_1^{*2} & \mathbf{0} \\ \mathbf{0} & \Omega \end{pmatrix}, \tag{A.2}$$

where $\sigma_1^{*2} = \sigma_1^2 - \mathbf{c}'\Omega^{-1}\mathbf{c}$.

The FB predictor of y_{1t}^* is

$$\tilde{y}_{1t}^* = \hat{\boldsymbol{\lambda}}_1^{*\prime}\hat{\mathbf{f}}_t, \tag{A.3}$$

and the LP predictor of y_{1t}^* is

$$\hat{y}_{1t}^* = \hat{\mathbf{w}}^{*\prime}\tilde{\mathbf{y}}_t, \tag{A.4}$$

where $\hat{\boldsymbol{\lambda}}_1^*$ is the PCA estimates of $\Lambda^* = (\boldsymbol{\lambda}_1^*, \boldsymbol{\lambda}_2, \cdots, \boldsymbol{\lambda}_N)'$, and $\hat{\mathbf{w}}^*$ is

$$\hat{\mathbf{w}}^* = \left(\sum_{t=1}^{T} \tilde{\mathbf{y}}_t\tilde{\mathbf{y}}_t'\right)^{-1} \sum_{t=1}^{T} \tilde{\mathbf{y}}_t y_{1t}^*. \tag{A.5}$$

When $(N,T) \to \infty$, $\hat{\boldsymbol{\lambda}}_1^* \to_p \boldsymbol{\lambda}_1^*$, $\hat{\mathbf{f}}_t \to_p \mathbf{f}_t$ and $\hat{\mathbf{w}}^* \to_p \mathbf{w}^*$. Both (A.3) and (A.4) are unbiased predictors of $y_{1t}^* = y_{1t} - \mathbf{c}'\Omega^{-1}\tilde{\mathbf{y}}_t$ with same asymptotic variance. Since $\hat{y}_{1t}$ and $\tilde{y}_{1t}$ are just $\hat{y}_{1t}^* + \mathbf{c}'\Omega^{-1}\tilde{\mathbf{y}}_t$ and $\tilde{y}_{1t}^* + \mathbf{c}'\Omega^{-1}\tilde{\mathbf{y}}_t$, the FB and LP predictors have the same asymptotic efficiency as $(N,T) \to \infty$ following the derivation when $\mathbf{c} = 0$.

A.2 Dominance of LP and FB when $\mathbf{c} \neq 0$ for $T \to \infty$ and N is finite

When $\mathbf{c} \neq \mathbf{0}$, the LP is equivalent to predicting $y_{1,T+h}$ by $\hat{y}_{1,T+h} = \hat{\mathbf{w}}^{*\prime}\tilde{\mathbf{y}}_{T+h}$ (where $\hat{\mathbf{w}}^*$ is given by (A.5)) with prediction error $y_{1,T+h} - \hat{y}_{1,T+h} = u_{1,T+h} - \mathbf{c}'\Omega^{-1}\tilde{\mathbf{u}}_{T+h}$, and $Var\left(u_{1,T+h} - \mathbf{c}'\Omega^{-1}\tilde{\mathbf{u}}_{T+h}\right) = Var\left(u_{1,T+h}|\tilde{\mathbf{u}}_{T+h}\right) \leq Var\left(u_{1,T+h}\right)$. Similar manipulation of (A.2) can be performed when N is finite and $T \to \infty$. Then just as shown in subsection 5.2, LP predictions have smaller error variance than FB predictions.

A.3 Derivation of FB Prediction Error Variance When both N and T are finite

It is shown by Anderson (1963), Bai (2003) that $\left(\hat{\Lambda}-\Lambda\right)$ or $\left(\hat{\mathbf{F}}-\mathbf{F}\right)$ is asymptotically normally distributed when $T\to\infty$ and $\frac{\sqrt{N}}{T}\to 0$, or $N\to\infty$ and $\frac{\sqrt{T}}{N}\to 0$. However, with finite N and T, the estimators of $\hat{\boldsymbol{\lambda}}_i$ and $\hat{\mathbf{f}}_{T+h}$ are biased. Following Bai and Ng (2002) and Bai (2003), we have the identity

$$\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{y}_t\mathbf{y}_t'\right)\hat{\Lambda}=\hat{\Lambda}\left(N\mathbf{V}_T\right), \tag{A.6}$$

where $\mathbf{V}_T$ is an $r\times r$ diagonal matrix consisting of the first r eigenvalues of $\frac{1}{T}\sum_{t=1}^{T}\mathbf{y}_t\mathbf{y}_t'$ arranged in decreasing order. Then

$$\begin{aligned}\hat{\Lambda}' &= \left(N\mathbf{V}_T\right)^{-1}\hat{\Lambda}'\left(\frac{1}{T}\sum_{t=1}^{T}\left(\Lambda\mathbf{f}_t+\mathbf{u}_t\right)\left(\Lambda\mathbf{f}_t+\mathbf{u}_t\right)'\right)\\ &= \mathbf{V}_T^{-1}\left[\begin{array}{c}\left(\frac{1}{N}\hat{\Lambda}'\Lambda\right)\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{f}_t\mathbf{f}_t'\right)\Lambda'+\left(\frac{1}{N}\hat{\Lambda}'\Lambda\right)\frac{1}{T}\sum_{t=1}^{T}\mathbf{f}_t\mathbf{u}_t'+\frac{1}{N}\hat{\Lambda}'\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{u}_t\mathbf{f}_t'\right)\Lambda'\\ +\frac{1}{N}\hat{\Lambda}'\frac{1}{T}\sum_{t=1}^{T}\mathbf{u}_t\mathbf{u}_t'\end{array}\right]\end{aligned} \tag{A.7}$$

