# Instrumented Principal Component Analysis*

Bryan Kelly
Yale, AQR, NBER

Seth Pruitt
Arizona State University

Yinan Su
Johns Hopkins University

December 17, 2020

## Abstract

We propose a new approach of latent factor analysis that, in addition to the main panel of interest, introduces other relevant data that serve as instruments for dynamic factor loadings. The method, called IPCA, provides a parsimonious means of incorporating vast conditioning information into factor model estimates. This improves the efficiency of estimates for the latent factors and their loadings, and helps to ascertain the economic relationships among factors and individuals via the observable instruments. The estimation is fast to calculate and accommodates unbalanced panels. We show consistency and asymptotic normality under general panel data generating processes. We demonstrate the advantages of IPCA in simulated data and in applications to equity asset pricing and international macroeconomics.

**Keywords:** IPCA, Latent Factors, Factor Loading Modeling, Instruments, Identification, Large Panel

*Kelly (corresponding author): bryan.kelly@yale.edu; Yale School of Management, 165 Whitney Ave., New Haven, CT 06511; (p) 203-432-2221; (f) 203-436-9604. Pruitt: seth.pruitt@asu.edu; W.P. Carey School of Business, 400 E Lemon St., Tempe, AZ 85287. Su: ys@jhu.edu; Carey Business School, 100 International Drive, Baltimore, MD 21202. We are grateful to Lars Hansen, Dacheng Xiu, Jonathan Wright, Yuehao Bai, and Federico Bandi as well as seminar and conference participants at ASU, Duke, and Montreal for helpful comments. We thank Rongchen Li for excellent research assistance.

Electronic copy available at: https://ssrn.com/abstract=2983919

# 1 Introduction

Latent factor analysis is an empirical workhorse in economics and finance. Its essence is parsimony: a small number of common factors drive the variation of a large cross section (Geweke, 1977; Sargent and Sims, 1977). A large literature has moved beyond using factor analysis merely for exploratory dimension-reduction and data preprocessing, to using it as an estimator for economic models that link hundreds, thousands, or millions of individual variables to a few aggregate economic forces. At the heart of these models are factor loadings—they quantify economic relationships as individuals' sensitivities to the aggregate factors.

High dimensional environments present an econometric challenge because the number of factor loadings to be estimated is increasing with the size of the cross section. The literature studies this issue predominantly in the case of static factor loadings (e.g., Bai, 2003; Stock and Watson, 2002). But the estimation challenges are magnified when factor loadings are time-varying, as called for in many important economic applications. In this case, the number of factor loadings may even exceed the size of the data panel.[1]

We propose instrumented principal component analysis, or IPCA, to overcome the challenges of recovering accurate estimates of latent factor models when panel dimensions are large and/or when loadings are time-varying. Our solution draws on (potentially vast) conditioning data, beyond that in the main panel of interest, to instrument for the factor loadings. This makes it possible to simultaneously bring more data to bear on the statistical model while introducing structure among factor loadings that *reduces* the model's degrees of freedom. In doing so, it also significantly expands the economic content of factor models.

For concreteness, let $t = 1, \ldots, T$ index time and $i = 1, \ldots, N$ index individuals. The IPCA model is

$$x_{i,t} = \beta_{i,t} f_t + \mu_{i,t}, \tag{1}$$

$$\beta_{i,t} = c_{i,t}\Gamma + \eta_{i,t}. \tag{2}$$

[1] Existing solutions for handling dynamic factor loadings rely on serial dependence of the loadings. Recent advances in state-space models view factor loadings as latent dynamic processes themselves (Primiceri, 2005; Pruitt, 2012; Del Negro and Otrok, 2008, among others). Su and Wang (2017) achieve smoothly varying loadings by repeatedly estimating static ones in moving kernels. But this does not relieve the difficulty associated with the need to estimate separate loading for each individual in a large cross section.

Electronic copy available at: https://ssrn.com/abstract=2983919

Equation (1) is a generic factor model in which $x_{i,t}$ is scalar panel data, $f_t$ and $\beta_{i,t}$ are the factors and their loadings ($K \times 1$ and $1 \times K$ vectors, respectively, where $K$ is the number of factors), and $\mu_{i,t}$ is the idiosyncratic error. Equation (2) is the core of IPCA and links information in an $L$-dimensional vector of instrumental variables, $c_{i,t}$, with the factor loadings $\beta_{i,t}$. The key restriction giving content to IPCA is that the mapping from instruments to loadings is fixed over time and across individuals. The mapping is linearly parameterized by the $L \times K$ matrix $\Gamma$, which is the primary target of estimation along with the latent factors $f_t$.[2]

With the advent of "big data," a wealth of new information is increasingly available for economic analysis. Drawing on this wealth, IPCA leverages instrumental conditioning data $c_{i,t}$, which can vary over time and across individuals, to better hone in on the latent factor structure in $x_{i,t}$. Economic theory often suggests that individuals possess ancillary observable attributes that convey information about their factor loadings. For example, Fama and French (1993) note that firm-level observable characteristics "such as size and book-to-market equity must proxy for sensitivity to ... risk factors in returns"—Kelly et al. (2019) (henceforth KPS) use IPCA to formalize this idea and estimate a conditional asset-pricing model with an unprecedented degree of empirical success in equity returns. As another example, option "Greeks" are sensitivities to underlying risk factors and are mechanically linked to observable and dynamic quantities such as the option's moneyness and maturity (Büchner and Kelly, 2020) and are generally observable aspects of the option contract and the option's traded price. As a third example, some macroeconomic models describe inter-firm trade as interaction of nodes on a network. As a firm's centrality increases or decreases it becomes more or less integral to overall economic activity, thus implying dynamic loadings on an aggregate growth factor (Acemoglu and Azar, 2020). Finally,

The common idea underlying each of these examples is that the individual's (option; stock; firm) constant identity becomes irrelevant once we condition on the appropriate attributes of the individual. These attributes (moneyness; book-to-market ratio; network centrality) that dictate the individual's exposure to aggregate fluctuations are often observable and thus ripe for harvesting in the estimation of latent factor models. IPCA operationalizes this logic through equation (2). The matrix $\Gamma$

[2] The error term $\eta_{i,t}$ allows for unobservable behavior of $\beta_{i,t}$ on top of what observable instruments can capture. An orthogonality condition between the instruments and errors, similar to the exclusion restriction in the method of instrumental variable regression, is necessary to achieve consistent model estimates.

Electronic copy available at: https://ssrn.com/abstract=2983919

determines how each of the $L$ attributes maps into each of the $K$ factor loadings. This is a constraint that IPCA imposes on the factor loadings in the generic factor model (1) by requiring that the $\Gamma$ mapping applies uniformly for all individuals in all time periods.[3]

Despite the fact that IPCA brings more data to bear on the factor model, the constraint in equation (2) renders IPCA an especially parsimonious factor model. The philosophy of IPCA is that individual loadings are an unnecessary excess, requiring a divergent count of $N \times K$ parameters in the commonly studied large cross section limit. IPCA displaces the need to estimate individual-specific loading parameters, instead requiring only an understanding of how individuals' attributes map into their factor loadings, summarized by the $L \times K$ parameters in $\Gamma$, whose size does not increase with either of the panel's dimensions. Accordingly, our theoretical analyses consider large panel asymptotics with $N, T$ approaching infinity while holding $L, K$ constant. The ability to assimilate large panel data in a factor model with fixed parameter dimensions means that IPCA's $\Gamma$ estimator converges $\sqrt{N}$ times faster than the individual static loading estimates derived from principal components analysis (PCA).[4]

At the same time, the constraint that IPCA imposes on factor loadings brings new economic content to factor modeling. It probes the economic content of a factor model by estimating which attributes in $c_{i,t}$ are important for capturing the factor structure in $x_{it}$. By evaluating the bindingness of constraints in (1), IPCA provides new tests of the economic determinants of individuals' sensitivities to aggregate shocks, which open a broad avenue of economic discovery. In summary, two directions of improvements—expanded input data and decreased model freedom—combine to form the flexible yet robust estimator of IPCA with deeper economic underpinnings than traditionally ascertained by statistical factor models.

**Estimation** In the following analysis, we often plug $\beta_{i,t}$ in equation (2) into equation (1) and combine their two errors into a new compound error term $e_{i,t}$, yielding an

[3] This logic shares an analogy with state-space models, which impose constraints on the serial dependence of $\beta_{i,t}$; the main difference is that in the typical state space model the dynamics of $\beta_{i,t}$ are inferred only from the main panel data, $x_{i,t}$.

[4] PCA $\beta$'s convergence rates is $\sqrt{T}$ according to Bai and Ng (2002); Bai (2003). IPCA $\Gamma$'s convergence rate is $\sqrt{NT}$ per Theorem 3.

Electronic copy available at: https://ssrn.com/abstract=2983919

equivalent representation of the IPCA model:

$$x_{i,t} = c_{i,t}\Gamma f_t + e_{i,t}, \qquad\qquad e_{i,t} := \eta_{i,t} f_t + \mu_{i,t}. \tag{3}$$

IPCA is estimated as a least squares problem. It minimizes the sample sum of squared errors (compound errors $e_{i,t}$) over parameters $\Gamma$ and $\{f_t\}$ jointly. This least squares estimation is inspired by PCA, which minimizes the sum of squared errors over $\{f_t\}$ and static loading parameters $\{\beta_i\}$.

The optimization does not admit an analytical solution in general, but is speedily solved numerically by an alternating least squares (ALS) algorithm. It iterates between minimizing over $\Gamma$ while holding $\{f_t\}$ fixed, and minimizing over $\{f_t\}$ while holding $\Gamma$ fixed, until convergence. Importantly, the two partial optimization subproblems are simple linear regressions. This has a number of practical benefits. For example, the procedure converges quickly and without the need for complicated numerical algorithms. And IPCA handles unbalanced panels as easily as pooled panel OLS, which is a great advantage in many applications.

Expanding the least squares problem to include nested constraints yields flexible tests of economic hypotheses. For example, restricting the $l$th row of $\Gamma$ to be all zeros aligns with a test for the marginal contribution of the $l$th instrument to overall model performance. As another example, restricting one of the $K$ factors to be some observable macroeconomic time series can be used to test the series' relevance for modeling covariation within the panel. In the context of cross-sectional asset pricing, another example restricts one of the factors to be a constant in order to test whether the factor space admits arbitrage opportunities.[5]

**Asymptotic Results** We show consistency as well as derive the rate of convergence and the limiting distributions of $\Gamma$ and $f_t$ estimations, within the framework of a large panel wherein $N, T$ simultaneously approaches infinity. Notably, $\Gamma$ converges at the rate of $\sqrt{NT}$. As a comparison, PCA loading's ($\beta$) rate is $\sqrt{T}$ since it relies on each individual's time-series information. IPCA additionally picks up cross-sectional information to estimate $\Gamma$—the benefit of bringing the wealth of instrumental information into the problem. Meanwhile, the estimation of $f_t$ achieves a convergence rate of $\sqrt{N}$, the same as PCA. This is because estimating $f_t$ relies on cross-sectional linear

[5] These tests are demonstrated in KPS.

Electronic copy available at: https://ssrn.com/abstract=2983919

regressions, which cannot be more accurate than $\sqrt{N}$.

The asymptotic results are connected to the panel literature if we view IPCA (Eq. 3) as a panel model with time-fixed effects $f_t$ and a structural parameter $\Gamma$. Gagliardini and Gourieroux (2014) study maximum likelihood estimators of general non-linear panel models with time-fixed effects, and arrive at the same convergence rates. However, in our linear factor model, asymptotics are built from the moment condition of the orthogonality between instruments and errors, rather than distributional assumptions.

**Rotational Identification** Rotational unidentification is a well-known issue in factor analysis. In the context of IPCA, the mapping matrix $\Gamma$ and the factors $\{f_t\}$ cannot be separately identified without further assumptions because a parameter pair $\Gamma, \{f_t\}$ and any of its "rotations" $\Gamma R, \{R^{-1}f_t\}$ generate identical sample fits.

It is well understood that the choice of identifying assumption, or "normalization," is not unique, though little work has investigated the effects of this choice on asymptotic properties of factor model estimators. We provide a new analysis that explicitly describes how the normalization choice affects stochastic rates of convergence and asymptotic distributions. This includes an asymptotic decomposition of parameter estimation error into a component arising from the estimation problem absent a normalization, and a separate component purely attributable to the choice of normalization. This decomposition can be applied not only to IPCA, but to latent factor estimators more generally, including PCA.[6] Based on this derivation, we show that the asymptotic results for IPCA discussed above hold under *any* valid identifying assumption.

**Outline** The paper proceeds as follows. Section 2 introduces the notion of a stochastic panel and uses that to describe the generating process of panel data considered by the paper. In Section 3, we discuss estimation and analyze choices of identifying normalization. Based on these preparations, Section 4 proves the consistency of $\Gamma$ estimation, Section 5 presents the asymptotic distributions of $\Gamma$ estimation errors, and Section 6 contains the asymptotics of the factor estimation. Section 7 examines the small sample properties of IPCA estimation with Monte Carlo simulations. Section 8

[6] Bai and Ng (2013) works out the asymptotics of PCA for a few specific normalizations other than the one in Bai (2003), while we develop a unified analysis of IPCA that is general to any normalization choice.

Electronic copy available at: https://ssrn.com/abstract=2983919

applies IPCA to the empirical analyses of international macroeconomics and returns of newly listed stocks. Section 9 concludes.

# 2 The Data Generating Process

## 2.1 Stochastic Panels

We construct what we call "stochastic panels" to model the generating process of two-dimensional data, which for concreteness we refer to as time series and cross-sectional dimensions, respectively. This extends the concept of a stochastic process defined as a transformation that traces out a sequence of sample points, to *two* recombining transformations that trace out a rectangular lattice of sample points.[7] The notion of a stochastic panel helps formalize the genesis of objects that are conventionally expressed with single or double subscripts. For example, variables with a single $t$ subscript can denote *common* or *aggregate* realizations that contain population information about the current cross section. This is the collection of events that is invariant to the "cross-sectional" transformation. Likewise, a single $i$ subscript would imply invariance to the time series transformation, while a double subscript $i, t$ implies dependence on transformations in both dimensions.

A *stochastic panel* is defined as a random variable (vector) $X$, on probability space $\{\Omega, \mathfrak{F}, Pr\}$, with two transformations $\mathbb{S}_{[\mathrm{d}]} : \Omega \to \Omega$ on two different directions $[\mathrm{d}] \in \{[\mathrm{cs}], [\mathrm{ts}]\}$, satisfying the following conditions:

SP.1 **Measurable:** Both transformations are $\mathfrak{F}$-measurable, i.e., $\forall \Lambda \in \mathfrak{F}$, $\mathbb{S}^{-1}_{[\mathrm{d}]}(\Lambda) \in \mathfrak{F}$.

SP.2 **Recombining:** $\mathbb{S}_{[\mathrm{d}]} \circ \mathbb{S}_{[-\mathrm{d}]} = \mathbb{S}_{[-\mathrm{d}]} \circ \mathbb{S}_{[\mathrm{d}]}$.[8]

SP.3 **Measure-preserving:** $\forall \Lambda \in \mathfrak{F}$, $Pr\left(\mathbb{S}^{-1}_{[\mathrm{d}]}(\Lambda)\right) = Pr(\Lambda)$, for either $d$.

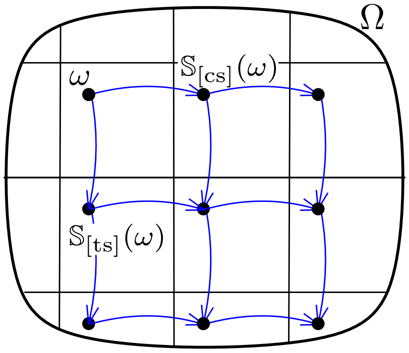
SP.4 **Jointly ergodic:** Any event invariant to both transformations has a probability of either zero or one, i.e., $\forall \Lambda$ s.t. $\Lambda = \mathbb{S}^{-1}_{[\mathrm{d}]}(\Lambda), \forall d, Pr(\Lambda) = 0$ or 1.

[7] For stochastic processes defined with a transformation on the sample space, see for example Hansen and Sargent (2013). This fundamental construction is similar to the one in Gagliardini et al. (2016). They work with a sample space for the time-series process and a continuum to represent individuals. Our sample space can be seen as the Cartesian product of these two sets, and we use two transformation to describe the sampling scheme from a single probability space.

[8] $[-\mathrm{d}]$ means the other direction with respect to $[\mathrm{d}]$.

Electronic copy available at: https://ssrn.com/abstract=2983919

Figure 1: Sample Stochastic Panel Generated by Lattice Points


$\Omega$
$\omega$
$\mathbb{S}_{[\text{cs}]}(\omega)$
$\mathbb{S}_{[\text{ts}]}(\omega)$

*Notes:* The horizontal blue arrows are $\mathbb{S}_{[\text{cs}]}$, vertical blue arrows are $\mathbb{S}_{[\text{ts}]}$. The random variables evaluated at the nine lattice points starting from $\omega$ constitute a sample stochastic panel with $N = 3, T = 3$. The square blocks represent partitions in $\mathfrak{F}$, the columns of blocks are $\mathfrak{F}^{[\text{ts}]}$'s partitions, and the rows blocks are $\mathfrak{F}^{[\text{cs}]}$'s partitions.

SP.5 **Cross-sectional exchangeable:** The sequence of random variables $\{X \circ \mathbb{S}_{[\text{cs}]}^{i-1}, i = 1, 2, \ldots\}$ is exchangeable.

Condition SP.1 is the familiar measurability condition from the definition of stochastic processes except applied in two directions. The interpretation of $\mathbb{S}_{[\text{ts}]}(\omega)$ is "the next period", and the interpretation of $\mathbb{S}_{[\text{cs}]}(\omega)$ is "the next individual" to be sampled, as illustrated in Figure 1.

Condition SP.2 guarantees that the next period's second individual is the same as second individual in the next period. Inductively, starting from an $\omega \in \Omega$, following the two transformations $N$ and $T$ times, respectively, one can trace out an $N \times T$ rectangular lattice of sample points (again see Figure 1). The set of random variables evaluated at the lattice of sample points defines the sample stochastic panel conventionally represented with double subscripts:

$$X_{i,t} := X \circ \mathbb{S}_{[\text{cs}]}^{i-1} \circ \mathbb{S}_{[\text{ts}]}^{t-1}, \qquad \forall i = 1, \cdots, N, \ t = 1, \cdots, T.$$

Furthermore, let sub-$\sigma$ algebra $\mathfrak{F}^{[\text{d}]} \subseteq \mathfrak{F}$ be the collection of events invariant under $\mathbb{S}_{[-\text{d}]}$, the transformation of the other direction. In Figure 1, imagine $\mathfrak{F}$ is generated with the square partitions, then $\mathfrak{F}^{[\text{ts}]}$ and $\mathfrak{F}^{[\text{cs}]}$ consists of the column blocks and row blocks, respectively.

Let $X^{[\text{d}]}$ be $\mathfrak{F}^{[\text{d}]}$-measurable random variables, and it is straightforward to see $X^{[\text{d}]} = X^{[\text{d}]} \circ \mathbb{S}_{[-\text{d}]}$. That is to say, $X_{i,t}^{[\text{ts}]}$ is constant for $i = 1, 2, \cdots$, so the $i$ subscript

Electronic copy available at: https://ssrn.com/abstract=2983919

is redundant and dropped. This provides a definition of random variables that are represented using only a single subscript.

$$
\begin{aligned}
X_t^{[\mathrm{ts}]} &:= X^{[\mathrm{ts}]}\mathbb{S}_{[\mathrm{ts}]}^{t-1}, \qquad && t = 1, \cdots, T, \\
X_i^{[\mathrm{cs}]} &:= X^{[\mathrm{cs}]}\mathbb{S}_{[\mathrm{cs}]}^{i-1}, \qquad && i = 1, \cdots, N.
\end{aligned}
$$

The $t$-subscript random variables are $\mathfrak{F}^{[\mathrm{ts}]}$-measurable. They represent common or aggregate realization and contain distributional information about the "current" cross-sectional population. The factor process $f_t$ is the main example. Symmetrically, $\mathfrak{F}^{[\mathrm{cs}]}$-measurable $i$-subscript variables are about individual $i$'s static characteristics. A static factor loading $\beta_i$ would be of this sort.

Condition SP.3 implies that each direction itself defines a stationary stochastic process in the traditional sense. Stationary stochastic processes admit the one-directional law of large numbers (LLN).[9]

**Lemma 1** (One directional LLN)**.** *Under Conditions SP.1–3 and if* $\mathbb{E}\left\|X\right\|^2 < \infty$, *then*

$$
\frac{1}{N}\sum_{i=1}^{N} X_{i,t} \xrightarrow{L^2} \mathbb{E}\left[X_{\cdot,t} \middle| \mathfrak{F}^{[\mathrm{ts}]}\right], \forall t, \qquad \textit{and} \qquad \frac{1}{T}\sum_{t=1}^{T} X_{i,t} \xrightarrow{L^2} \mathbb{E}\left[X_{i,\cdot} \middle| \mathfrak{F}^{[\mathrm{cs}]}\right], \forall i.
$$

Notice the right-hand sides are $\mathfrak{F}^{[\mathrm{ts}]}$ or $\mathfrak{F}^{[\mathrm{cs}]}$-measurable. That means, for example, the cross-sectional average converges to a time-specific aggregate, which can be written with a single-$t$ subscript. This allows $X_{i,t}$ to be non-ergodic in either direction. For example, if $\omega$ and the next individual $\mathbb{S}_{[\mathrm{cs}]}(\omega)$ are in different $\mathbb{S}_{[\mathrm{ts}]}$-invariant events (different row blocks in Figure 1), $X_{1,t}$ will not repeat the course of events of $X_{2,t}$, no matter how long time lasts. We intentional leave this possibility as a realistic feature of panel data. For example, for the application in Section 8.1, although all countries' import/export shares are varying over time, some countries are inherently more trade-reliant than others. As a result, their time-series averages converge to different limits given by the expectations conditional on country identity ($\mathfrak{F}^{[\mathrm{cs}]}$).

Condition SP.4 introduces and imposes "panel-wise" ergodicity when jointly considering both directions. This allows panel-wise averages to converge to deterministic limits, even as we intentionally allow non-deterministic one-directional limits. The

[9] We write a mean-square convergence result, because it is used for following derivations. An almost-sure convergence result is also obvious.

Electronic copy available at: https://ssrn.com/abstract=2983919

convergence result to deterministic limits paves the foundation for frequentist inference on stochastic panel-based models.

**Lemma 2** (Panel LLN). *Under Conditions SP.1–4, and if* $\mathbb{E}\left\|X\right\|^2 < \infty$*, then as* $N, T \to \infty$*,*

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} X_{i,t} \xrightarrow{L^1} \mathbb{E}X.$$

The four conditions discussed so far are symmetric in the two directions. However, we expect the cross section to be independent in some sense, while the time series evolutions are serially dependent. SP.5 formalizes these properties by strengthening stationarity in SP.3 to exchangeability only for the [cs] direction.[10] Under this condition, $\mathfrak{F}^{[\mathrm{cs}]}$-measurable single $i$-subscript variables (for instance the fixed loadings $\beta_i$ in PCA) are cross-sectionally i.i.d. And, the double-subscript variables are i.i.d. *conditional on* $\mathfrak{F}^{[\mathrm{ts}]}$.[11] For example, $\textit{book-to-market}_{i,t}$ is i.i.d. across stocks *conditional* on the common time-specific information. Unconditionally, it is not independent (only exchangeable) due to aggregate fluctuations of the value ratio.

## 2.2 IPCA Assumptions

We can now formally state the assumptions of the IPCA model using the machinery of stochastic panels. Let $c, \mu, \eta, f^0$ be the random variables of a stochastic panel. Among them, $f^0$ is $\mathfrak{F}^{[\mathrm{ts}]}$-measurable (no $i$ subscript). Based on those primitive variables, equations (1)-(3) hold for each $\omega$, defining $e, \beta, x$ accordingly.[12]

### 2.2.1 Assumptions for Consistency

For consistency, we make the following three assumptions.

**Assumption A** (Instrument orthogonal to error). $\mathbb{E}\left[c_{i,t}^{\top} e_{i,t} \middle| \mathfrak{F}^{[\mathrm{ts}]}\right] = \mathbf{0}_{L\times 1}$.

[10] We note SP.5 is not required for consistency results, but only for asymptotic normality.

[11] The first claim is because ergodic exchangeable processes are i.i.d. The second is de Finetti decomposition of exchangeable processes onto sub-$\sigma$-algebras of invariant events (see Hansen and Sargent, 2013).

[12] When referring to the true parameters, we use a zero superscript, for example $\Gamma^0$ and $f_t^0$, and omit the zero superscript $(\Gamma, f_t)$ for generic parameters depending on the context.

Electronic copy available at: https://ssrn.com/abstract=2983919

This assumption is IPCA's analogue to the exclusion restriction in instrumental variable regression and is a key condition giving content to model equation (2). It is necessary for consistent estimation of $\Gamma$.[13]

**Assumption B** (Moments)**.** *The following moments exist:* (1) $\mathbb{E}\left\|f_t^0 f_t^{0\top}\right\|^2$ (2) $\mathbb{E}\|c_{i,t}e_{i,t}\|^2$ (3) $\mathbb{E}\left\|c_{i,t}^\top c_{i,t}\right\|^2$ (4) $\mathbb{E}\left[\left\|c_{i,t}^\top c_{i,t}\right\|^2 \|f_t^0\|^2\right]$.

**Assumption C.** (1) *The parameter space* $\Psi$ *of* $\Gamma$ *is compact and away from rank deficient:* $\det \Gamma^\top \Gamma > \epsilon$ *for some* $\epsilon > 0$. (2) *Almost surely,* $c_{i,t}$ *is bounded, and* (3) *Define* $\Omega_t^{cc} := \mathbb{E}\left[c_{i,t}^\top c_{i,t}\middle|\mathfrak{F}^{[\text{ts}]}\right]$*, then almost surely* $\det \Omega_t^{cc} > \epsilon$ *for some* $\epsilon > 0$.

The two assumptions above are regularity conditions for consistency. Assumption B lists the required finite moments to apply the panel Law of Large Numbers (Lemma 2). Assumption C guarantees matrix $\Gamma^\top C_t^\top C_t \Gamma$, whose inverse frequently appears, remains nonsingular ($C_t$ denotes the $N \times L$ matrix that stacks up the cross section of $c_{i,t}$).

### 2.2.2 Assumptions for Asymptotic Normality

For asymptotic normality and deriving the asymptotic variance, we impose the following additional assumptions.

**Assumption D** (Central Limit Theorems)**.**
(1) *As* $N, T \to \infty$, $\frac{1}{\sqrt{NT}} \sum_{i,t} \operatorname{vect}\left(c_{i,t}^\top e_{i,t} f_t^{0\top}\right) \xrightarrow{d} \operatorname{Normal}\left(0, \Omega^{cef}\right)$.
(2) *For any* $t$*, as* $N \to \infty$, $\frac{1}{\sqrt{N}} C_t^\top e_t \xrightarrow{d} \operatorname{Normal}\left(\mathbf{0}, \Omega_t^{ce}\right)$.
(3) *As* $N, T \to \infty$, $\frac{1}{\sqrt{T}} \sum_t \operatorname{vecb}\left(f_t^0 f_t^{0\top} - V^{ff}\right) \xrightarrow{d} \operatorname{Normal}\left(\mathbf{0}, \mathbb{V}^{[3]}\right)$*, where* $V^{ff} := \mathbb{E}\left[f_t^0 f_t^{0\top}\right]$.[14]

**Assumption E** (Bounded Dependence)**.** *There exists an* $M < \infty$*, such that* $\forall N, T$, $\frac{1}{NT} \sum_{i,j,t,s} \|\tau_{ij,ts}\| \leq M$*, where* $\tau_{ij,ts} := \mathbb{E}\left[c_{i,t}^\top e_{i,t} e_{j,s} c_{j,s}\right]$.

Assumption D(1) is a panel-wise central limit theorem. It gives the asymptotic distribution of the dominant term in the score function $S(\Gamma^0)$ (see Lemma 3). Assumption D(2) is a cross-sectional central limit theorem conditional on aggregate time

[13] Notice, an alternative to Assumption A is its sufficient condition of imposing orthogonality with respect to each of the two primitive errors: $\mathbb{E}\left[c_{i,t}^\top \eta_{i,t}\middle|\mathfrak{F}^{[\text{ts}]}\right] = \mathbf{0}$ and $\mathbb{E}\left[c_{i,t}^\top \mu_{i,t}\middle|\mathfrak{F}^{[\text{ts}]}\right] = \mathbf{0}$.

[14] vect$(A)$ is an operator that vectorizes matrix $A$'s elements to a column vector by going right first, then down. For a square matrix $A$, vecb$(A)$ vectorizes $A$'s upper triangular entries, *not* including the diagonal, to a vector by going right first, then down. See details with an example in Appendix A.

Electronic copy available at: https://ssrn.com/abstract=2983919

series information ($\mathfrak{F}^{[\mathrm{ts}]}$). It gives factor estimation's asymptotic distribution (see Theorem 4.b), and works in a similar way as Bai's (2003) Assumption F3. Assumption D(3) is used by the results in Appendix B, which is related to how orthogonalizing factors (as a normalization) introduces "contamination" to estimation.

Assumption E sets a bound for the times series and cross-sectional dependency of $c_{i,t}^{\top} e_{i,t}$ for proving asymptotic normality (see Lemma 3). It is analogous to Bai's (2003) Assumption C4.

**Assumption F** (Constant Second Moment of Instruments)**.** $\Omega_t^{cc}$ *is constant at* $\Omega^{cc}$*.*

Assumption F shuts down the variation of $c_{i,t}$'s cross-sectional second moment, in order to retain concise expressions for the asymptotic variances in the specific identification cases (Theorem 3). The essence of the asymptotic theory does not depend on Assumption F. For example, the convergence rates would not change if it were relaxed.

# 3 Estimator as Normalized Optimizer

Estimation consists of two steps: optimizing the objective function to find the set of equivalent solutions, followed by a normalization of the solutions that selects a unique element of this set to serve as the parameter estimate. The second step is necessary because IPCA has a rotational unidentification issue well-known from other latent factor settings. While our identification approach in some sense follows convention in the literature, we more formally construct identification conditions. This formality is not just for defining the estimator, but for more rigorously expounding the notion of a "true" parameter and the corresponding measurement of estimation errors. As a result, we develop a general asymptotic analysis framework for estimators normalized by *any* valid identifying assumption, which is also applicable in contexts beyond IPCA (such as in the asymptotic analysis of PCA).

Electronic copy available at: https://ssrn.com/abstract=2983919

## 3.1 Optimization

IPCA is estimated as a least squares problem that minimizes the sample sum of squared errors (SSE) over parameters $\Gamma$ and $\{f_t\}$:

$$\min_{\Gamma,\{f_t\}} \sum_{i,t} \left(x_{i,t} - c_{i,t}\Gamma f_t\right)^2 . \tag{4}$$

This objective is inspired by PCA which also optimizes the sample SSE but over parameters $\{\beta_i\}$ and $\{f_t\}$.

We use an Alternating Least Squares (ALS) method for the numerical solution of the optimization because, unlike PCA, the IPCA optimization problem does not have a solution through an eigen-decomposition.[15] The SSE target in (4) is quadratic in either $\Gamma$ or $\{f_t\}$ when the other is fixed. This property allows for analytical optimization of $\Gamma$ and $\{f_t\}$ one at a time. Given any $\Gamma$, factors $\{f_t\}$ are $t$-separable and solved with cross-sectional OLS for each $t$:

$$\widehat{f}_t(\Gamma) := \arg\min_{f_t} \sum_{i} \left(x_{i,t} - c_{i,t}\Gamma f_t\right)^2 = \left(\Gamma^\top C_t^\top C_t \Gamma\right)^{-1} \Gamma^\top C_t^\top x_t.\text{[16]} \tag{5}$$

Symmetrically, given $\{f_t\}$, the optimizing $\Gamma$ (vectorized as $\gamma$) is solved with pooled panel OLS of $x_{i,t}$ onto $LK$ regressors, $c_{i,t} \otimes f_t^\top$:

$$\arg\min_{\gamma} \sum_{i,t} \left(x_{i,t} - c_{i,t}\Gamma f_t\right)^2 = \left(\sum_{i,t} \left(c_{i,t}^\top \otimes f_t\right) \left(c_{i,t} \otimes f_t^\top\right)\right)^{-1} \left(\sum_{i,t} \left(c_{i,t}^\top \otimes f_t\right) x_{i,t}\right). \tag{6}$$

Throughout the paper, the lower case $\gamma$ represents vectorized $\Gamma^\top$ matrix, $\gamma = \text{vect}\left(\Gamma\right)$ (likewise for $\gamma^0$, $\widehat{\gamma}$, and so forth).

[15] In the special case that $C_t^\top C_t$ is constant across $t$, the solution of the IPCA estimation optimization reduces to the SVD of the interaction of $x_{i,t}$ and $c_{i,t}$, which is similar to Fan et al. (2016) projected principal component analysis; however, they did not consider time-varying instruments nor time-varying loadings. The SSE target can be transformed to a sum of multivariate Rayleigh quotients with inverted matrices $\left(\Gamma^\top C_t^\top C_t \Gamma\right)^{-1}$. Were $C_t^\top C_t$ invariant, these matrices are invariant and can be pulled out of the $t$-summation, leading to a case similar to PCA where the SSE can be minimized with SVD. A previous version of this paper and the Appendix of KPS detail the solution under this special case.

[16] Here, $x_t$ is the $N \times 1$ vector of $x_{i,t}$, $C_t$ is the $N \times L$ matrix of $c_{i,t}$. Notice $\widehat{f}_t(\Gamma)$ depends on the sample cross-section and is not $\mathfrak{F}^{[\text{ts}]}$-measurable (although single $t$ subscript). A more proper but cumbersome indexing would be $\widehat{f}_{N,t}$

Electronic copy available at: https://ssrn.com/abstract=2983919

The ALS algorithm begins with an initial guess and iterates between updates of $\Gamma$ and $\{f_t\}$ while holding the other fixed according to the above first-order conditions. The program stops when the optimality conditions are satisfied up to a predetermined tolerance.[17] The ALS method can be seen as a Block Coordinate Descent algorithm with $\Gamma$ and $\{f_t\}$ as the two blocks. In practice, it converges after a few hundred iterations in a matter of seconds in our empirical and simulated datasets. The speedy calculation is a notably important feature that will allow for widespread application of IPCA. For example, it is easy to repeatedly estimate the model for a large number of generated samples as required in simulation-based inference procedures.

Notice the numerical method easily adapts to unbalanced panels, because both the partial optimizations are regressions which tolerate missing values. In particular, we can rewrite Equations (4)-(6) with the minor changes for all the summation signs to summing over *only* the observed $i, t$ panel entries. Then the same procedure follows through.[18]

The following sections focus on $\Gamma$ first by concentrating out $f_t$ as an intermediate object. Define *target function* $G(\Gamma)$ as the concentrated SSE objected in (4) with respect to $\Gamma$ only,

$$G(\Gamma) = \frac{1}{2NT} \sum_{i,t} \left( x_{i,t} - c_{i,t} \Gamma \widehat{f}_{t(\Gamma)} \right)^2, \tag{7}$$

where $\widehat{f}_{t(\Gamma)}$ is the optimal factor given any $\Gamma$ according to (5). And, define *score function* as the derivative: $S(\widehat{\Gamma}) = \frac{\partial G(\Gamma)}{\partial \gamma}$. Therefore, the two-argument joint minimization problem (Eq. 4) is equivalent to minimizing $G(\Gamma)$ or solving the first order condition $S(\Gamma) = \mathbf{0}_{LK\times 1}$, with respect to $\Gamma$ only. The asymptotic analysis proceeds by first analyzing $S$ to characterize the asymptotics of $\widehat{\Gamma}$ while treating factor estimate $\widehat{f}_t(\widehat{\Gamma})$ as an implicit intermediate. Once the asymptotic analysis of $\widehat{\Gamma}$ is done, Section 6 comes back to the factor asymptotics by plugging $\widehat{\Gamma}$ in.

[17] We can prove uniqueness for the asymptotic score function up to rotation in the sense of Proposition 2, which provides a theoretical foundation of the ALS method. In simulations we see that convergence is unique and fast unless data generating errors are simulated to be unreasonably large.

[18] In practice, we first "fill up" data $\{x_{i,t}, c_{i,t}\}$ at any unobserved $i, t$ entry with zeros, and then use the same ALS program for the completed panel. It is easy to show this process is equivalent to summing over only the observed entries.

Electronic copy available at: https://ssrn.com/abstract=2983919

## 3.2 Normalization

In population, a class of "rotated" true parameters $\Gamma^0 R$ yields exactly the same data generating process for $x_{i,t}, c_{i,t}$, so long as true factors are rotated inversely as $R^{-1} f_t^0$ (for any full rank $K \times K$ matrix $R$). The counterpart sample issue is that the target and score are invariant to rotation: if $\widehat{\Gamma}$ solves $S(\Gamma) = \mathbf{0}$, then any rotation $\widehat{\Gamma} R$ is also a solution.

IPCA estimation deals with the issue by following the convention in the latent factor literature—it selects a normalization from the set of many rotational equivalent optimizers as the unique estimator. The choice of normalization is often made based on economic interpretability, algebraic elegance, or computational convenience.

The convention in the literature is to derive asymptotic theory of an estimator conditional on a *specific* normalization. In contrast with this convention, we provide a general method to characterize asymptotic properties under *any* valid normalization. To this end, we construct the following concepts for a generic normalization. We will discuss two specific normalization examples towards the end of this section. The asymptotic analysis in the following sections are first derived with the generic normalization, based on which the specific cases become straightforward evaluations.

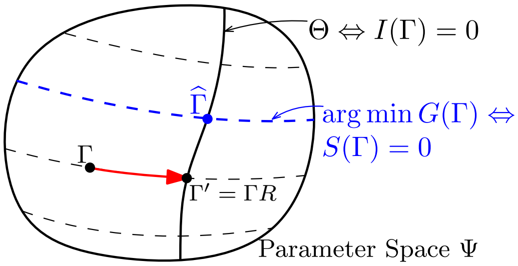
To define key concepts, let *parameter space* $\Psi$ be the set of all possible parameters $\Gamma$. A parameter $\Gamma$ is called *rotationally equivalent* to $\Gamma'$ if there exists an invertible matrix $R$ such that $\Gamma = \Gamma' R$. Define an *identification condition* as a subset $\Theta \subset \Psi$, such that for any $\Gamma \in \Psi$ there is a unique $\Gamma' \in \Theta$ which is rotationally equivalent to $\Gamma$. Associated with an identification condition, the *normalization* is the mapping from any $\Gamma$ to the (unique) equivalent $\Gamma' \in \Theta$. Define *identification function* $I(\Gamma)$ as a (vector-valued) function on the parameter space such that its solution set constitute an identification condition: $\Theta = \{\Gamma \mid I(\Gamma) = \mathbf{0}\}$. The identification function operationalizes a normalization, and is analyzed in tandem with the score function in our asymptotic analysis.

Figure 2 illustrates these concepts. Dashed lines through the parameter space $\Psi$ represent sets of rotationally equivalent parameters. The data generating process is unchanged for parameters in the same set. In addition, the target function $G$ is invariant within each set. The particular set that minimize $G$ is drawn in blue.

We have plotted a generic identification condition $\Theta$ as the solid black line which intersects each dashed line at only one point. There are many possible normalizations, representable as different $\Theta$'s that "cut through" all of the rotationally equivalent

Electronic copy available at: https://ssrn.com/abstract=2983919

Figure 2: Identification Condition, Normalization, and Estimation



*Notes:* The dashed lines are sets of rotationally equivalent parameters. A particular one (in blue) minimizes the target function. The black curve represents the identification condition $\Theta$. The intersection of sets $\Theta$ and $\arg\min G(\Gamma)$ defines the estimator $\widehat{\Gamma}$. The red arrow represents the normalization of a parameter to $\Theta$.

parameter sets *each only once*. We will analyze two concrete examples of $\Theta$ further blow.

## 3.3 Estimator Defined as Normalized Optimizer

The *estimator* is defined as the optimizer of target function $G(\Gamma)$ that is normalized by $\Theta$:

$$\widehat{\Gamma}(\Theta) = \arg\min_{\Gamma \in \Theta} G(\Gamma). \tag{8}$$

When there is no emphasis on the specific normalization choice, we omit "$(\Theta)$" and simply write $\widehat{\Gamma}$.

Despite first appearances, (8) is not a constrained optimization because the constraint "$\Gamma \in \Theta$" never restrains the target from achieving its global optimum—it only picks a unique representation out of the (rotationally equivalent) set of $\Gamma$'s that all achieve the same minimum. Equivalently, $\widehat{\Gamma}$ is the solution of the simultaneous equations $S(\Gamma) = \mathbf{0}$ and $I(\Gamma) = \mathbf{0}$, shown as the intersection of the two corresponding curves in Figure 2. Based on this representation, we will characterize the asymptotics of $\widehat{\Gamma}$ by analyzing the score and identification functions.