Similarly, we can write

$$\Lambda'=\left(N\mathbf{V}\right)^{-1}\Lambda'\left(\Lambda\Sigma_f\Lambda'+\tilde{\Omega}\right), \tag{A.8}$$

where $\mathbf{V}$ is the $r\times r$ diagonal matrix consisting of the first r eigenvalues of $E\left(\mathbf{y}_t\mathbf{y}_t'\right)$ arranged in decreasing order and $\tilde{\Omega}=E\left(\mathbf{u}_t\mathbf{u}_t'\right)$.

Using the facts that $\mathbf{V}_T-\mathbf{V}=O_p\left(\frac{1}{\sqrt{T}}\right)$, $\frac{1}{T}\sum_{t=1}^{T}\mathbf{u}_t\mathbf{u}_t'-\tilde{\Omega}=O_p\left(\frac{1}{\sqrt{T}}\right)$, $\frac{1}{T}\sum_{t=1}^{T}\mathbf{f}_t\mathbf{f}_t'-\Sigma_f=O_p\left(\frac{1}{\sqrt{T}}\right)$, and $\frac{1}{N}\left(\hat{\Lambda}-\Lambda\right)\Lambda'=O_p\left(\frac{1}{\sqrt{T}}\right)$, then

$$\hat{\Lambda}'-\Lambda'=O_p\left(\frac{1}{T^{3/2}}\right)+\mathbf{V}_T^{-1}\left[\left(\frac{1}{N}\hat{\Lambda}'\Lambda\right)\frac{1}{T}\sum_{t=1}^{T}\mathbf{f}_t\mathbf{u}_t'+\frac{1}{N}\hat{\Lambda}'\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{u}_t\mathbf{f}_t'\right)\Lambda'\right]. \tag{A.9}$$

Let $\hat{\Lambda}=\hat{\Lambda}+\Lambda-\Lambda$, then $E\left(\frac{1}{N}\Lambda'\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{u}_t\mathbf{f}_t'\right)\Lambda'\right)=0$ and

$$\begin{aligned}\frac{1}{N}\left(\hat{\Lambda}-\Lambda\right)'\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{u}_t\mathbf{f}_t'\right)\Lambda' &= \frac{1}{N}\left(\mathbf{F}'\mathbf{F}\right)^{-1}\mathbf{F}'\left(\mathbf{u}_1,\ldots,\mathbf{u}_N\right)\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{u}_t\mathbf{f}_t'\right)\Lambda'\\ &= \frac{1}{N}\Sigma_f^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{f}_t\mathbf{u}_t'\right)\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{u}_t\mathbf{f}_t'\right)\Lambda',\end{aligned} \tag{A.10}$$

when $\hat{\Lambda}$ is estimated by $\hat{\Lambda}'=\left(\mathbf{F}'\mathbf{F}\right)^{-1}\sum_{t=1}^{T}\mathbf{f}_t\mathbf{y}_t'$.[27]

[27] We ignore the higher order bias term due to the estimated $\hat{\mathbf{f}}_t$.

Under Assumption 1-3, from (A.9), we have

$$\begin{aligned} E\left(\hat{\boldsymbol{\lambda}}_i - \boldsymbol{\lambda}_i\right) &= \frac{1}{N}\Sigma_f^{-1}\left(\frac{1}{T^2}\sum_{t=1}^{T}\mathbf{f}_t E\left(\mathbf{u}_t'\mathbf{u}_t\right)\mathbf{f}_t'\right)\boldsymbol{\lambda}_i \\ &= \frac{C}{T}\frac{\sum_{i=1}^{N}\sigma_i^2}{N} = O\left(\frac{1}{T}\right). \end{aligned}$$

Similarly, one can show that the bias of $\hat{\mathbf{f}}_{T+h}$ is $O\left(\frac{1}{N}\right)$. Ignoring the cross-product term in (5.21), which is $O\left(\frac{1}{NT}\right)$, the bias of predicting $y_{1,T+h}$ by $\tilde{y}_{1,T+h} = \hat{\boldsymbol{\lambda}}_1'\hat{\mathbf{f}}_{T+h}$ is biased of order $O\left(\frac{1}{T}\right) + O\left(\frac{1}{N}\right)$. When N and T are finite, $\frac{1}{N}$ or $\frac{1}{T}$ is likely to be a nonzero constant.