In addition, we note that a normalization $\Theta$ must be known to the econometrician, meaning it can depend on the sample but cannot depend on the underlying population parameters.

Electronic copy available at: https://ssrn.com/abstract=2983919

## 3.4 Two Normalization Examples

We give two specific normalization examples that are convenient and interpretable, called $\Theta_{\text{X}}$ and $\Theta_{\text{Y}}$. We will keep returning to these two examples for concrete asymptotic analysis, and always use "X" and "Y" subscripts when referring to the constructions. For example, $I_{\text{X}}(\Gamma)$ and $I_{\text{Y}}(\Gamma)$ are identification functions for the two cases.

$\Gamma$ has $LK$ degrees of freedom in total, within which $K^2$ degrees are unidentified since the rotation $R$ is $K \times K$. So an identification condition must restrict $K^2$ degrees of freedom. The first example pins down the top $K \times K$ block of $\Gamma$ as the identity matrix, and leaves the lower $L-K$ rows free. Specifically, define $\Theta_{\text{X}} := \{\Gamma \in \Psi \mid Block_{1:K}(\Gamma) = \mathbb{I}_K\}$, where $Block_{1:K}(\cdot)$ is a function of a matrix that cuts out its first $K$ rows. The associated normalization mapping is $\Gamma' = \Gamma Block_{1:K}(\Gamma)^{-1}$.

This identification condition forces a one-to-one correspondence between the $K$ factors and the first $K$ instruments, which gives the factors an economic interpretation dictated by those instruments.[19] Using a familiar asset pricing example Fama and French (1993), suppose there are three factors ($K = 3$), that $x_{i,t}$ consists of monthly stock returns, and that the first three instruments are lagged size, book-to-market ratio, and momentum of a stock ($c_{i,t} = [size_{i,t-1}, bm_{i,t-1}, mom_{i,t-1}, ...]$). We know that $\beta_{i,t}$ is given by $c_{i,t}\Gamma$. Under $\Theta_{\text{X}}$, the first loading is $\beta_{i,t,1} = 1 \cdot size_{i,t-1} + 0 \cdot bm_{i,t-1} + 0 \cdot mom_{i,t-1} + \textit{effects of other instruments}$—loadings on the first factor are driven one-for-one by the first instrument (size), but are unaffected by book-to-market and momentum, giving the first factor a direct interpretation as a size factor analogous to the "SMB" factor of Fama and French (1993). Likewise, the second and third factors are book-to-market and momentum factors. The lower $L-K$ rows of $\Gamma$ tell us how the remaining $L-K$ instruments affect loadings. Hence, this normalization is reminiscent of, and a less ad-hoc alternative to, the popular characteristics-sorted portfolios that the empirical asset pricing literature has adopted for understanding systematic risks.[20]

[19] Without loss of generality, we can reorder the instruments however we wish.

[20] This identification condition $\Theta_{\text{X}}$ appears similar to identifying restriction PC3 in Bai and Ng (2013), but carries an important distinction due to IPCA's utilization of instruments. In the static-loading PCA estimator, the PC3 restriction says that the first $K$ *individuals* have a direct correspondence to the $K$ factors. Continuing the asset-pricing example, it would imply that the factors are aligned with the first $K$ individual stocks which, from a financial economics point-of-view, is ad-hoc and not intuitive for understanding the factor space.

Electronic copy available at: https://ssrn.com/abstract=2983919

For the second normalization, define $\Theta_{\mathrm{Y}}$ as the set of $\Gamma \in \Psi$ such that [1] $\Gamma$ is ortho-normal: $\Gamma^\top \Gamma = \mathbb{I}_K$; and [2] $\widehat{f}_t(\Gamma)$ is orthogonal: $\frac{1}{T}\sum_t \widehat{f}_t(\Gamma)\widehat{f}_t^\top(\Gamma)$ is diagonal, with distinct and descending entries. Appendix C.4 verifies $\Theta_{\mathrm{Y}}$ is indeed an identification condition that satisfies the uniqueness condition, and writes out the associated normalization procedure.[21] This identification condition is a familiar PCA convention and is implemented for IPCA in KPS. It chooses a set of orthogonal factors to represent the factor space. Notice [1] and [2] pin down $\frac{1}{2}K(K+1)$ and $\frac{1}{2}K(K-1)$ degrees of freedom, respectively, for a total of $K^2$. Importantly, identification condition $\Theta_{\mathrm{Y}}$ depends on the sample, because $\widehat{f}_t(\Gamma)$ in [2] is a sample-dependent function. This sample-dependence brings some complexity in the asymptotic analysis which we will address in the following sections.

# 4 Consistency of $\Gamma$ Estimation

This section demonstrates that IPCA estimates the mapping matrix $\Gamma$ consistently. Since the estimate $\widehat{\Gamma}$ is the simultaneous solution of the first order condition (score function) and the identification function, its consistency is based on the (uniform) convergence of these two functions, following the standard strategy for analyzing $M$-estimators (Newey and McFadden, 1994). Relative to the classical framework, we must confront two additional challenges: simultaneous $N, T$ convergence and the identification issue. To address the former, we rely on the large sample properties of stochastic panels developed in Section 2.1. For the latter, we build on the normalization concepts constructed in Section 3.2.

## 4.1 Score Function Convergence

We show the score function uniformly converges to a deterministic limiting function:

**Proposition 1** (Uniform Convergence of the Score Function)**.**

[21] An additional nuance is that the signs of the $K$ $\Gamma$ columns (and the corresponding $K$ factors) need one more restriction to pin down. If the additional restriction is not suitable, it brings problems in finite sample simulations (for example, the one Stock and Watson (2002) proposed). Appendix D explains the issue in detail and constructs a sign restriction [3] that is theoretically sound and practically easy to use in simulations.

Electronic copy available at: https://ssrn.com/abstract=2983919

*Under Assumptions A–C, the score function converges uniformly in probability:*

$$\sup_{\Gamma\in\Psi}\left\|S(\Gamma)-S^0(\Gamma)\right\|\xrightarrow{p}0,\qquad\qquad N,T\to\infty.$$[22]

Moreover, we verify that the limiting function $S^0$ is solved *only* by $\Gamma^0$ and its unidentified rotations:

**Proposition 2** (Only True Parameters Solve Limiting Score)**.** *Under Assumption C, $S^0(\Gamma)=\mathbf{0}$ if and only if $\Gamma$ is rotationally equivalent with $\Gamma^0$.*

These two results are the foundation of IPCA consistency. When combined, they imply a sense of *set* convergence regardless of the identification issue—the set of SSE minimizers $\{\Gamma \text{ s.t. } S(\Gamma)=0\}$ converges to the set of unidentified true parameters. Once we build the counterparts of these results for the identification function in the next subsection, score and identification functions together will lead to estimator convergence.

But before that, let us explain the intuition behind Proposition 1, which is essential for IPCA's large sample theory. By taking the derivative of the target, we find $S(\Gamma)=\frac{1}{NT}\sum_t \operatorname{vect}\left(C_t^\top\widehat{e}_t(\Gamma)\widehat{f}_t^\top(\Gamma)\right)$, where $\widehat{f}_t(\Gamma)$ and $\widehat{e}_t(\Gamma)$ are the coefficients and errors, respectively, of the cross-sectional OLS regression of $x_t$ onto $C_t\Gamma$.[23] The sample OLS estimate $\widehat{f}_t(\Gamma)$ misses the true $f_t^0$ for two reasons. For one, it uses a $\Gamma$ that is not the true $\Gamma^0$, and thereby the instrumented factor loadings $C_t\Gamma$ are off. Second, even if we knew the true $\Gamma^0$ and thus the sample $\beta$'s were right, OLS in the finite cross section does not exactly reveal the true $f_t^0$. To formalize this intuition, we construct $\widetilde{f}_t(\Gamma)$ as the "population" counterpart of the same $x_t$-on-$C_t\Gamma$ regression. By "population" we mean conditional on $\mathfrak{F}^{[\mathrm{ts}]}$: that is, the collection of all "aggregate" events which contain the information of how the "current" cross section would be generated. Specifically,

$$\widetilde{f}_t(\Gamma):=\mathbb{E}\left[\Gamma^\top c_{i,t}^\top c_{i,t}\Gamma\middle|\mathfrak{F}^{[\mathrm{ts}]}\right]^{-1}\mathbb{E}\left[\Gamma^\top c_{i,t}^\top x_{i,t}\middle|\mathfrak{F}^{[\mathrm{ts}]}\right]=\left(\Gamma^\top\Omega_t^{cc}\Gamma\right)^{-1}\Gamma^\top\Omega_t^{cc}\Gamma^0 f_t^0,\tag{9}$$

with the shorthand $\Omega_t^{cc}:=\mathbb{E}\left[c_{i,t}^\top c_{i,t}\middle|\mathfrak{F}^{[\mathrm{ts}]}\right]$.[24] Therefore, the forementioned two sources

[22] The expression of $S^0$ is in Appendix C.6 together with the proof.

[23] Eq. (5) defined $\widehat{f}_t(\Gamma)$, and $\widehat{e}_t(\Gamma):=x_t-C_t\Gamma\widehat{f}_t(\Gamma)$.

[24] Assumption A is used in Eq. (9).

Electronic copy available at: https://ssrn.com/abstract=2983919

of $\widehat{f}_{t(\Gamma)}$ error are formally represented with the decomposition

$$\widehat{f}_{t(\Gamma)} = \widetilde{f}_{t(\Gamma)} + \left(\Gamma^\top C_t^\top C_t \Gamma\right)^{-1} \Gamma^\top C_t^\top \widetilde{e}_{t(\Gamma)}, \tag{10}$$

where $\widetilde{e}_{t(\Gamma)} := x_t - C_t \Gamma \widetilde{f}_{t(\Gamma)}$ is the corresponding population OLS error. Importantly, given a rotation $R$, $\widetilde{f}_t\left(\Gamma^0 R\right) = R^{-1} f_t^0$ and $\widetilde{e}_t\left(\Gamma^0 R\right) = e_t$. This means if one knew the true $\Gamma^0$ (even if not knowing the particular rotation), functions $\widetilde{f}_{t(\Gamma)} and \widetilde{e}_{t(\Gamma)}$ would be able to reveal the true factor structure. The second term in (10) captures the remaining error caused only by the finite cross section.

With this knowledge, the score function can be broken down by representing $\widehat{f}_{t(\Gamma)}, \widehat{e}_{t(\Gamma)}$ with $\widetilde{f}_{t(\Gamma)}, \widetilde{e}_{t(\Gamma)}$. We save the exact expression to Appendix C.6, except noting that the score function inherits the same decomposition: into a part that is only "off" due to the "wrong" $\Gamma$, and another part due to finite cross section. Intuitively, the first part converges to the probability limit $S^0(\Gamma)$, while the the second vanishes in the large-$N, T$ limit.[25] Finally, the limiting score $S^0(\Gamma)$ can recover the true $\Gamma^0$ (up to a rotation) according to Proposition 2.

## 4.2 Identification Function Convergence

As defined in Subsection 3.2, an identification function is solved by the normalized parameters: $\Theta = \{\Gamma \text{ s.t. } I(\Gamma) = \mathbf{0}\}$. We define the following identification functions for the the two examples $\Theta_{\mathrm{X}}$ and $\Theta_{\mathrm{Y}}$, respectively:

$$I_{\mathrm{X}}(\Gamma) := \operatorname{vect}\left(Block_{1:K}(\Gamma) - \mathbb{I}_K\right), \quad I_{\mathrm{Y}}(\Gamma) := \begin{bmatrix} \operatorname{veca}\left(\Gamma^\top \Gamma - \mathbb{I}_K\right) \\ \operatorname{vecb}\left(\frac{1}{T}\sum_t \widehat{f}_{t(\Gamma)} \widehat{f}_t^\top{}_{(\Gamma)} - V^{ff}\right) \end{bmatrix}.^{26}$$

Notice $\widehat{f}_{t(\Gamma)}$ involves sample data, and hence $I_{\mathrm{Y}}$ is a random function.

We show $I_{\mathrm{Y}}$ converges uniformly to a deterministic function $I_{\mathrm{Y}}^0$, which also constitutes an identification function. The same claim trivially applies to $I_{\mathrm{X}}$ as it is

[25] The proof is in Appendix C.6, which deals with additional complication of *uniform* large-$N, T$ convergence across $\Gamma$, which is necessary for the convergence discussed next.

[26] Mappings veca and vecb vectorize the upper triangular entries of a square matrix. The difference is veca includes the diagonal entries, and vecb does not. See details with an example in Appendix A. $V^{ff}$ is the factor's *population* second moment matrix as defined in Assumption D. It is actually redundant to write "$-V^{ff}$" in the expression of $I_{\mathrm{Y}}$, as long as the assumed $V^{ff}$ is a diagonal matrix, since vecb ignores the diagonal entries. However, we leave it in to note $\Theta_{\mathrm{Y}}$ can be easily adjusted at applied researcher's discretion with the asymptotic analysis largely unchanged. For example, one can specify $V^{ff} = \mathbb{I}_K$ and switch "veca" and "vecb" to standardize factors instead of $\Gamma$.

Electronic copy available at: https://ssrn.com/abstract=2983919

deterministic to begin with. In detail, Appendix C.8 verifies both $I_{\mathrm{X}}, I_{\mathrm{Y}}$ satisfy the following conditions about a generic identification function $I$:

IF.1 **Uniform convergence:** There exists a deterministic function $I^0(\Gamma)$ such that

$$\sup_{\Gamma\in\Psi}\left\|I(\Gamma)-I^0(\Gamma)\right\|\xrightarrow{p}0,\qquad\qquad N,T\to\infty.$$

IF.2 **Limit is Identification Condition:**

Set $\Theta^0 := \{\Gamma\in\Psi \mid I^0(\Gamma)=\mathbf{0}\}$ is an identification condition.

In addition, we maintain the identification assumption about the true $\Gamma^0$:

IA **Identification Assumption:** $\Gamma^0\in\Theta^0$, i.e., $I^0(\Gamma^0)=\mathbf{0}$.

IF.2 and IA combined imply $\Gamma^0$ solves $I^0$, and it is the *only* solution among all of its equivalent rotations.

The above convergence and uniqueness conditions about $I$ are symmetric to Propositions 1 and 2 about $S$. Since the estimator is the intersection of $S$ and $I$, these conditions together form the premise of estimator consistency, which will be formally stated in Subsection 4.4. Therefore, IF.1–2 can be used to establish if any normalization, besides $\Theta_{\mathrm{X}}$ and $\Theta_{\mathrm{Y}}$, yields a consistent parametric estimation.

We note IA is an innocuous assumption because the true parameter can always be normalized to satisfy any identification condition, including $\Theta^0$, without changing the data generating process. Next, we sort out some ambiguity brought about by having different, but rotationally equivalent, true parameters.

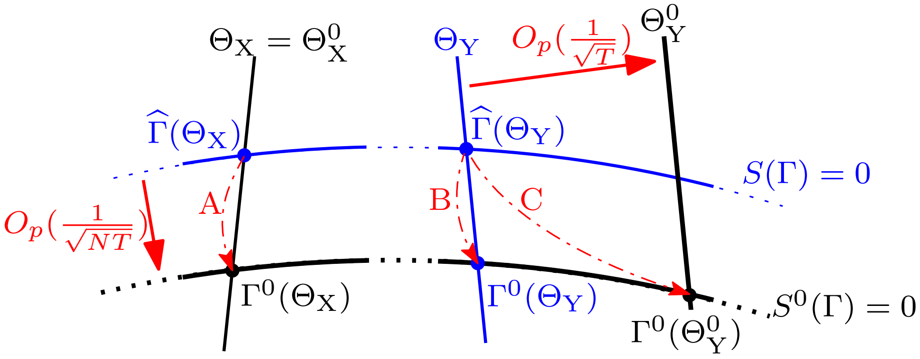
## 4.3 Which True Parameter?

This subsection clarifies some subtle issues involved in normalizing the true parameter. Let us first review how the points talked about so far are shown in Figure 3. Based on this visual representation, we explain the roles played by the different true parameters in asymptotic analysis.

Proposition 1 establishes that the set of optimizers (shown as the upper "$S(\Gamma)=\mathbf{0}$" curve) converges to the set of rotationally-equivalent true parameters (the lower "$S^0(\Gamma)=\mathbf{0}$" curve). Meanwhile, IF.1 says the identification conditions are also converging. The blue vertical line describes the normalization $\Theta_{\mathrm{Y}}$ for a specific sample.

Electronic copy available at: https://ssrn.com/abstract=2983919

Figure 3: Estimated and True Parameters in the Two Normalization Cases


$\Theta_{\rm X} = \Theta^0_{\rm X}$
$\Theta_{\rm Y}$
$O_p(\frac{1}{\sqrt{T}})$
$\Theta^0_{\rm Y}$
$\widehat{\Gamma}(\Theta_{\rm X})$
$\widehat{\Gamma}(\Theta_{\rm Y})$
$S(\Gamma) = 0$
$O_p(\frac{1}{\sqrt{NT}})$
A
B
C
$\Gamma^0(\Theta_{\rm X})$
$\Gamma^0(\Theta_{\rm Y})$
$\Gamma^0(\Theta^0_{\rm Y})$
$S^0(\Gamma) = 0$

*Notes:* The top horizontal curve is the set of optimizers. The lower curve is the set of rotationally equivalent true parameters. The three vertical lines are identification conditions. All the black objects are deterministic, the blue ones are sample-dependent. The two solid red arrows point from the random sets to their deterministic limits, labeled with the rates of convergence. The dashed red arrows A, B, C mark three specific estimation errors, whose asymptotic distributions are in Theorems 3.a, 3.b, 5, respectively.

It varies across samples but converges to the limit $\Theta^0_{\rm Y}$, shown as the vertical line on the right. On the other hand, the limit of $\Theta_{\rm X}$ is itself (since it is deterministic), shown as the vertical line on the left. The estimators $\widehat{\Gamma}(\Theta_{\rm X})$ and $\widehat{\Gamma}(\Theta_{\rm Y})$ are at the intersections of the optimizer set and the identification-condition set, where the parenthesis clarifies which identification condition it is in.

We make two main points. First, assumption IA means that the true $\Gamma^0$ under the two normalization cases $\Theta_{\rm X}$ and $\Theta_{\rm Y}$ are *not* the same. This can be seen by noting that the limiting identification conditions $\Theta^0$ are different for the two cases—one is $\Theta_{\rm X}$ itself and the other is $\Theta^0_{\rm Y}$. To avoid ambiguity, we write the true parameter under the two cases as $\Gamma^0(\Theta_{\rm X})$ and $\Gamma^0(\Theta^0_{\rm Y})$, while $\Gamma^0$ is reserved as the shorthand when there is no emphasis on the specific case. The two are rotationally equivalent, but satisfy different limiting identification conditions as clarified in the parentheses. The true parameter is deterministic, since it is the intersection of two deterministic sets: $S^0(\Gamma) = \mathbf{0}$ and $\Theta^0$. This deterministic truth is the parameter in IPCA definition (Eq. 2). Later on, it will serve as the fixed reference point in asymptotic expansion. Finally, the asymptotic variance is expressed with regard to the generic $\Gamma^0$, which is subsequently evaluated at either $\Gamma^0(\Theta_{\rm X})$ or $\Gamma^0(\Theta^0_{\rm Y})$ depending on the specific case.

Our second main point is that we measure estimation error against the true parameter normalized in the same way as the estimate is normalized. We call this the *normalized true* parameter and denote it as $\Gamma^0(\Theta)$, where $\Theta$ is the normalization used

Electronic copy available at: https://ssrn.com/abstract=2983919

for estimation.[27] Arrows A and B in Figure 3 are examples of this choice of estimation error. The two arrows represent the two estimation errors involved in our asymptotic expansions: $\widehat{\Gamma}(\Theta_{\mathrm{X}}) - \Gamma^0(\Theta_{\mathrm{X}})$ and $\widehat{\Gamma}(\Theta_{\mathrm{Y}}) - \Gamma^0(\Theta_{\mathrm{Y}})$. Each pair is an estimate minus a true, both under a common normalization.

A subtle distinction is that the normalized truth can be *made* sample-dependent by the normalization. This is the case of Arrow B in particular, where the reference point $\widehat{\Gamma}(\Theta_{\mathrm{Y}})$ is sample-dependent. On the other hand, in the case of $\Theta_{\mathrm{X}}$ there is no distinction for the normalized truth because $\Theta_{\mathrm{X}}$ and its limit are one and the same.

Measuring estimation errors against the normalized truth is well-justified, although perhaps peculiar at first sight that a "true" parameter can be random. Indeed, it follows and formalizes the convention of asymptotic analysis in factor analysis.[28] Following previous literature, our primary interest is on Arrows A and B, since they are essential about the accuracy of IPCA estimation in recovering a true parameter, albeit a rotated true parameter. In contrast, the asymptotic analysis of $\widehat{\Gamma}(\Theta_{\mathrm{Y}}) - \Gamma^0(\Theta^0_{\mathrm{Y}})$ (Arrow C) would reveal the complete asymptotic distribution of the estimator $\widehat{\Gamma}(\Theta_{\mathrm{Y}})$ since the reference point $\Gamma^0(\Theta^0_{\mathrm{Y}})$ is fixed. We will show that Arrow C converges at a slower rate than Arrows A or B due to an additional stochastic wedge. The wedge exists because the estimate is normalized according to $\Theta_{\mathrm{Y}}$ while the reference point is normalized by its limit $\Theta^0_{\mathrm{Y}}$.[29] This slows down the convergence of Arrow C, because it also relies on the convergence of $\Theta_{\mathrm{Y}}$ to $\Theta^0_{\mathrm{Y}}$ (which happens at the slower rate $\sqrt{T}$).

We derive these and other facts rigorously in the analysis below. We focus our asymptotic analysis on the errors labeled as Arrows A and B, and relegate theoretical and simulation results for Arrow C to Appendix B.

To recap, we have defined the following notations. The estimate is $\widehat{\Gamma}(\Theta)$, or $\widehat{\Gamma}$ in short when there is no emphasize on the specific normalization $\Theta$. The true parameter is $\Gamma^0(\Theta^0)$, or $\Gamma^0$ in short similarly. And, the normalized true parameter is $\widehat{\Gamma}(\Theta)$, against which the estimation errors are referenced. The instances of these

[27] Since $\Theta$ is an identification condition, per its definition, such a $\Gamma^0(\Theta)$ is unique.

[28] Bai and Ng (2002) and Bai (2003) are prominent examples that estimation errors are measured against the normalize true in the literature. They study PCA estimation error $\widehat{\lambda}_i - H^{-1}\lambda^0_i$, where $H$ is a *sample-dependent* rotation imposed on the true $\lambda^0_i$. (In their notation, $\lambda_i$ is factor loadings, equivalent to our $\beta_i$.) Although in Bai and Ng (2013), such rotation matrix $H$ is avoided, the true parameters used as comparison targets are still normalized according to sample realizations. For example, they directly restrict the *sample* second moment matrix of the true factors in their PC1 and PC2. This is comparable to our $\Theta_{\mathrm{Y}}$ condition [2].

[29] One cannot normalize the estimate by $\Theta^0_{\mathrm{Y}}$, simply because it is of population quantities. It is only a theoretical construction for large sample theory.

Electronic copy available at: https://ssrn.com/abstract=2983919

three objects in the two normalization examples can be found in Figure 3. Lastly, the same notations with a small $\gamma$ represent the vectorized variants.

### 4.4 Consistency of $\Gamma$ Estimation

Now we combine the previous preparations to show the consistency of $\Gamma$ estimation.

Per the estimator definition, $\widehat{\Gamma}$ solves the simultaneous equations $[S; I](\Gamma) = \mathbf{0}$.[30] Proposition 1 and IF.1 combined means $[S; I](\Gamma)$ uniformly converges to $[S^0; I^0](\Gamma)$. Proposition 2, IF.1, and IA together imply $\Gamma^0$ is the unique solution of $[S^0; I^0](\Gamma) = \mathbf{0}$. Since the solution of a uniformly converging function converges to the limit's *unique* solution (Newey and McFadden, 1994), we have $\widehat{\Gamma} \xrightarrow{p} \Gamma^0$. Meanwhile, $\Gamma^0(\Theta)$ is the solution of $[S^0; I](\Gamma)$, and we can similarly show it approaches the same fixed limit $\Gamma^0$. Bridged by the common limit $\Gamma^0$, the difference between the estimation and the normalized truth must converge.

**Theorem 1** ($\Gamma$ Estimation Consistency – Generic Normalization)**.** *Under Assumptions A–C, and if the identification condition $\Theta$ has an associated identification function $I(\Gamma)$ that satisfies IF.1–2, then as $N, T \to \infty$, $\widehat{\Gamma} - \Gamma^0(\Theta) \xrightarrow{p} \mathbf{0}$.*

The theorem above is the consistency result for a generic identification condition $\Theta$. For the two specific cases ($\Theta_{\mathrm{X}}$ and $\Theta_{\mathrm{Y}}$), we already verified that both identification functions $I_{\mathrm{X}}(\Gamma)$ and $I_{\mathrm{Y}}(\Gamma)$ satisfy Condition IF.1–2. Therefore, the estimators in these two specific cases are consistent as well.

**Corollary 1** ($\Gamma$ Estimation Consistency — Specific Cases)**.** *Under Assumptions A–C, IPCA $\widehat{\Gamma}$ estimated under normalization $\Theta_{\mathrm{X}}$ or $\Theta_{\mathrm{Y}}$ are consistent with respect to the accordingly normalized true parameters: as $N, T \to \infty$, $\widehat{\Gamma}(\Theta_{\mathrm{X}}) - \Gamma^0(\Theta_{\mathrm{X}}) \xrightarrow{p} \mathbf{0}$, and $\widehat{\Gamma}(\Theta_{\mathrm{Y}}) - \Gamma^0(\Theta_{\mathrm{Y}}) \xrightarrow{p} \mathbf{0}$.*

## 5 Asymptotic Distributions of $\Gamma$ Estimation Error

This section analyzes the asymptotic distribution of estimation error $\widehat{\gamma} - \gamma^0(\Theta)$. We first provide the general results for a generic identification condition $\Theta$, and then calculate the asymptotic distributions under the two specific normalization cases.

[30] $[S; I](\Gamma)$ means the vector-valued function by vertically stacking up $S(\Gamma)$ and $I(\Gamma)$. Same for a few other stacked functions below.

Electronic copy available at: https://ssrn.com/abstract=2983919

## 5.1 Asymptotic Error — Generic Normalization

**Theorem 2** (Asymptotic Error — Generic Normalization)**.** *Under the conditions of Theorem 1, and assuming both $S(\Gamma)$ and $I(\Gamma)$ are continuously differentiable in a neighborhood around $\gamma^0$, where $\gamma^0$ satisfy IA, as $N, T \to \infty$,*

$$\widehat{\gamma} - \gamma^0 = -\left(H^{0\top} H^0 + J^{0\top} J^0\right)^{-1}\left(H^{0\top} S(\gamma^0) + J^{0\top} I(\gamma^0)\right) + o_p\left(S(\gamma^0) + I(\gamma^0)\right), \tag{2.a}$$

$$\widehat{\gamma} - \gamma^0(\Theta) = -\left(H^{0\top} H^0 + J^{0\top} J^0\right)^{-1} H^{0\top} S(\gamma^0) + o_p\left(S(\gamma^0)\right), \tag{2.b}$$

*where* $H^0 := \left.\frac{\partial S^0(\Gamma)}{\partial \gamma^\top}\right|_{\gamma=\gamma^0}$ *and* $J^0 := \left.\frac{\partial I^0(\Gamma)}{\partial \gamma^\top}\right|_{\gamma=\gamma^0}$.

The proof is built on Newey and McFadden's (1994) analysis of a $M$-estimator by linearizing the score function. We extend that result to the not-locally-identified situation by appending the identification function to the score, and linearizing both symmetrically.[31] This analytical method, and the results it affords, could more broadly be used to analyze other estimators that require a normalization step.

The theorem offers a clear decomposition of the sources of randomness in parameter estimation. On the right-hand sides of 2.a and 2.b, the only random terms are the score and identification functions (both evaluated at $\Gamma^0$), representing the sample inaccuracies emanating from optimization and normalization. The theorem says $\widehat{\gamma} - \gamma^0$ loads on both sources of randomness while $\widehat{\gamma} - \gamma^0(\Theta)$ loads only on the randomness of normalization.

The intuition can be illustrated by Figure 3. Look at the small neighborhood around $\Gamma^0(\Theta^0_{\mathrm{Y}})$, and imagine both the "$S(\Gamma) = \mathbf{0}$" curve and the $\Theta_{\mathrm{Y}}$ curve are "wobbling" around their deterministic limits: they represent the two sources of randomness affecting estimation. Since the estimate $\widehat{\Gamma}(\Theta_{\mathrm{Y}})$ is the intersection of the two curves, it must load on the randomness of both (shown as Arrow C). However, since both $\widehat{\Gamma}(\Theta_{\mathrm{Y}})$ and $\Gamma^0(\Theta_{\mathrm{Y}})$ are under the same identification $\Theta_{\mathrm{Y}}$, their difference does not depend on the randomness of $\Theta_{\mathrm{Y}}$ (shown as Arrow B).

## 5.2 Asymptotic Error — Specific Cases

This section brings the general asymptotic result in Theorem 2.b to the specific cases denoted by Arrows A and B. For the reasons discussed in Section 4.3, asymptotic

[31] The proof that details our way of analysis is in Appendix C.10.

Electronic copy available at: https://ssrn.com/abstract=2983919

analysis of Arrow C, which is a special case of 2.a, is relegated to Appendix B.

To evaluate 2.b, it only remains to calculate the three right-hand side terms: $H^0$, $J^0$, and the asymptotic distribution of $S(\gamma^0)$. As we have discussed in Subsection 4.3, the general expressions in Theorem 2 need to be evaluated locally at either $\Gamma^0(\Theta_{\mathrm{X}})$ or $\Gamma^0(\Theta^0_{\mathrm{Y}})$. As mentioned before, we use use subscripts X and Y to denote the values calculated for their corresponding normalizations.

The Hessian-like matrices $H^0_{\mathrm{X}}$, $J^0_{\mathrm{X}}$ and $H^0_{\mathrm{Y}}$, $J^0_{\mathrm{Y}}$ are calculated by taking the derivatives of the limiting functions, which are detailed in Appendix C.12–C.13. The more interesting result is the asymptotic distribution of $S(\gamma^0)$. We know $S(\gamma^0) \xrightarrow{p} \mathbf{0}$ from Propositions 1 and 2. The following lemma says that the convergence happens at the rate of $\sqrt{NT}$ and gives the asymptotic distributions.

**Lemma 3** (Asymptotic Distribution of the Score Evaluated at $\Gamma^0$)**.**
*Under Assumptions A–F, as $N, T \to \infty$ such that $T/N \to 0$,*

$$\sqrt{NT}S\left(\gamma^0(\Theta_{\mathrm{X}})\right) \xrightarrow{d} \text{Normal}\left(\mathbf{0}, \mathbb{V}^{[1]}_{\mathrm{X}}\right), \quad \sqrt{NT}S\left(\Gamma^0(\Theta^0_{\mathrm{Y}})\right) \xrightarrow{d} \text{Normal}\left(\mathbf{0}, \mathbb{V}^{[1]}_{\mathrm{Y}}\right).$$ [32]

This lemma implies the convergence rate of $\Gamma$ estimation error is $\sqrt{NT}$, regardless of the normalization choice. The $\sqrt{NT}$-convergence rate highlights IPCA's ability to harness not just the time-series, but also the cross-sectional information. The ability ultimately comes from the assumption that the instrumental mapping is common across individuals. In contrast, without modeling the structure of the factor loadings, even were factors observed then the loading estimation could only rely on time-series information and thereby achieve convergence at rate $\sqrt{T}$.

From the perspective of the panel literature, the $\sqrt{NT}$ convergence rate can be understood by viewing $\Gamma$ as a common structural parameter and $f_t$ as the time fixed-effects. This view raises the concern that estimation could be asymptotically biased in the presence of "incidental parameters" whose number increases with the sample size (Neyman and Scott, 1948; Lancaster, 2000). Gagliardini and Gourieroux (2014) note the incidental parameter problem is much less pronounced in the case of time fixed-effects with large cross sections. We follow them and focus on the $T/N \to 0$ case, in which the asymptotic distribution is centered around zero and no asymptotic bias correction is needed. Should $T/N$ converge to a positive number, we conjecture the asymptotic distribution is still normal but with a non-zero mean—exact analysis

[32] The expressions of $\mathbb{V}^{[1]}_{\mathrm{X}}$ and $\mathbb{V}^{[1]}_{\mathrm{Y}}$ is with the proof in appendix C.11.

Electronic copy available at: https://ssrn.com/abstract=2983919

is left to future research. The simulation results we report below show no noticeable bias.

Finally, assembling the previous calculations into 2.b, we have the asymptotic distributions for Arrows A and B. The expressions are symmetric for the two cases and the specific values are numerically verified by Monte Carlo simulation in Section 7.

**Theorem 3** (Asymptotic Error — Specific Cases)**.** *Under Assumptions A–F, as $N, T \to \infty$ such that $T/N \to 0$,*

$$\sqrt{NT}\left(\widehat{\gamma}(\Theta_{\mathrm{X}}) - \gamma^0(\Theta_{\mathrm{X}})\right) \xrightarrow{d} -\left(H_{\mathrm{X}}^{0\top} H_{\mathrm{X}}^0 + J_{\mathrm{X}}^{0\top} J_{\mathrm{X}}^0\right)^{-1} H_{\mathrm{X}}^{0\top} \mathrm{Normal}\left(\mathbf{0}, \mathbb{V}_{\mathrm{X}}^{[1]}\right), \qquad (3.\mathrm{a})$$

$$\sqrt{NT}\left(\widehat{\gamma}(\Theta_{\mathrm{Y}}) - \gamma^0(\Theta_{\mathrm{Y}})\right) \xrightarrow{d} -\left(H_{\mathrm{Y}}^{0\top} H_{\mathrm{Y}}^0 + J_{\mathrm{Y}}^{0\top} J_{\mathrm{Y}}^0\right)^{-1} H_{\mathrm{Y}}^{0\top} \mathrm{Normal}\left(\mathbf{0}, \mathbb{V}_{\mathrm{Y}}^{[1]}\right). \qquad (3.\mathrm{b})$$

# 6 Asymptotic Analysis of Factor Estimation

We have so far treated factor estimation $\widehat{f}_t(\widehat{\Gamma})$ implicitly, because we concentrated-out $f_t$ in the target function (Eq. 7). Now, with the asymptotics of $\widehat{\Gamma}$ in hand, we come back to factor estimation and lay out its asymptotic distribution with relative ease.

We measure estimation error against the normalized true factor for the reasons discussed in 4.3. Given $\Gamma^0(\Theta)$ is the normalized true $\Gamma$, the correspondingly normalized true factor is $\widetilde{f}_t\left(\Gamma^0(\Theta)\right)$.[33]

This factor estimation error can be decomposed into two sources following the discussion around Eq. (10),

$$\widehat{f}_t(\widehat{\Gamma}) - \widetilde{f}_t\left(\Gamma^0(\Theta)\right) = \left(\widetilde{f}_t(\widehat{\Gamma}) - \widetilde{f}_t\left(\Gamma^0(\Theta)\right)\right) + \left(\widehat{\Gamma}^\top C_t^\top C_t \widehat{\Gamma}\right)^{-1} \widehat{\Gamma}^\top C_t^\top \widetilde{e}_t(\widehat{\Gamma}). \qquad (11)$$

The first term captures the part of factor estimation error only due to the inaccuracy of $\Gamma$ estimation. The previous section concluded that $\widehat{\Gamma}$ and $\Gamma^0(\Theta)$ converge at the rate of $\sqrt{NT}$. Hence, the first term has the same rate of convergence. The second term captures the remaining inaccuracy from the cross-sectional regression's finite sample. That is, even were $\Gamma^0(\Theta)$ known, the second term still would dominate at

[33] Given a (sample-dependent) $R$, such that the normalized true $\Gamma$ is represented as $\Gamma^0(\Theta) = \Gamma^0 R$, the correspondingly normalized true factor is inversely rotated as $R^{-1} f_t^0 = R^{-1}\widetilde{f}_t\left(\Gamma^0\right) = \widetilde{f}_t\left(\Gamma^0 R\right) = \widetilde{f}_t\left(\Gamma^0(\Theta)\right)$. In other words, the two pairs, $\left\{\Gamma^0, f_t^0\right\}$ and $\left\{\Gamma^0(\Theta), \widetilde{f}_t\left(\Gamma^0(\Theta)\right)\right\}$, are rotationally equivalent to each other.

Electronic copy available at: https://ssrn.com/abstract=2983919

rate $\sqrt{N}$. Based on this decomposition, we arrive at the follow results about the convergence of factor estimation error.

**Theorem 4** (Factor Estimation Asymptotics)**.**
*Under Assumptions A–F, with an identification* $\Theta$ *that satisfies IF.1–2,*
(4.a) *Factor estimation is consistent: as* $N, T \to \infty$, $\widehat{f}_t(\widehat{\Gamma}) - \widetilde{f}_t\left(\Gamma^0(\Theta)\right) \xrightarrow{p} \mathbf{0}$.
(4.b) *Factor estimation error centered against the normalized true factor converges to a normal distribution at the rate of* $\sqrt{N}$*: as* $N, T \to \infty$, $\forall t$,

$$\sqrt{N}\left(\widehat{f}_t(\widehat{\Gamma}) - \widetilde{f}_t\left(\Gamma^0(\Theta)\right)\right) \xrightarrow{d} \text{Normal}\left(\mathbf{0}, \mathbb{V}_t^{[2]}\right).$$ [34]

The theorem gives the factor's asymptotic distribution under a generic normalization $\Theta$. Similar to $\Gamma$ estimation, we can evaluate $\mathbb{V}_t^{[2]}$ at either the $\Gamma^0(\Theta_{\mathrm{X}})$ or $\Gamma^0(\Theta_{\mathrm{Y}}^0)$ normalizations.[35]

# 7 Simulations

This section presents a concise set of simulations that illustrate the behavior of the IPCA estimation in finite samples, and assess the accuracy of approximation based on the asymptotic theory derived above. To summarize, we find that estimation errors are well-approximated with a normal distribution. This is true even in rather small samples, and when the true generating process has errors with large variance. This shows that applied researchers can confidently assume normality for confidence intervals and hypothesis tests when applying IPCA, as it verifies the asymptotic derivations in the previous section. We present details below.

For given $N, T$, we generate a stochastic panel of $c, f^0, e$ and use these to assemble the $x$ panel. We calibrate the simulated data to the IPCA model (fixing $K = 2$ and $L = 10$) estimated from US monthly stock returns in KPS.[36]

Simulations proceed according to the following steps:

[34] The expression of the asymptotic variance $\mathbb{V}_t^{[2]}$ is with the proof in Appendix C.14.

[35] Appendix B.3 contains the asymptotics when centered by the original $f_t^0$. This situation corresponds to Arrow C for $\Gamma$ estimation, and the the additional stochastic wedge introduced by a sample-based normalization affects the asymptotic distribution.

[36] In particular, we use the ten most significant instruments from KPS as calibration targets. They are market capitalization, total assets, market beta, short-term reversal, momentum, turnover, price relative to its 52-week high, long-term reversal, unexplained volume, and idiosyncratic volatility with respect to the Fama-French three factor model.

Electronic copy available at: https://ssrn.com/abstract=2983919

1. **Generate factors**. Fit a VAR(1) process to estimated IPCA factors from KPS. Simulate $f_t^0$ according to the estimated VAR employing normal innovations.

2. **Generate instruments**. For each stock, calculate the time-series averages of the instruments. Pool the demeaned characteristics into a panel and estimate a ten variable panel VAR(1). Next, for each $i$, generate the means of $c_{i,t}$ as an i.i.d. draw from the empirical distribution of the time series means. Then, simulate the dynamic component of $c_{i,t}$ from the estimated VAR with normal innovations.[37]

3. **Generate errors**. Elements of the error panel $e$ are simulated from an i.i.d. normal distribution whose variance is calibrated so that the population $R^2$, defined as $1-\mathbb{E}e^2/\mathbb{E}x^2$, equals 20%, matching the empirical $R^2$ in the estimated model in KPS.

4. **Generate main panel**. We fix $\widehat{\Gamma}(\Theta_{\mathrm{Y}})$ at its empirically estimated value and calculate $x_{i,t}$ according to model equation (3).

We produce 200 simulated sample panels of dimension $N = 200, T = 200$. For each simulated sample, we estimate $\Gamma$ under the two identification conditions $\Theta_{\mathrm{X}}$ and $\Theta_{\mathrm{Y}}$. Figure 4 reports the histograms of the estimation errors and overlay them with the theoretical distributions for comparison. Given the data generating process, the theoretical distributions of estimation errors are approximated by the asymptotic distributions given in Theorem 3.[38]

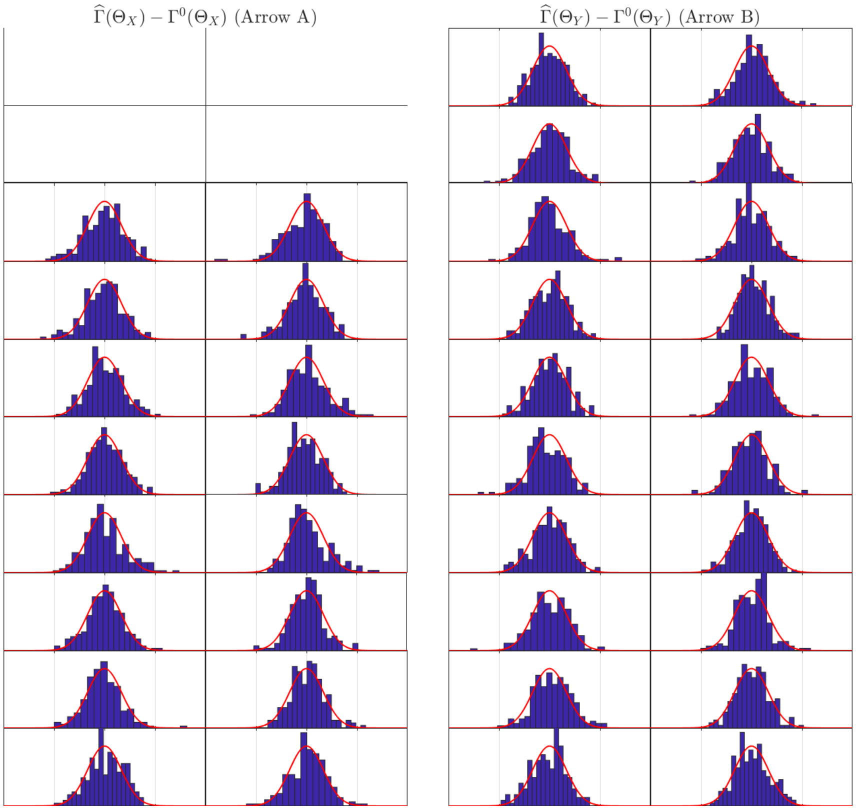
The first panel reports estimation error $\widehat{\Gamma}(\Theta_{\mathrm{X}}) - \Gamma^0(\Theta_{\mathrm{X}})$. It corresponds to Theorem 3.a, or Arrow A in Figure 3. The four entries at the top are empty, because the corresponding four entries of $\Gamma$ are pinned down by the identification condition and do not need to be estimated. The second panel corresponds to Theorem 3.b, or Arrow B. It reports $\widehat{\Gamma}(\Theta_{\mathrm{Y}}) - \Gamma^0(\Theta_{\mathrm{Y}})$. Note that while there are $K^2$ more distributions presented on the right, the two normalizations have exactly the same degrees of

[37] We generate $c_{i,t}$ with ex-ante i.i.d. means, so that each individual's time-series process is non-ergodic. This is deliberate so that $c_{i,t}$ admits the flexible property allowed by stochastic panels and resembles real-world instrument data.

[38] The required population moments, $\Omega^{cc}, \Omega^{cef}, \mathbb{V}^{[3]}$ etc., are calculated by large sample Monte Carlo. In the process of calculating these population quantities via Monte Carlo, we rotate the data generating process ex ante according to the required identification assumption IA. For example, $f_t^0$ needs to be inversely rotated when $\Gamma^0$ is normalized from $\Gamma^0(\Theta_{\mathrm{Y}}^0)$ to $\Gamma^0(\Theta_{\mathrm{X}})$, resulting in a different value for $\Omega^{cef}$.

Electronic copy available at: https://ssrn.com/abstract=2983919

Figure 4: $\Gamma$ Estimation Errors — Simulated v.s. Asymptotic Approximation



*Note:* This figure reports the small sample distributions of $\Gamma$ estimation errors under the two example normalization cases. We conduct 200 simulations with sample dimensions $N = 200, T = 200$. The left panel reports the distribution of $\widehat{\Gamma}(\Theta_{\mathrm{X}}) - \Gamma^0(\Theta_{\mathrm{X}})$ (Arrow A). The right panel reports the distribution of $\widehat{\Gamma}(\Theta_{\mathrm{Y}}) - \Gamma^0(\Theta_{\mathrm{Y}})$ (Arrow B). Each subplot corresponds to one element in the $L \times K$ $(10 \times 2)$ matrix of $\Gamma$. Each histogram plots the simulated estimation errors, which is overlaid with the asymptotic distributions from Theorem 3 as finite sample approximations. The horizontal axis range is ±6 theoretical standard deviations, the tick marks are at ±3 theoretical standard deviations. The vertical axes are probability density for the bell curves or frequency density for the bars.

freedom. In other words, the top four distributions on the right duplicate information in the distributions plotted below them.

In all cases, the simulated distributions are centered around zero and match the theoretical distributions well. For some entries, we observe some skewness and tail heaviness. These are due to the small sample size and relatively large error variance in the generating process. In untabulated simulations with $N = 1000, T = 1000$, the asymptotics more-fully kick in and the distributions become visually indistinguishable from a normal. Hence, the simulation results suggest that even with panels of only

Electronic copy available at: https://ssrn.com/abstract=2983919

moderate size, our asymptotic approximations are accurate.

# 8 Applications

This section describes two empirical applications of IPCA to demonstrate its broad usefulness for analyzing economic data. The first is an application to international macroeconomics, where IPCA makes it easy analyze many nations' evolving relationships to global business cycles using country-level instruments. The second application builds on KPS and uses IPCA to analyze a dynamic model of asset risk and expected returns.

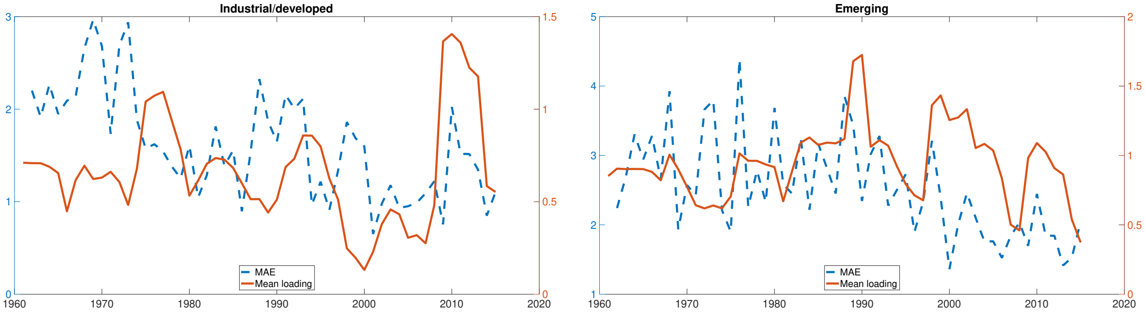
## 8.1 International Macroeconomics

Country-level macroeconomic fluctuations are globally connected (Backus et al., 1992). Using a static state-space model, Gregory et al. (1997) use maximum likelihood to document this for the G7 countries. Likewise, Kose et al. (2003) use a static state-space model estimated with Bayesian methods to disentangle global from regional and country-specific growth factors for a panel of countries over 30 years. Recently Kose et al. (2012) (henceforth KOP) used data from the World Development Indicators and estimated their static state-space model for a panel of countries before and after 1985 in order to analyze the convergence or decoupling of global business cycles. In essence, they ask: have countries' relationships to global growth changed as the countries themselves have evolved? This question is ideally suited for investigation with IPCA.

We use IPCA to analyze the global factor structure in GDP growth using data from the World Development Indicators database. We include as many countries as possible from the "industrial/developed" and "emerging" country groups studied by KOP. After reasonable data filters on indicators and countries, which we detail in the Appendix E, we are left with 45 countries. Within this sample there are nine variables that are available for most countries, so we use these as our instruments. The first two instruments are the import and export share of GDP—natural indicators of a country's economic connectedness with the rest of the world. Next we use the proportion of GDP relative to world GDP to measure the nation's relative size. We account for dynamics in capital intensity using gross capital formation, and we use popula-

Electronic copy available at: https://ssrn.com/abstract=2983919

Figure 5: MAE and Average Loading



*Notes:* The left axis shows the IPCA model mean absolute error in units of percentage annual growth and corresponds to the blue dotted line. The right axis corresponds to the equally weighted cross-sectional mean factor loading in each group, shown in red.

tion growth to account for growth in the labor force. To account for recent economic growth and risks, we include the 5-year rolling mean and volatility of the nation's GDP growth and its rate of inflation. Finally, we include a constant characteristic. Next, we double the set of instruments to 18 by interacting the nine variable above with an indicator for whether a country is in industrial/developed group. Our annual data run from 1961 to 2015 so that $T = 55$, and we demean the growth rates following KOP. About 91% of the 5,280 possible country-year observations are non-missing.

We study latent factor models with $K = 1$ and compare IPCA to the static-loading PCA estimator. We find a panel $R^2$ from the IPCA model of 32%, capturing roughly triple the variation in demeaned country growth explained by KOP. The $R^2$ from PCA is 22%, or two-thirds that achieved by IPCA. When making a head-to-head comparison of PCA and IPCA it is important to keep in mind two major differences between the estimators. The first is their stark difference in parameterization. IPCA achieves its fit using only 18 parameters to estimate its loadings, or 60% fewer parameters than the 45 used by PCA.[39] Second, IPCA accommodates dynamics in each country's loading on the global growth factor while PCA estimates static loadings. If countries converge or decouple as they evolve, IPCA's dynamic loadings are capable of detecting this. PCA's static loadings, on the other hand, cannot detect such dynamics and will instead try to fit an evolving system with a static model, and this type of misspecification is difficult to diagnose.

Results from IPCA show that beta dynamics are indeed critical to understanding

[39] These parameter counts are net of the 45 country-specific means used to demean the data in both IPCA and PCA.

Electronic copy available at: https://ssrn.com/abstract=2983919

Table 1: Global Growth Model Estimates

| | | | |
|---|---|---|---|
| GDP | -0.27* | Ind×GDP | 0.44* |
| Capital Formation | -0.09 | Ind×Capital Formation | 0.21 |
| Exports | 0.45 | Ind×Exports | -0.48 |
| Imports | -0.34 | Ind×Imports | 0.32 |
| Inflation | 0.23* | Ind×Inflation | -0.05 |
| Pop. Growth | -0.31 | Ind×Pop. Growth | 0.17 |
| Growth Vol. | 0.79* | Ind×Growth Vol. | 0.06 |
| Mean Growth | 0.00 | Ind×Mean Growth | -0.19* |
| Constant | -0.60* | Ind×Constant | 0.69* |

*Notes:* Estimated $\Gamma$ coefficients scaled by the panel standard deviation of each instrument (the exceptions are Constant and Ind×Constant which are unscaled). The instruments are the log of GDP as a fraction of world GDP, gross capital formation, export and import share of GDP, inflation, population growth, 5-year rolling mean and volatility of GDP growth, and a constant. Each instrument is also interacted with an indicator for inclusion in the "industrial/developed" country group. An asterisk denotes statistical significance at the 10% level or better (using a bootstrap test following KPS).

the global business cycle. Figure 5 shows the time series of loadings on the global growth factor broken out by industrial/developed economies and emerging economies. For readability, we report the equally weighted cross-sectional average loading within each group of countries.

We see substantial variation in global growth sensitivity in each group. This is underpinned by an interesting state dependence in loadings—they rise sharply in economic downturns. While this is visually evident in the plot for industrial/developed countries, the precise nature of state dependence in loadings can be read from the estimated $\Gamma$ matrix, which is shown in Table 1. To make estimates more interpretable, we scale each element of $\Gamma$ to describe the effect on factor loadings from a one standard deviation increase in the associated instrument.

First, the constant and its interaction with the industrial dummy show that loadings in the industrial/developed group are significantly higher than those in emerging economies. But the largest and perhaps most interesting finding is the role of growth volatility in describing state dependence in global growth sensitivity. A well documented pattern in the business cycle literature is the spike in growth volatility associated with recessions (Bloom, 2014).Table 1 shows that such rises in volatility are accompanied by concomitant rises in sensitivity to the global growth factor. It also shows that the dependence of growth sensitivity is one of the few instruments

Electronic copy available at: https://ssrn.com/abstract=2983919

that plays a similar role in both developed and emerging countries. For the emerging group, a one standard deviation increase in growth volatility associates with an increased loading of 0.79 on the global growth factor, versus 0.85 for the industrial group (the difference is insignificant). The only other instrument that the two groups agree on is inflation. We see that higher inflation associates with high global sensitivity, that this effect is slightly stronger in emerging economies (estimate of 0.23), but that the difference versus developed economies is insignificant. For the remaining instruments, our estimates show significant differences in the drivers of global exposure. Emerging economies are especially sensitive to global fluctuations if they have high exports, low imports, are relatively small, have high inflation, and have low population growth. In industrial/developed group, the effects of most instruments other than volatility net to nearly zero. One other significant effect in industrial/developed countries is that global sensitivity rises when recent growth has been low (based on the significant coefficient of $-0.19$ on recent mean growth). This compounds the jump in global sensitivity associated with a rise in volatility, because recent growth volatility and recent mean growth are negatively correlated.

Finally, we see some broad evidence of global convergence from the dotted lines shown in Figure 5. These describe the mean absolute error (MAE) from the IPCA model each year. They show that over time the global growth factor has become increasingly successful at describing the full panel of growth rates. This is evident from the downward trend in MAE among both industrial and emerging economies. Overall, IPCA results illustrate an important role for dynamic loadings in global business cycle models and show that IPCA's ability to accommodate such dynamics ultimately delivers a more accurate description of the data than leading alternatives.

## 8.2 Asset Pricing

KPS apply IPCA to describe systematic risk and associated risk premium (cost of capital) of US stocks, where systematic risk is defined as dynamic loadings on latent factors that are instrumented by stock characteristics. Here, we expand upon their asset pricing context and use IPCA to calculate systematic risk and cost of capital for *newly listed* firms that the model has not seen before. This is motivated by the challenging question of how to value private firms and if it is possible to use information in publicly traded equity prices for this purpose. Like in the macroeconomic example,

Electronic copy available at: https://ssrn.com/abstract=2983919

IPCA is ideally suited to address this question because the model parameterizes risk and cost of capital as a function of firm characteristics. IPCA finds the mapping between the return behavior of traded firms and their characteristics, which can then be extrapolated to non-traded firms to approximate their as-if traded value. Note that this is not possible with standard empirical asset pricing methodologies, which require the history of publicly traded prices for individual assets to infer their future risk premia (and thus cost of capital).

Our data consists of over 2.9 million stock-month observations of excess stock returns ($x$) and 93 associated firm characteristics ($z$) from 1965-2018.[40] We use lagged firm characteristics to instrument for the conditional systematic risk loadings ($c_{i,t} = z_{i,t-1}$).[41] The out-of-sample evaluation is performed for newly listed firms, defined as the first 12 months following a firm's initial public offering (IPO). In our data, 21,275 firms have an IPO, comprising 249,414 out-of-sample stock-month observations. In-sample estimation is based on the complementary 2.6 million observations of incumbent stocks that excludes those in the test sample of new listings.

We calculate the total $R^2$ and predictive $R^2$, as defined in KPS, to evaluate the estimates of stocks' systematic risk and expect returns, respectively. The procedure starts by first estimating $\widehat{\Gamma}$ and $\{\widehat{f}_t\}$ within the in-sample data set of incumbent stocks. Then, the estimates are brought to the data of new listings to calculate fitted values as

$$\widehat{x}_{i,t+1}^{\text{Tot}} := z_{i,t}\widehat{\Gamma}\widehat{f}_{t+1}, \qquad \widehat{x}_{i,t+1}^{\text{Pred}} := z_{i,t}\widehat{\Gamma}\widehat{\lambda}$$

for all $i, t$ in the out-of-sample data set. The first term $\widehat{x}_{i,t+1}^{\text{Tot}}$ reconstructs the realized return as the factor model's fitted value (that is, using the factor realization, $\widehat{f}_{t+1}$). The second term is a prediction of the new listing return and replaces the factor realization with the factor's estimated mean $\widehat{\lambda}$, directly following KPS.[42] Based on these fits, we calculate the out-of-sample total and predictive $R^2$ as the explained variation in $x_{i,t+1}$ due to $\widehat{x}_{i,t+1}^{\text{Tot}}$ and $\widehat{x}_{i,t+1}^{\text{Pred}}$, respectively.

Table 2 reports the results for the IPCA model (with $K = 4$, as advocated by KPS). The close similarity of the in-sample and out-of-sample total $R^2$ indicates the

[40] The dataset is from Gu et al. (2020). Firm characteristics are transformed into ranks on the interval $[-0.5, 0.5]$ as in KPS. Any missing characteristic is assigned the value 0, which is replacement with the cross-sectional mean/median.

[41] The information content of unexpected idiosyncratic return shocks is formalized by Assumption A.

[42] $\widehat{\lambda}$ is calculated as the in-sample time series mean of $\widehat{f}_{t+1}$.

Electronic copy available at: https://ssrn.com/abstract=2983919

Table 2: Explained Variation of Stock Returns

| | Total $R^2$ | Predictive $R^2$ |
|---|---|---|
| Incumbent stocks (in-sample) | 15.66 | 0.25 |
| New listings (out-of-sample) | 13.44 | 0.22 |

*Notes:* $R^2$ in percentage. Based on IPCA with $K = 4$ for the 1965-2018 US stock-month panel.

same characteristics that determine the systematic riskiness of incumbent stocks also determine the riskiness of new listings. In other words, once we condition on firm characteristics, the model finds a highly similar description for the common variation among returns on newly listed stocks compared to the common variation in returns on incumbents.

While the total $R^2$ is indicative of the model's ability to describe systematic risks of new listings, the predictive $R^2$ summarizes the model's description of their expected returns (or, in equivalent terms, cost of capital or discount rates). That is, the predictive $R^2$ is especially informative about the usefulness of the model for asset valuation. The close similarity of the in-sample and out-of-sample predictive $R^2$ indicates that the IPCA model is as effective at "pricing" new listings as it is for pricing incumbent stocks. The most important takeaway is that the model does this without using the individual return history of the new listings (which is of course unavailable and the crux of the research question), but manages to price them nonetheless by extrapolating what it learns from data on incumbent stocks.[43]

# 9 Conclusion

This paper has introduced a new approach of modeling and estimating the latent factor structure of panel data, called Instrumented Principal Component Analysis (IPCA). The key innovation is using additional panel data to instrument for the dynamic factor loadings. Mainly, each individual's time-varying factor loading is related to instrumental data according to a common and constant mapping.

Estimating this mapping, rather than the factor loadings directly, has many econometric advantages compared to other latent variable estimators like PCA. On one

[43] This performance is even more remarkable when we recognize that new firms' stock returns are more variable than incumbents.

Electronic copy available at: https://ssrn.com/abstract=2983919

hand, the mapping's parameterization tends to be more parsimonious. Its degrees of freedom are fixed and not increasing with the size of the cross section, which improves the rate of convergence and tends to avoid over-fit problems in high-dimensional applications. At the same time, IPCA brings broad possibilities of economic discovery by harnessing the wealth of additional panel information as well as relying on economic theories that indicate relationships between observable covariates and factor exposures. These advantages are exemplified with two applications to equity returns and international macroeconomics respectively.

Our main theoretical contribution is to derive the statistical properties of the IPCA estimator. We show consistency and asymptotic normality under general data generating processes. We emphasize generality with two main theoretical innovations. First, the fundamental construction of stochastic panels establishes a collection of large panel asymptotic results under general conditions without resorting to higher-level assumptions. Second, our method deals with the well-known rotational unidentification issue in latent factor analysis in general terms rather than under specific normalization assumptions. We show how the choice of normalization affects the rate of convergence and the asymptotic distribution. This identification framework is applicable to estimators other than IPCA that also require additional normalization.

Electronic copy available at: https://ssrn.com/abstract=2983919

# References

Acemoglu, D. and Azar, P. D. (2020). Endogenous Production Networks. *Econometrica*, 88(1):33–82. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA15899.

Backus, D. K., Kehoe, P. J., and Kydland, F. E. (1992). International Real Business Cycles. *Journal of Political Economy*, 100(4):745–775. Publisher: University of Chicago Press.

Bai, J. (2003). Inferential Theory for Factor Models of Large Dimensions. *Econometrica*, 71(1):135–171.

Bai, J. and Ng, S. (2002). Determining the Number of Factors in Approximate Factor Models. *Econometrica*, 70(1):191–221.

Bai, J. and Ng, S. (2013). Principal components estimation and identification of static factors. *Journal of Econometrics*, 176(1):18–29.

Büchner, M. and Kelly, B. (2020). A Factor Model for Option Returns. *Yale University Working Paper*.

Bloom, N. (2014). Fluctuations in Uncertainty. *Journal of Economic Perspectives*, 28(2):153–176.

Del Negro, M. and Otrok, C. (2008). Dynamic factor models with time-varying parameters: measuring changes in international business cycles. *FRB of New York Staff Report*, (326).

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56.

Fan, J., Liao, Y., and Wang, W. (2016). Projected principal component analysis in factor models. *The Annals of Statistics*, 44(1):219–254.

Gagliardini, P. and Gourieroux, C. (2014). EFFICIENCY IN LARGE DYNAMIC PANEL MODELS WITH COMMON FACTORS. *Econometric Theory*, 30(5):961–1020. Publisher: Cambridge University Press.

Electronic copy available at: https://ssrn.com/abstract=2983919

Gagliardini, P., Ossola, E., and Scaillet, O. (2016). Time-Varying Risk Premium in Large Cross-Sectional Equity Data Sets. *Econometrica*, 84(3):985–1046.

Geweke, J. (1977). The dynamic factor analysis of economic time series. *Latent variables in socio-economic models*.

Gregory, A., Head, A., and Raynauld, J. (1997). Measuring world business cycles. *International Economic Review*, 38(3):677–701.

Gu, S., Kelly, B., and Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies*, 33(5):2223–2273. Publisher: Oxford Academic.

Hansen, L. P. and Sargent, T. J. (2013). Risk, Uncertainty, and Value.

Kelly, B. T., Pruitt, S., and Su, Y. (2019). Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics*, 134(3):501–524.

Kose, A., Otrok, C., and Prasad, E. (2012). Global business cycles: convergence or decoupling? *International Economic Review*, 53(2):511–538.

Kose, M. A., Otrok, C., and Whiteman, C. H. (2003). International Business Cycles: World, Region, and Country-Specific Factors. *American Economic Review*, 93(4):1216–1239.

Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of Econometrics*, 95(2):391–413.

Newey, W. K. and McFadden, D. (1994). Chapter 36 Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, volume 4, pages 2111–2245. Elsevier.

Neyman, J. and Scott, E. L. (1948). Consistent Estimates Based on Partially Consistent Observations. *Econometrica*, 16(1):1–32. Publisher: [Wiley, Econometric Society].

Primiceri, G. E. (2005). Time Varying Structural Vector Autoregressions and Monetary Policy. *The Review of Economic Studies*, 72(3):821–852.

Pruitt, S. (2012). Uncertainty Over Models and Data: The Rise and Fall of American Inflation. *Journal of Money, Credit and Banking*, 44(2-3):341–365.

Electronic copy available at: https://ssrn.com/abstract=2983919

Sargent, T. J. and Sims, C. A. (1977). Business cycle modeling without pretending to have too much a priori economic theory. *New methods in business cycle research*, 1:145–168.

Stock, J. H. and Watson, M. W. (2002). Forecasting Using Principal Components from a Large Number of Predictors. *Journal of the American Statistical Association*, 97(460):1167–1179.

Su, L. and Wang, X. (2017). On time-varying factor models: Estimation and testing. *Journal of Econometrics*, 198(1):84–101.

Electronic copy available at: https://ssrn.com/abstract=2983919

# Appendix

## A Notation Details

Matrix Operations: $\text{vect}\,(A)$ vectorizes matrix $A$ to a column by going right first, then down. Throughout the paper, $\gamma$ is always $\text{vect}\,(\Gamma)$, same for all the other decorated $\widehat{\Gamma}, \Gamma^0(\Theta)$, etc. Related, $\text{veca}\,(A)$ stacks $A$'s upper triangle entries, including the diagonal, in a vector by going right first, then down; $\text{vecb}\,(A)$ stacks $A$'s upper triangle entries, *not* including the diagonal, in a vector by going right first, then down. For example, let $A = \Big[1, 2, 3; 4, 5, 6; 7, 8, 9\Big]$, then $\text{vect}\,(A) = [1, 2, 3, 4, 5, 6, 7, 8, 9]^\top$, $\text{veca}\,(A) = [1, 2, 3, 5, 6, 9]^\top$, $\text{vecb}\,(A) = [2, 3, 6]^\top$.

$[A; B]$ means the two matrices vertically stacked.

$\|\cdot\|$ is the Euclidean norm of a vector, or the Frobenius norm if the input is a matrix. $\|\cdot\|^2$ is the square of $\|\cdot\|$, or the sum of squares across all entries of the vector or matrix.

Summations like $\sum_i, \sum_t$ are always in the range of the panel sample: $i = 1$ to $N$, $t = 1$ to $T$, without writing out the details.

## B Results About Arrow C

### B.1 Asymptotic Distribution of $\widehat{\Gamma}(\Theta_\text{Y})$

This subsection works out the asymptotic distribution of $\widehat{\Gamma}(\Theta_\text{Y}) - \Gamma^0(\Theta^0_\text{Y})$, or Arrow C in Figure 3. Since the reference point $\Gamma^0(\Theta^0_\text{Y})$ is deterministic, this analysis provides the complete asymptotic distribution of the estimator $\widehat{\Gamma}(\Theta_\text{Y})$. In contrast to Arrow B (studied in Section 5.2), Arrow C also depends on the randomness of the identification $\Theta_\text{Y}$. We rely on the general result 2.a for the asymptotic distribution. Subsection 5.2 has already calculated three of the four right-hand side inputs: $H^0_\text{Y}, J^0_\text{Y}$ and the asymptotic distribution of $S\left(\Gamma^0(\Theta^0_\text{Y})\right)$. The fourth piece that needs to be analyzed is the asymptotic distribution $I_\text{Y}\left(\Gamma^0(\Theta^0_\text{Y})\right)$.

Identification function $I_\text{Y}(\Gamma)$, as defined in 4.2 is stacked up by two parts. The top $\frac{1}{2}K(K+1)$ entries are from ortho-normalizing $\Gamma$, which is irrelevant of sample information. Therefore, similar to the $\Theta_\text{X}$ case, the top part of $I_\text{Y}\left(\Gamma^0(\Theta^0_\text{Y})\right)$ is still deterministic. The bottom $\frac{1}{2}K(K-1)$ rows of $I_\text{Y}(\Gamma)$ are from diagonalizing the sample

Electronic copy available at: https://ssrn.com/abstract=2983919

second moment matrix of $\widehat{f}_t$, a process that introduces a rotational contamination with convergence rate $\sqrt{T}$. To see why it is this rate, suppose the true $f_t^0$ is observed, which follows a diagonal second moment matrix *in population*. However, its sample second moment would not be exactly diagonal, off by a sample error that is in the order of $1/\sqrt{T}$. Imposing the sample normalization requires rotating $f_t^0$ slightly just to offset that error.

**Lemma 4** (Asymptotic distribution of $I_{\mathrm{Y}}(\Gamma^0)$)**.** *Under Assumptions A–F,*

$$\sqrt{T} I_{\mathrm{Y}}\left(\Gamma^0(\Theta_{\mathrm{Y}}^0)\right) \xrightarrow{d} \text{Normal}\left(\mathbf{0}_{K^2\times 1}, \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbb{V}^{[3]} \end{bmatrix}\right),$$

*where $\mathbb{V}^{[3]}$ is specified in Assumption D(3).*[44]

Lemma 4 implies the "contamination effect" that the normalization step introduces to $\widehat{\Gamma}(\Theta_{\mathrm{Y}})$ converges at the rate of $\sqrt{T}$. Surprisingly, this is even slower than that of the core optimization step, which converges at $\sqrt{NT}$ according Lemma 3. Hence it solely shows up in the asymptotic distribution of $\widehat{\Gamma}(\Theta_{\mathrm{Y}})$ as the dominant term. In summary, the convergence rate of Arrow C is $\sqrt{T}$, while those of Arrows A and B are both $\sqrt{NT}$.

**Theorem 5** (Asymptotic Distribution of Arrow C)**.** *Under Assumptions A–F, as $N, T \to \infty$,*

$$\sqrt{T}\left(\widehat{\gamma}(\Theta_{\mathrm{Y}}) - \gamma^0(\Theta_{\mathrm{Y}}^0)\right) \xrightarrow{d} -\left(H_{\mathrm{Y}}^{0\top} H_{\mathrm{Y}}^0 + J_{\mathrm{Y}}^{0\top} J_{\mathrm{Y}}^0\right)^{-1} J_{\mathrm{Y}}^{0\top} \text{Normal}\left(\mathbf{0}, \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbb{V}^{[3]} \end{bmatrix}\right).$$

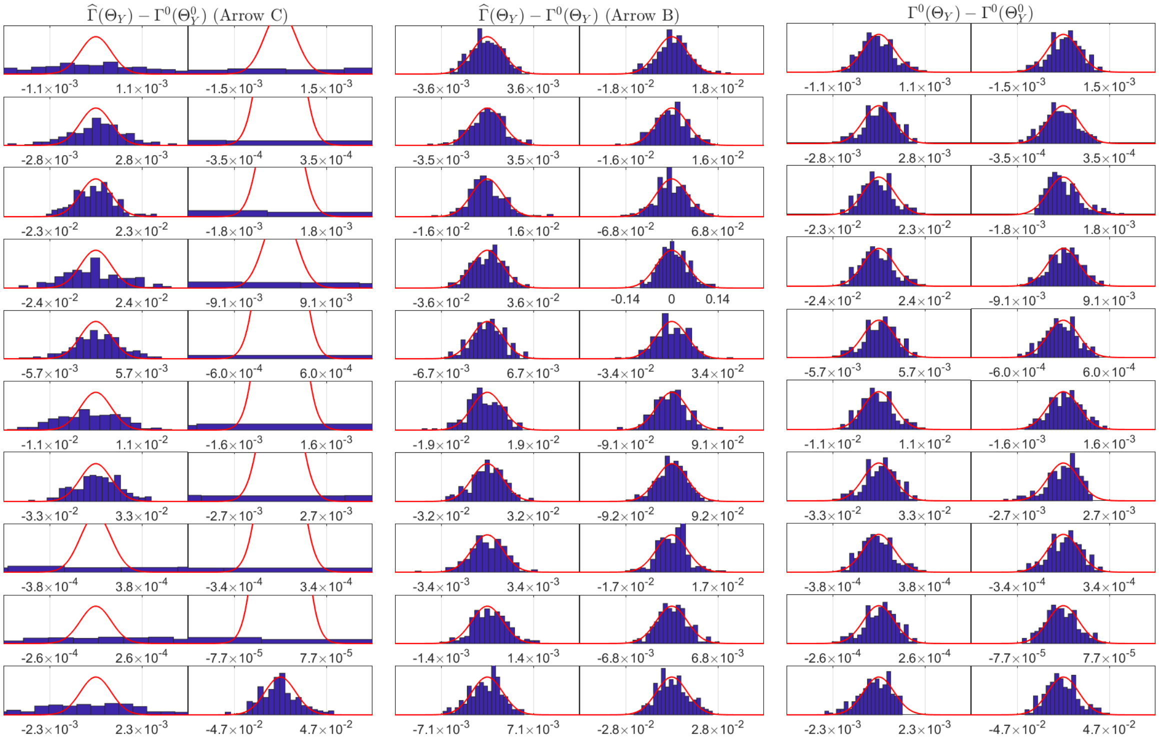
## B.2 Simulation Results about Arrow C

Figure 6 presents the simulation results about Arrow C. The results are from the same simulations from Section 7. The first panel reports the simulated error of Arrow C and compare that with the theoretical ones from Theorem 5. Except for some entries that roughly match, the simulated small-sample distributions are wider than the theoretical distributions by orders of magnitude. This is purely a finite sample phenomenon, the asymptotics have no problem.

[44] In the variance matrix, the $\mathbf{0}$ block in the top left is $\frac{1}{2}K(K+1) \times \frac{1}{2}K(K+1)$, top right is $\frac{1}{2}K(K+1) \times \frac{1}{2}K(K-1)$, lower left is $\frac{1}{2}K(K-1) \times \frac{1}{2}K(K+1)$. Together with $\mathbb{V}^{[3]}$ of $\frac{1}{2}K(K-1) \times \frac{1}{2}K(K-1)$, they make up the variance matrix of $K^2 \times K^2$.

Electronic copy available at: https://ssrn.com/abstract=2983919

Figure 6: Γ Estimation Errors — Simulated v.s. Asymptotic Approximation



*Note:* This figure is constructed with the same simulation exercises in Figure 4, and also follows the same format. The simulated distribution of a new set of random variables (written at the top of each panel) are reported. The corresponding theoretical distributions are from Theorems 5, 3.b, and 5 respectively. Additionally, the x-axes ticks are marked with numbers for comparison. Again, the x-axes range is ±6 theoretical standard deviations, and the tick numbers mark ±3 theoretical standard deviations.

To diagnose the issue, we decompose Arrow C as $\widehat{\Gamma}(\Theta_{\mathrm{Y}}) - \Gamma^0(\Theta_{\mathrm{Y}}^0) = (\widehat{\Gamma}(\Theta_{\mathrm{Y}}) - \Gamma^0(\Theta_{\mathrm{Y}})) + (\Gamma^0(\Theta_{\mathrm{Y}}) - \Gamma^0(\Theta_{\mathrm{Y}}^0)) = \mathcal{O}_p\left(1/\sqrt{NT}\right) + \mathcal{O}_p\left(1/\sqrt{T}\right)$, and respectively plot the two terms in the second and third panels. Although the first term (arrow B) is asymptotically small, its level, as reported in the second panel, is still large at $N = 200, T = 200$.[45] As a result, the first term messes up the small sample distribution of Arrow C, which theoretically is only driven by the second term. In the third panel, we ignore the first term and only look at the dominant second term. It conforms to the asymptotic theory from Theorem 5 well. In this sense, the third panel verifies the derivation of large sample asymptotics of the theorem.

So $N, T$-orders aside, why the asymptotic variance of the first term (Arrow B) is

[45] The second panel of Figure 6 is a copy of the second panel of Figure 4, except we now include the x-axis labels for comparing theoretical variances.

Electronic copy available at: https://ssrn.com/abstract=2983919

much larger than the second? Retracing the analyses, Arrow B hinges on the accuracy of the first order condition, which eventually depends on the $N, T$-CLT of $c^\top e f^{0\top}$ (Assumption D(1)). While the second depends on the accuracy of the identification condition, which is about the time-series CLT of $f^0 f^{0\top}$ only (Assumption D(3)). So the relative levels of the variances of $e$ to $f^0$ is critical here, which can be roughly measured and calibrated to empirical data by $R^2$. As a check, scaling down the variance of $e$ to an unrealistic level makes the first panel also converge well (detailed plot omitted).

### B.3 Factor Estimation Error Measured Against $\{f_t^0\}$

Section 6 analyzed the factor estimation error against the normalized true $\widetilde{f}_t\left(\Gamma^0(\Theta_{\mathrm{Y}})\right)$. Now we measured factor estimation against $f_t^0$. We know $f_t^0 = \widetilde{f}_t\left(\Gamma^0\right)$. The analogue of Eq. (11) is

$$
\begin{aligned}
&\widehat{f}_t\left(\widehat{\Gamma}(\Theta_{\mathrm{Y}})\right) - f_t^0 \\
&= \left(\widetilde{f}_t\left(\widehat{\Gamma}(\Theta_{\mathrm{Y}})\right) - \widetilde{f}_t\left(\Gamma^0(\Theta_{\mathrm{Y}}^0)\right)\right) + \left(\widehat{\Gamma}(\Theta_{\mathrm{Y}})^\top C_t^\top C_t \widehat{\Gamma}(\Theta_{\mathrm{Y}})\right)^{-1} \widehat{\Gamma}^\top C_t^\top \widetilde{e}_t\left(\widehat{\Gamma}(\Theta_{\mathrm{Y}})\right).
\end{aligned}
$$

The only difference happens at the first term, which is now driven by the estimation error between $\widehat{\Gamma}(\Theta_{\mathrm{Y}})$ and $\Gamma^0(\Theta_{\mathrm{Y}}^0)$ (Arrow C), which has a convergence rate of $\sqrt{T}$ according to Subsection B.1. This implies the additional stochastic wedge introduced by a sample-based normalization might no longer be dominated by the second term. The following Theorem says $\widehat{f}_t\left(\widehat{\Gamma}(\Theta_{\mathrm{Y}})\right)$ and $f_t^0$ converge the slower of $\sqrt{N}$ or $\sqrt{T}$.

**Theorem 6** (Factor Estimation Error Measured Against $\{f_t^0\}$)**.** *Under Assumptions A–F, as $N, T \to \infty$,*

$$
\widehat{f}_t\left(\widehat{\Gamma}(\Theta_{\mathrm{Y}})\right) - f_t^0 = \mathcal{O}_p\left(\max\left\{1/\sqrt{N}, 1/\sqrt{T}\right\}\right).
$$

# C Proofs of All Theoretical Results

## C.1 Proof of Lemma 1

*Proof.* This is a direct application of Birkhoff Law of Large Numbers (Hansen and Sargent, 2013, Theorem 2.5.1). Notice in the two right-hand sides, the dot "$\cdot$" can take

Electronic copy available at: https://ssrn.com/abstract=2983919

any natural number without changing the value of the conditional expectation. □

## C.2 Lemma 5 and its Proof

The joint ergodicity condition SP.4 first implies the single-subscript random variables are ergodic, and hence their averages converge to the unconditional expectations:

**Lemma 5** (Single-subscript LLN)**.** *Under Conditions SP.1–4, let* $X^{[\mathrm{d}]}$ *be* $\mathfrak{F}^{[\mathrm{d}]}$*-measurable (single-subscript) random vectors, and if* $\mathbb{E}\left\|X^{[\mathrm{d}]}\right\|^2 < \infty$*, then*

$$\frac{1}{N}\sum_{i=1}^{N} X_i^{[\mathrm{cs}]} \xrightarrow{L^2} \mathbb{E}X^{[\mathrm{cs}]},\ \textit{as } N\to\infty \quad \textit{and} \quad \frac{1}{T}\sum_{t=1}^{T} X_t^{[\mathrm{ts}]} \xrightarrow{L^2} \mathbb{E}X^{[\mathrm{ts}]},\ \textit{as } T\to\infty.$$

*Proof.* Apply Lemma 1, we have $\frac{1}{N}\sum_{i=1}^{N} X_i^{[\mathrm{cs}]} \xrightarrow{L^2} \mathbb{E}\left[X^{[\mathrm{cs}]}\middle|\mathfrak{F}^{[\mathrm{ts}]}\right]$. It remains to show $X_i^{[\mathrm{cs}]}$ is ergodic. We know $X^{[\mathrm{cs}]}$ is measured by $\mathfrak{F}^{[\mathrm{cs}]}$, within which every event is $\mathbb{S}_{[\mathrm{ts}]}$-invariant. So an $\mathbb{S}_{[\mathrm{cs}]}$-invariant event within $\mathfrak{F}^{[\mathrm{cs}]}$ must be invariant to both transformations. By condition SP.4, it has probability either 0 or 1. That is to say, $\left\{X_i^{[\mathrm{cs}]}\right\}$ is an ergodic stochastic process (in the traditional one-directional sense). That implies $\mathbb{E}\left[X^{[\mathrm{cs}]}\middle|\mathfrak{F}^{[\mathrm{ts}]}\right] = \mathbb{E}X^{[\mathrm{cs}]}$ w.p. 1, which in turn gives the desired m.s. convergence.

The other direction is symmetric. □

## C.3 Proof of Lemma 2

Combining Lemmas 1 and 5 yields a result that the panel-wise average can recover the population moment when both $N$ and $T$ are large.

*Proof.*

$$\frac{1}{NT}\sum_t\sum_i X_{i,t} = \frac{1}{T}\sum_t\left(\frac{1}{N}\sum_i X_{i,t} - \mathbb{E}\left[X_{\cdot,t}\middle|\mathfrak{F}^{[\mathrm{ts}]}\right]\right) + \frac{1}{T}\sum_t \mathbb{E}\left[X_{\cdot,t}\middle|\mathfrak{F}^{[\mathrm{ts}]}\right] \tag{12}$$

We are going to show first term $\xrightarrow{L^1} \mathbf{0}$ ($N\to\infty$, uniformly across $T$), second term $\xrightarrow{L^1} \mathbb{E}X$, as $N, T\to\infty$ ($T\to\infty$, uniformly across $N$).

Electronic copy available at: https://ssrn.com/abstract=2983919

For any $T$, apply $\lim_{N\to\infty} \mathbb{E} \|\cdot\|$ to the first term

$$\lim_{N\to\infty} \mathbb{E} \left\| \frac{1}{T} \sum_t \left( \frac{1}{N} \sum_i X_{i,t} - \mathbb{E}\left[ X_{\cdot,t} \middle| \mathfrak{F}^{[\text{ts}]} \right] \right) \right\| \tag{13}$$

$$\leq \lim_{N\to\infty} \frac{1}{T} \sum_t \mathbb{E} \left\| \frac{1}{N} \sum_i X_{i,t} - \mathbb{E}\left[ X_{\cdot,t} \middle| \mathfrak{F}^{[\text{ts}]} \right] \right\| \tag{14}$$

$$= \lim_{N\to\infty} \mathbb{E} \left\| \frac{1}{N} \sum_i X_{i,t} - \mathbb{E}\left[ X_{\cdot,t} \middle| \mathfrak{F}^{[\text{ts}]} \right] \right\| = 0. \tag{15}$$

We applied a triangular inequality, stationarity by condition SP.3, and Lemma 1 (notice $L^2$ convergence implies $L^1$ convergence), in that order.

For the second term, notice the summand $\mathbb{E}\left[ X_{\cdot,t} \middle| \mathfrak{F}^{[\text{ts}]} \right]$ is $\mathfrak{F}^{[\text{ts}]}$-measurable. Therefore, Lemma 5 implies $L^2$ convergence which in turn implies $L^1$ convergence towards the unconditional expectation. □

## C.4 Verify $\Theta_{\text{Y}}$ is an identification condition

**Lemma 6.** *Set $\Theta_{\text{Y}}$ is an identification condition. I.e., for any $\Gamma \in \Psi$ there is a unique $\Gamma' \in \Theta_{\text{Y}}$ that is rotationally equivalent to $\Gamma$.*

*Proof.* We proof by first constructing the normalization $\Gamma'$ first, and then show it is unique. Before that, we notice a relationship:

$$\frac{1}{T} \sum_t \widehat{f}_t \left( \Gamma R \right) \widehat{f}_t^{\top} \left( \Gamma R \right) = R^{-1} \frac{1}{T} \sum_t \widehat{f}_t(\Gamma) \widehat{f}_t^{\top}(\Gamma) R^{-1\top} \tag{16}$$

**normalization procedure:** Starting from a $\Gamma$, we look for a $\Gamma R \in \Theta_{\text{Y}}$.

1. Find Cholesky decomposition: $[\text{Chol}]^{\top} [\text{Chol}] = \Gamma^{\top}\Gamma$.
2. Calculate: $[\text{OldV}] = \frac{1}{T} \sum_t \widehat{f}_t(\Gamma) \widehat{f}_t^{\top}(\Gamma)$.
3. Find Eigen-decomposition: $[\text{Chol}] [\text{OldV}] [\text{Chol}]^{\top} = [\text{Orth}] [\text{Diag}] [\text{Orth}]^{\top}$ such that $[\text{Diag}]$ is has descending diagonal entries.
4. Finally, we find the normalization as $\Gamma' = \Gamma [\text{Chol}]^{-1} [\text{Orth}]$.

(5) In addition, it might be required to align the signs of the $K$ columns of $\Gamma'$ following the discussion in Appendix D.

To verify the normalization procedure is correct, we need to check $\Gamma' \in \Theta_{\text{Y}}$. We verify that: $\Gamma'^{\top}\Gamma' = [\text{Orth}]^{\top} [\text{Chol}]^{\top} \Gamma^{\top}\Gamma [\text{Chol}]^{-1} [\text{Orth}] = \mathbb{I}_K$. And, based on Eq. 16,

Electronic copy available at: https://ssrn.com/abstract=2983919

we verify that:

$$\frac{1}{T}\sum_{t}\widehat{f}_t\left(\Gamma'\right)\widehat{f}_t^{\top}\left(\Gamma'\right)=[\text{Orth}]^{-1}[\text{Chol}][\text{OldV}][\text{Chol}]^{\top}[\text{Orth}]^{-1\,\top}=[\text{Diag}] \tag{17}$$

**normalization is unique:**

We need to show such an $\Gamma'$ is unique. Since $\Gamma' = \Gamma R$, we just need to show such an $R$ is unique. We inspect the restrictions of $\Theta_{\text{Y}}$ and shrink the set of possible $R$ down to singularity, following the above normalization procedure:

First, given $\Theta_{\text{Y}}$ restriction [1], $R^{\top}\Gamma^{\top}\Gamma R = \mathbb{I}_K$. The possible $R$ must have decomposition: $R = [\text{Chol}]^{-1}[\text{Orth}]$, where [Orth] can only be ortho-normal matrices ($[\text{Orth}]^{\top}[\text{Orth}] = \mathbb{I}$). Plug this decomposition into $\Theta_{\text{Y}}$ restriction [2], we need an [Orth] that also satisfies $[\text{Orth}]^{-1}[\text{Chol}][\text{OldV}][\text{Chol}]^{\top}[\text{Orth}]^{-1\,\top} = [\text{Diag}]$. We have found setting [Orth] as the eigen-decomposition satisfies. Importantly, when we restrict the eigen-decomposition with distinct and decreasing eigen-values such an eigen-decomposition is unique. □

## C.5 Some Lemmas for Uniform Large-$N,T$ Convergence

Let $x_{N,t}(\Gamma)$ represent a sequence (indexed by $N$) of stochastic process functions of $\Gamma$ with finite dimensions. Define several large $N$ limiting conditions that it can be subject to. These conditions are all *uniformly* across $\Gamma$, and prefixed with "U-".

$$\lim_{N\to\infty}\mathbb{E}\left[\sup_{\Gamma\in\Psi}\left\|x_{N,t}(\Gamma)\right\|\right]=0 \qquad \text{(U-mean converging)}$$

$$\lim_{N\to\infty}\mathbb{E}\left[\sup_{\Gamma\in\Psi}\left\|x_{N,t}(\Gamma)\right\|^2\right]=0 \qquad \text{(U-mean square converging)}$$

$$\exists M, N^*, \text{s.t.} \qquad \mathbb{E}\left[\sup_{\Gamma\in\Psi}\left\|x_{N,t}(\Gamma)\right\|^2\right]<M \qquad \forall N>N^*$$

$$\text{(U-mean square bounded)}$$

$$\exists M, N^*, \text{s.t.} \qquad Pr\left\{\sup_{\Gamma}\left\|x_{N,t}(\Gamma)\right\|<M\right\}=1, \qquad \forall N>N^* \qquad \text{(U-a.s. bounded)}$$

Next, Lemma 7 is the upgrade version of Lemma 2 to deal with "uniform across $\Gamma$".

**Lemma 7.** *If $x_{N,t}(\Gamma)$ is stationary in $t$ and U-mean converging, then its time-series average is converging in the large-$N$ probability limit uniformly for any $T$. That is,*

Electronic copy available at: https://ssrn.com/abstract=2983919

$\forall \epsilon, \delta > 0, \exists N^{[1]}$ *s.t.* $\forall T$ *and* $\forall N > N^{[1]}$

$$Pr\left\{\sup_{\Gamma\in\Psi}\left\|\frac{1}{T}\sum_t x_{N,t}(\Gamma)\right\| > \epsilon\right\} < \delta. \tag{18}$$

*Proof.* Start by an inequality exchanging the order of sup and $\sum$:

$$\sup_{\Gamma\in\Psi}\left\|\frac{1}{T}\sum_t x_{N,t}(\Gamma)\right\| \leq \sup_{\Gamma\in\Psi}\frac{1}{T}\sum_t \|x_{N,t}(\Gamma)\| \leq \frac{1}{T}\sum_t \sup_{\Gamma\in\Psi}\|x_{N,t}(\Gamma)\|. \tag{19}$$

Apply expectation on both sides, and by stationarity, $\forall T$

$$\mathbb{E}\sup_{\Gamma\in\Psi}\left\|\frac{1}{T}\sum_t x_{N,t}(\Gamma)\right\| \leq \mathbb{E}\frac{1}{T}\sum_t \sup_{\Gamma\in\Psi}\|x_{N,t}(\Gamma)\| = \mathbb{E}\sup_{\Gamma\in\Psi}\|x_{N,t}(\Gamma)\|. \tag{20}$$

The last term $\mathbb{E}\sup_{\Gamma\in\Psi}\|x_{N,t}(\Gamma)\|$ is irrelevant of $T$, and it converges to zero as $N \to \infty$, according to the precondition about U-mean converging. Hence, the first term is $T$-uniform large $N$-convergence: $\forall \epsilon > 0, \exists N^{[1]}$ s.t. $\forall T$ and $\forall N > N^{[1]}$

$$\mathbb{E}\sup_{\Gamma\in\Psi}\left\|\frac{1}{T}\sum_t x_{N,t}(\Gamma)\right\| < \epsilon. \tag{21}$$

Therefore, by Chebyshev's inequality, we can make conclusion statement about $T$-uniform $N$-convergence in probability. □

Lemma 7 is critical to establish large $N, T$ simultaneous convergence. Notice in the conclusion statement of the Lemma, $N^{[1]}$ does not depend on $T$ but only on $\epsilon, \delta$. Hence, this statement implies large $N, T$ simultaneous convergence, a property that will be used in the proof of Proposition 1 later.

Lemma 7 also shows U-mean converging is important since it is the necessary condition for large-$N, T$ simultaneous convergence. The next lemma gives some calculation rules to reach a U-mean converging sequence.

**Lemma 8.** *If* $x^{[1]}_{N,t}(\Gamma)$ *is U-mean square converging,* $x^{[2]}_{N,t}(\Gamma)$ *is U-mean square bounded, and* $x^{[3]}_{N,t}(\Gamma)$ *is U-a.s. bounded, then*

1. $x^{[1]}_{N,t}(\Gamma)x^{[2]}_{N,t}(\Gamma)$ *is U-mean converging, which implies*

2. $x^{[1]}_{N,t}(\Gamma)$ *by itself is also U-mean converging.*

Electronic copy available at: https://ssrn.com/abstract=2983919

*3.* $x^{[1]}_{N,t}(\Gamma)x^{[3]}_{N,t}(\Gamma)$ *is still U-mean square converging.*

*4.* $x^{[2]}_{N,t}(\Gamma)x^{[3]}_{N,t}(\Gamma)$ *is still U-mean square bounded.*

*Proof.* 1. For each $\omega$, we have

$$\sup_{\Gamma} \|x_{N,t}(\Gamma)\| = \sup_{\Gamma} \left\|x^{[1]}_{N,t}(\Gamma)x^{[2]}_{N,t}(\Gamma)\right\| \leq \sup_{\Gamma} \left\|x^{[1]}_{N,t}(\Gamma)\right\| \sup_{\Gamma} \left\|x^{[2]}_{N,t}(\Gamma)\right\|. \tag{22}$$

So, put inside expectation:

$$\begin{aligned} \mathbb{E}\sup_{\Gamma} \|x_{N,t}(\Gamma)\| &\leq \mathbb{E}\left[\sup_{\Gamma} \left\|x^{[1]}_{N,t}(\Gamma)\right\| \sup_{\Gamma} \left\|x^{[2]}_{N,t}(\Gamma)\right\|\right] && (23) \\ &\leq \left(\mathbb{E}\left[\sup_{\Gamma} \left\|x^{[1]}_{N,t}(\Gamma)\right\|^2\right] \mathbb{E}\left[\sup_{\Gamma} \left\|x^{[2]}_{N,t}(\Gamma)\right\|^2\right]\right)^{1/2} && (24) \end{aligned}$$

by Cauchy-Schwarz inequality. Then it is easy to wrap up the proof with deterministic limit analysis. Namely the product of a sequence converging to zero and a bounded sequence is also converging to zero, and the square root of a converging sequence converges to the square root.

2. Trivial.

3. By a matrix version of the Cauchy-Schwarz inequality:

$$\|x_{N,t}(\Gamma)\|^2 = \left\|x^{[1]}_{N,t}(\Gamma)x^{[3]}_{N,t}(\Gamma)\right\|^2 \leq \left\|x^{[1]}_{N,t}(\Gamma)\right\|^2 \left\|x^{[3]}_{N,t}(\Gamma)\right\|^2 \tag{25}$$

Apply $E\sup_{\Gamma}$ on both sides:

$$\begin{aligned} \mathbb{E}\sup_{\Gamma} \|x_{N,t}(\Gamma)\|^2 &\leq \mathbb{E}\sup_{\Gamma} \left\|x^{[1]}_{N,t}(\Gamma)\right\|^2 \left\|x^{[3]}_{N,t}(\Gamma)\right\|^2 && (26) \\ &\leq \mathbb{E}\left[\sup_{\Gamma} \left\|x^{[1]}_{N,t}(\Gamma)\right\|^2 \sup_{\Gamma} \left\|x^{[3]}_{N,t}(\Gamma)\right\|^2\right] && (27) \end{aligned}$$

For any $\omega$, if $\sup_{\Gamma\in\Psi} \left\|x^{[3]}_{N,t}(\Gamma)\right\| < M$, then $\sup_{\Gamma\in\Psi} \left\|x^{[3]}_{N,t}(\Gamma)\right\|^2 < M^2$. That means $\left\|x^{[3]}_{N,t}(\Gamma)\right\|^2$ is also U-a.s. bounded for large enough $N$'s. Plug the bound, $M^2$, back into the expectation calculation for any finite $N$ we had above:

$$\mathbb{E}\sup_{\Gamma} \|x_{N,t}(\Gamma)\|^2 \leq \mathbb{E}\left[\sup_{\Gamma} \left\|x^{[1]}_{N,t}(\Gamma)\right\|^2\right] M^2 \tag{28}$$

Electronic copy available at: https://ssrn.com/abstract=2983919

Take large-$N$ limits on both sides:

$$\lim_N \mathbb{E} \sup_\Gamma \|x_{N,t}(\Gamma)\|^2 \leq \lim_N \mathbb{E} \left[ \sup_\Gamma \left\| x_{N,t}^{[1]}(\Gamma) \right\|^2 \right] M^2 = 0. \tag{29}$$

4. Almost the same the as the previous proof. Just change $x_{N,t}^{[1]}(\Gamma)$ to $x_{N,t}^{[2]}(\Gamma)$ everywhere until the last three lines. In the last three lines, just change "$\lim_N = 0$" to "$\limsup_N < \infty$". □

Lemma 9 builds the necessary conditions for U-mean square converging from low level conditions for the primitives for example Assumption A. It is closely related to Lemma 1.

**Lemma 9.** *If $x_{i,t}$ is a stochastic panel with zero time-series conditional expectation with bounded unconditional second moment:*

$$\mathbb{E}\left[x \middle| \mathfrak{F}^{[\text{ts}]}\right] = \mathbf{0} \text{ and } \mathbb{E}\|x\|^2 < +\infty, \tag{30}$$

*then its cross-sectional average $\frac{1}{N}\sum_i x_{i,t}$ is U-mean square converging.* [46]

*Proof.* Apply Lemma 1 to $x$:

$$\frac{1}{N}\sum_{i=1}^{N} X_{i,t} \xrightarrow{L^2} \mathbf{0}, \forall t. \tag{31}$$

Because $x$ is irrelevant of $\Gamma$, it is easy to see the mean squared convergence result above implies U-mean square converging.

□

## C.6 Proposition 1

### C.6.1 Complete the statement of Proposition 1

We first state the complete version of the Proposition 1 with the addition of an intermediate large-$N$ result and writing the expressions of the probability limits. .

[46] notice here $\Gamma$ does not enter in the random function.

Electronic copy available at: https://ssrn.com/abstract=2983919

**Proposition 1** (Uniform Convergence of the Score Function)**.**
*Under Assumptions A–C, the score function converges uniformly in probability:*

$$\sup_{\Gamma\in\Psi} \|S(\Gamma) - S_T(\Gamma)\| \xrightarrow{p} 0, \qquad N \to \infty, \forall T, \tag{32}$$

$$\sup_{\Gamma\in\Psi} \left\|S(\Gamma) - S^0(\Gamma)\right\| \xrightarrow{p} 0, \qquad N, T \to \infty, \tag{33}$$

*where*

$$S_T(\Gamma) := \frac{1}{T}\sum_t \text{vect}\left(\Omega_t^{cc}\Pi_{t(\Gamma)} f_t^0 \widetilde{f}_t^\top(\Gamma)\right), \tag{34}$$

$$S^0(\Gamma) := \mathbb{E}\ \text{vect}\left(\Omega_t^{cc}\Pi_{t(\Gamma)} f_t^0 \widetilde{f}_t^\top(\Gamma)\right), \tag{35}$$

$$\Pi_{t(\Gamma)} := \left(\mathbb{I}_L - \Gamma\left(\Gamma^\top\Omega_t^{cc}\Gamma\right)^{-1}\Gamma^\top\Omega_t^{cc}\right)\Gamma^0. \tag{36}$$

### C.6.2 Preparations for the Proof of Proposition 1

The proof of Proposition 1 is quite involved. The first part manipulates the expression of the score function to a form consisting of primitive random variables. Second, Lemma 10 deals with the cross-section convergence. Then, Lemma 11 deals with the time-series dimension. In the final step, the results are put together to finish the proof of for the large $N, T$ convergence in Proposition 1.

The first part of proof manipulates the expression of the score function to a form consisting of primitive random variables. Then, we analyze its uniform convergence. We already have

$$S(\Gamma) = \frac{1}{NT}\sum_t \left(C_t^\top \otimes \widehat{f}_{t(\Gamma)}\right)\left(x_t - C_t\Gamma\widehat{f}_{t(\Gamma)}\right) = \frac{1}{NT}\sum_t \text{vect}\left(C_t^\top \widehat{e}_{t(\Gamma)}\widehat{f}_t^\top(\Gamma)\right), \tag{37}$$

and

$$\widehat{f}_{t(\Gamma)} = \widetilde{f}_{t(\Gamma)} + \left(\Gamma^\top C_t^\top C_t\Gamma\right)^{-1}\Gamma^\top C_t^\top \widetilde{e}_{t(\Gamma)}. \tag{38}$$

Population OLS error:

$$\widetilde{e}_{t(\Gamma)} = x_t - C_t\Gamma\widetilde{f}_{t(\Gamma)} = e_t + C_t\Pi_{t(\Gamma)} f_t^0,$$

with the shorthand $\Pi_{t(\Gamma)} := \left(\mathbb{I}_L - \Gamma\left(\Gamma^\top\Omega_t^{cc}\Gamma\right)^{-1}\Gamma^\top\Omega_t^{cc}\right)\Gamma^0$. Combined with Eq.

Electronic copy available at: https://ssrn.com/abstract=2983919

(10), we have

$$\begin{aligned}\widehat{e}_{t(\Gamma)} &= e_t + C_t\Pi_{t(\Gamma)}f_t^0 - C_t\Gamma\left(\Gamma^\top C_t^\top C_t\Gamma\right)^{-1}\Gamma^\top C_t^\top \widetilde{e}_{t(\Gamma)}\\ \widehat{f}_{t(\Gamma)} &= \widetilde{f}_{t(\Gamma)} + \left(\Gamma^\top C_t^\top C_t\Gamma\right)^{-1}\Gamma^\top C_t^\top \widetilde{e}_{t(\Gamma)}\end{aligned}\tag{39}$$

Plug those back to the score. Each summand in equation (37) yields $3\times 2 = 6$ terms:

$$\begin{aligned} & C_t^\top \widehat{e}_{t(\Gamma)}\widehat{f}_t^\top{}_{(\Gamma)} \\ =& C_t^\top e_t && \widetilde{f}_t^\top{}_{(\Gamma)} \\ +& C_t^\top C_t\Pi_{t(\Gamma)}f_t^0 && \widetilde{f}_t^\top{}_{(\Gamma)} \\ -& C_t^\top C_t\Gamma\left(\Gamma^\top C_t^\top C_t\Gamma\right)^{-1}\Gamma^\top C_t^\top\widetilde{e}_{t(\Gamma)} && \widetilde{f}_t^\top{}_{(\Gamma)} \\ +& C_t^\top e_t && \widetilde{e}_t^\top{}_{(\Gamma)}C_t\Gamma\left(\Gamma^\top C_t^\top C_t\Gamma\right)^{-1} \\ +& C_t^\top C_t\Pi_{t(\Gamma)}f_t^0 && \widetilde{e}_t^\top{}_{(\Gamma)}C_t\Gamma\left(\Gamma^\top C_t^\top C_t\Gamma\right)^{-1} \\ -& C_t^\top C_t\Gamma\left(\Gamma^\top C_t^\top C_t\Gamma\right)^{-1}\Gamma^\top C_t^\top\widetilde{e}_{t(\Gamma)} && \widetilde{e}_t^\top{}_{(\Gamma)}C_t\Gamma\left(\Gamma^\top C_t^\top C_t\Gamma\right)^{-1}. \end{aligned}\tag{40}$$

Call the six terms $S_t^{[1]}(\Gamma)$ to $S_t^{[6]}(\Gamma)$, so that

$$S(\Gamma) = \frac{1}{NT}\sum_t \text{vect}\left(S_t^{[1]}(\Gamma) + \cdots + S_t^{[6]}(\Gamma)\right).\tag{41}$$

Before jumping into the rigorous proof, we give a loose description of the rationale. Given the expression of score function in Eq. (41), Proposition 1 states the score function's uniform probability limit. First, taking $N \to \infty$ at a fixed $t$, we have the three modular results, $\frac{1}{N}C_t^\top e_t \to \mathbf{0}$, $\frac{1}{N}\Gamma C_t^\top \widetilde{e}_{t(\Gamma)} \to \mathbf{0}$, and $\frac{1}{N}C_t^\top C_t = \mathcal{O}_p(1)$, by cross-sectional LLN. Plugging these into the score expression (40), we find that $\frac{1}{N}S_t^{[p]}(\Gamma) \to \mathbf{0}$ for $p = 1, 3, 4, 5, 6$. The second term is an exception as the only one not involving an $\widetilde{e}_{t(\Gamma)}$ or $e_t$ error term. It correspond to the first source of $\widehat{f}_t$ decomposition purely from a "wrong" $\Gamma$ rather than the finite sample. We see $S_t^{[2]}$ increases with $N$. We have $\frac{1}{N}S_t^{[2]}(\Gamma) \to \Omega_t^{cc}\Pi_{t(\Gamma)}f_t^0\widetilde{f}_t^\top{}_{(\Gamma)}$. The cross-sectional limit is $\mathfrak{F}^{[\text{ts}]}$ measurable (Lemma 1). Taking a finite time-series average yields the finite-$T$ large-$N$ convergence of the score given by Eq. (32).[47] Finally, taking $T \to \infty$ delivers the score's convergence to the unconditional expectation, as Eq. (33) report.

The proof, to a large extent, follows the steps of proving Lemmas 5 and 2, with

[47] This result could be used to construct finite-$T$ large-$N$ inference – we leave that for future work.

Electronic copy available at: https://ssrn.com/abstract=2983919

the complication of uniform convergence across $\Gamma$, which is necessary for solution convergence proved further below.

**Lemma 10** (Large $N$ Cross-sectional Convergence at each $t$)**.**

$$\frac{1}{N}\sum_i \left(S_t^{[1]}(\Gamma) + \cdots + S_t^{[6]}(\Gamma)\right) - \Omega_t^{cc}\Pi_{t(\Gamma)} f_t^0 \widetilde{f}_t^\top{}_{(\Gamma)} \tag{42}$$

*is U-mean converging.*

*Proof.* The cross-section convergence is the bulk of the analysis. We proceed by analyzing the six terms one by one. We list out the statements in each step and provide in-line proofs of the statements.

1. $\frac{1}{N}C_t^\top e_t$ is U-mean square converging.

   This is by Lemma 9, treating $c_{i,t}e_{i,t}$ as the $x_{i,t}$ in the lemma. The conditions are met given assumptions A, B.

2. $\left(\Gamma^\top\Omega_t^{cc}\Gamma\right)^{-1}\Gamma^\top\Omega_t^{cc}\Gamma^0$ is U-a.s. bounded.

   This is because it is a continuous function w.r.t. $\Gamma, \Omega_t^{cc}$, and $\Gamma^0$ whose domains are all bounded and away from singularity given assumptions C.

3. $\widetilde{f}_t^\top{}_{(\Gamma)}$ is U-mean square bounded.

   $\widetilde{f}_{t(\Gamma)} = \left(\Gamma^\top\Omega_t^{cc}\Gamma\right)^{-1}\Gamma^\top\Omega_t^{cc}\Gamma^0 f_t^0$, in which $\left(\Gamma^\top\Omega_t^{cc}\Gamma\right)^{-1}\Gamma^\top\Omega_t^{cc}\Gamma^0$ is U-a.s. bounded by the previous statement, $f_t^0$ is U-mean square bounded by assumption B1. Then apply lemma 8.4.

4. $\frac{1}{N}\sum_i\left(c_{i,t}^\top c_{i,t} - \Omega_t^{cc}\right)$ is U-mean square converging.

   The argument is the same as statement number 1 above. Treat $c_{i,t}^\top c_{i,t} - \Omega_t^{cc}$ as the $x_{i,t}$ and apply Lemma 9. The conditions are met given the definition of $\Omega_t^{cc}$ and assumption B.

5. $\frac{1}{N}S_t^{[1]}(\Gamma)$ is U-mean converging.

   Notice decomposition $\frac{1}{N}S_t^{[1]}(\Gamma) = \left[\frac{1}{N}C_t^\top e_t\right]\left[\widetilde{f}_t^\top{}_{(\Gamma)}\right]$, use the two previous statements about the two parts and apply lemma 8.1.

Electronic copy available at: https://ssrn.com/abstract=2983919

6. $\Pi_{t(\Gamma)} f_t^0 \widetilde{f}_t^{\top}{}_{(\Gamma)}$ is U-mean square bounded.

$$\Pi_{t(\Gamma)} f_t^0 \widetilde{f}_t^{\top}{}_{(\Gamma)} = \left[\left(\mathbb{I}_L - \Gamma\left(\Gamma^\top \Omega_t^{cc} \Gamma\right)^{-1} \Gamma^\top \Omega_t^{cc}\right)\Gamma^0\right]\left[f_t^0 f_t^{0\top}\right]\left[\Gamma^{0\top}\Omega_t^{cc}\Gamma\left(\Gamma^\top\Omega_t^{cc}\Gamma\right)^{-1}\right] \tag{43}$$

The third term, according to statement number 2 above, is U-a.s. bounded. By the same arguments, so is the first term. The middle term is U-mean square bounded by assumption B(1). Then by Lemma 8.4, the three things together is U-mean square bounded.

7. $\frac{1}{N} S_t^{[2]}(\Gamma) - \Omega_t^{cc} \Pi_{t(\Gamma)} f_t^0 \widetilde{f}_t^{\top}{}_{(\Gamma)}$ is U-mean converging.

$$\frac{1}{N} S_t^{[2]}(\Gamma) = \frac{1}{N} C_t^\top C_t \Pi_{t(\Gamma)} f_t^0 \widetilde{f}_t^{\top}{}_{(\Gamma)} \tag{44}$$

$$\frac{1}{N} S_t^{[2]}(\Gamma) - \Omega_t^{cc} \Pi_{t(\Gamma)} f_t^0 \widetilde{f}_t^{\top}{}_{(\Gamma)} = \left[\frac{1}{N}\sum_i \left(c_{i,t}^\top c_{i,t} - \Omega_t^{cc}\right)\right]\left[\Pi_{t(\Gamma)} f_t^0 \widetilde{f}_t^{\top}{}_{(\Gamma)}\right] \tag{45}$$

Then, straightforward application of the previous two statements on Lemma 8.1.

8. $C_t^\top C_t \Gamma \left(\Gamma^\top C_t^\top C_t \Gamma\right)^{-1}$ is U-a.s. bounded.

The term equals $\left[\frac{1}{N} C_t^\top C_t \Gamma\right]\left[\left(\Gamma^\top \frac{1}{N} C_t^\top C_t \Gamma\right)^{-1}\right]$. Obviously the first part is U-a.s. bounded. Treat the second term as a non-linear function in the form of $\left(\Gamma^\top \Omega \Gamma\right)^{-1}$, which is a continuous function for non-singular inputs. It remains to show that the domain of the inputs are bounded and away from singularity so that the output is bounded. We know $\frac{1}{N} C_t^\top C_t$ is not only bounded, but also uniformly approaches $\Omega_t^{cc}$ a.s., which is invertible a.s. by Assumption C.2. So for large enough $N$, $\frac{1}{N} C_t^\top C_t$ is invertible a.s. as well. Also $\Gamma$ is full rank according to Assumption C.1.

9. $\frac{1}{N} \Gamma^\top C_t^\top \widetilde{e}_{t(\Gamma)}$ is U-mean square converging.

$$\frac{1}{N} \Gamma^\top C_t^\top \widetilde{e}_{t(\Gamma)} = \frac{1}{N} \Gamma^\top C_t^\top e_t + \frac{1}{N} \Gamma^\top C_t^\top C_t Q_t^{\top}{}_{(\Gamma)} \Gamma^0 f_t^0 \tag{46}$$

$$= \frac{1}{N} \Gamma^\top C_t^\top e_t + \frac{1}{N} \Gamma^\top \sum_i \left(c_{i,t}^\top c_{i,t} - \Omega_t^{cc}\right) Q_t^{\top}{}_{(\Gamma)} \Gamma^0 f_t^0 \tag{47}$$

The first term is U-mean square converging, according to statement 1, assump-

Electronic copy available at: https://ssrn.com/abstract=2983919

tion C, and Lemma 8.3. We want to show so is the second. We put a vect $(\cdot)$ operator to the summand, which does not affect the norm. Rearrange it as,

$$\text{vect}\left(\left(c_{i,t}^{\top}c_{i,t} - \Omega_t^{cc}\right) Q_t^{\top}{}_{(\Gamma)}\Gamma^0 f_t^0\right) = \left(\left(c_{i,t}^{\top}c_{i,t} - \Omega_t^{cc}\right) \otimes f_t^{0\top}\right) \text{vect}\left(Q_t^{\top}{}_{(\Gamma)}\Gamma^0\right) \tag{48}$$

So the second term all together equals:

$$\Gamma^{\top}\left[\frac{1}{N}\sum_i \left(c_{i,t}^{\top}c_{i,t} - \Omega_t^{cc}\right) \otimes f_t^{0\top}\right] \text{vect}\left(Q_t^{\top}{}_{(\Gamma)}\Gamma^0\right) \tag{49}$$

The first and third part is U-a.s. bounded. The middle part is U-mean square converging, lemma 9, given assumptions B(4). Then, we can just apply Lemma 8.4 twice.

10. $\frac{1}{N}S_t^{[3]}(\Gamma)$ is U-mean converging.

$$\frac{1}{N}S_t^{[3]}(\Gamma) = \left[C_t^{\top}C_t\Gamma\left(\Gamma^{\top}C_t^{\top}C_t\Gamma\right)^{-1}\right]\left[\frac{1}{N}\Gamma^{\top}C_t^{\top}\widetilde{e}_{t(\Gamma)}\right]\left[\widetilde{f}_t^{\top}{}_{(\Gamma)}\right] \tag{50}$$

The three term are U-a.s. bounded, U-mean square converging, and U-mean square bounded, and apply Lemma 8.

11. $\frac{1}{N}S_t^{[4]}(\Gamma)$, $\frac{1}{N}S_t^{[5]}(\Gamma)$, $\frac{1}{N}S_t^{[6]}(\Gamma)$ are all U-mean converging.

For the last three terms, given the similarities to the situations above, we just write out the decompositions. The remaining arguments about repeatedly applying Lemma 8 are omitted.

$$\frac{1}{N}S_t^{[4]}(\Gamma) = \left[\frac{1}{N}C_t^{\top}e_t\right]\left[\frac{1}{N}\widetilde{e}_t^{\top}{}_{(\Gamma)}C_t\Gamma\right]\left(\Gamma^{\top}\frac{1}{N}C_t^{\top}C_t\Gamma\right)^{-1} \tag{51}$$

$$\frac{1}{N}S_t^{[5]}(\Gamma) = \left[\frac{1}{N}C_t^{\top}C_t\Pi_{t(\Gamma)}\right]\left[f_t^0\right]\left[\frac{1}{N}\widetilde{e}_t^{\top}{}_{(\Gamma)}C_t\Gamma\right]\left(\Gamma^{\top}\frac{1}{N}C_t^{\top}C_t\Gamma\right)^{-1} \tag{52}$$

$$\frac{1}{N}S_t^{[6]}(\Gamma) = \left[\frac{1}{N}C_t^{\top}C_t\Gamma\left(\Gamma^{\top}\frac{1}{N}C_t^{\top}C_t\Gamma\right)^{-1}\right]\left[\frac{1}{N}\widetilde{e}_t^{\top}{}_{(\Gamma)}C_t\Gamma\right]\left[\frac{1}{N}\Gamma^{\top}C_t^{\top}\widetilde{e}_{t(\Gamma)}\right] \tag{53}$$

$$\left(\Gamma^{\top}\frac{1}{N}C_t^{\top}C_t\Gamma\right)^{-1} \tag{54}$$

Finally, given the analysis above of $S_t^{[1]}\cdots S_t^{[6]}$, we can conclude the required state-

Electronic copy available at: https://ssrn.com/abstract=2983919

ment. □

**Lemma 11** ($S_T$ Convergence). $\sup_{\Gamma \in \Psi} \|S_T(\Gamma) - S^0(\Gamma)\| \xrightarrow{p} 0, T \to \infty$.

*Proof.* This is a familiar case in the sense that it only has the time-series dimension — this is a stationary and ergodic time-series average analysis. The only twist is it requires uniform convergence over $\Gamma \in \Psi$. We proceed by applying Lemma 2.4 in Newey and McFadden (1994). It requires to construct the $\Gamma$-irrelevant random variable $d_t$, and verify it dominates $t$ and has finite expectation. Notice, $S_T(\Gamma) = \frac{1}{T}\sum_t \text{vect}\,(t(\Gamma))$ and $t(\Gamma) = \Omega_t^{cc}\Pi_{t(\Gamma)} f_t^0 \widetilde{f}_t^{\top}(\Gamma)$

$$\|\text{vect}\,(t(\Gamma))\| = \|t(\Gamma)\| = \left\|[\Omega_t^{cc}\Pi_{t(\Gamma)}]\left[f_t^0 f_t^{0\top}\right]\left[\Gamma^{0\top}\Omega_t^{cc}\Gamma\left(\Gamma^{\top}\Omega_t^{cc}\Gamma\right)^{-1}\right]\right\| \tag{55}$$

$$\leq \|\Omega_t^{cc}\Pi_{t(\Gamma)}\|\left\|\Gamma^{0\top}\Omega_t^{cc}\Gamma\left(\Gamma^{\top}\Omega_t^{cc}\Gamma\right)^{-1}\right\|\left\|f_t^0 f_t^{0\top}\right\| \tag{56}$$

$$\leq M\left\|f_t^0 f_t^{0\top}\right\| := d_t \tag{57}$$

where $M$ is the a.s. bound, such that

$$Pr\left\{\sup_{\Gamma\in\Psi}\|\Omega_t^{cc}\Pi_{t(\Gamma)}\|\left\|\Gamma^{0\top}\Omega_t^{cc}\Gamma\left(\Gamma^{\top}\Omega_t^{cc}\Gamma\right)^{-1}\right\| < M\right\} = 1. \tag{58}$$

A finite $M$ exists, because the two norms within sup are continuous functions on compact domains, given Assumption C. We have thus constructed $d_t$, and shown that $\|\text{vect}\,(t(\Gamma))\| \leq \|d_t\|$. Moreover, $\mathbb{E}d_t < \infty$ given Assumption B(1). □

### C.6.3 The Main Proof of Proposition 1

After preparing the lemmas above, we are finally ready for the main proof of Proposition 1.

*Proof.* According to Lemma 10, $\frac{1}{N}\sum_i \left(S_t^{[1]}(\Gamma) + \cdots + S_t^{[6]}(\Gamma)\right) - \Omega_t^{cc}\Pi_{t(\Gamma)} f_t^0 \widetilde{f}_t^{\top}(\Gamma)$ is U-mean converging. We have the time-series average defined as

$$S_T(\Gamma) := \frac{1}{T}\sum_t \text{vect}\left(\Omega_t^{cc}\Pi_{t(\Gamma)} f_t^0 \widetilde{f}_t^{\top}(\Gamma)\right) \tag{59}$$

Apply Lemma 7, we have $\sup_{\Gamma\in\Psi}\|S(\Gamma) - S_T(\Gamma)\| \xrightarrow{p} 0, N \to \infty, \forall T$. That is to say,

Electronic copy available at: https://ssrn.com/abstract=2983919

$\forall \epsilon, \delta > 0$, $\exists N^{[1]}$, s.t. $\forall T$ and $\forall N > N^{[1]}$

$$Pr\left\{\sup_{\Gamma \in \Psi} \|S(\Gamma) - S_T(\Gamma)\| < \delta\right\} > 1 - \epsilon. \tag{60}$$

By Lemma 11, $\sup_{\Gamma \in \Psi} \|S_T(\Gamma) - S^0(\Gamma)\| \xrightarrow{p} 0, T \to \infty$. That is to say, $\forall \epsilon, \delta > 0$, $\exists T^{[1]}$, s.t. irrelevant of $N$, $\forall T > T^{[1]}$

$$Pr\left\{\sup_{\Gamma \in \Psi} \left\|S_T(\Gamma) - S^0(\Gamma)\right\| < \delta\right\} > 1 - \epsilon. \tag{61}$$

Combined, $\forall N > N^{[1]}, T > T^{[1]}$

$$\begin{aligned} & Pr\left\{\sup_{\Gamma \in \Psi} \left\|S(\Gamma) - S^0(\Gamma)\right\| < 2\delta\right\} && (62) \\ \geq & Pr\left\{\sup_{\Gamma \in \Psi} \|S(\Gamma) - S_T(\Gamma)\| < \delta \text{ AND } \sup_{\Gamma \in \Psi} \left\|S_T(\Gamma) - S^0(\Gamma)\right\| < \delta\right\} && (63) \\ \geq & 1 - 2\epsilon. && (64) \end{aligned}$$

That means the required conclusion: $\sup_{\Gamma \in \Psi} \|S(\Gamma) - S^0(\Gamma)\| \xrightarrow{p} 0, N, T \to \infty$ □

## C.7 Proof of Proposition 2

*Proof.* "If": It is easy to verify that $\Gamma^0$ and all of its rotations solve $S^0(\Gamma) = \mathbf{0}$, because they solve $\Pi_t(\Gamma) = \mathbf{0}$, $\forall \omega$.

"Only if": All we need to show is that $\forall \Gamma$ not rotationally equivalent to $\Gamma^0$, $S^0(\Gamma) \neq \mathbf{0}$. The random term in $S^0$:

$$\begin{aligned} \Omega_t^{cc} \Pi_{t(\Gamma)} f_t^0 \widetilde{f}_t^\top(\Gamma) &= \Omega_t^{cc} \left(\mathbb{I}_L - \Gamma \left(\Gamma^\top \Omega_t^{cc} \Gamma\right)^{-1} \Gamma^\top \Omega_t^{cc}\right) \Gamma^0 f_t^0 f_t^{0\top} \Gamma^{0\top} \Omega_t^{cc} \Gamma \left(\Gamma^\top \Omega_t^{cc} \Gamma\right)^{-1} \\ &:= A R f_t^0 f_t^{0\top} R^\top B, \end{aligned} \tag{65}$$

where $R$ is rotation s.t. $\mathbb{E} R f_t^0 f_t^{0\top} R$ is diagonal with positive entries, and $A$ and $B$ are the shorthands of $L \times K$ and $K \times K$ respectively. Notice $\forall \Gamma$ not rotationally equivalent to $\Gamma^0$, $\Pi_{t(\Gamma)} \neq \mathbf{0}$, so $A$, $B$ are full rank $K$ w.p. 1, by Assumption C. One can construct constant $p, q$ of lengths $L, K$ s.t. the signs of each entry in $p^\top A$ and $Bq$ are always the same. As a result, $p^\top \mathbb{E}\left[A R f_t^0 f_t^{0\top} R B\right] q > 0$. □

Electronic copy available at: https://ssrn.com/abstract=2983919

## C.8 Specific Identification Functions Satisfy IF.1–2

**Lemma 12** (Verify Condition IF.1–2)**.** *The identification functions $I_{\mathrm{X}}(\Gamma)$ and $I_{\mathrm{Y}}(\Gamma)$ both satisfy Condition IF.1–2.*

*Proof.* It is obvious for $\Theta_{\mathrm{X}}$, because it is deterministic. The identification function and its limit are the same $I_{\mathrm{X}}(\Gamma) = I^0_{\mathrm{X}}(\Gamma) = \mathrm{vect}\left(Block_{1:K}(\Gamma) - \mathbb{I}_K\right)$.

For $\Theta_{\mathrm{Y}}$, construct $I^0_{\mathrm{Y}}$ as

$$I^0_{\mathrm{Y}}(\Gamma) := \begin{bmatrix} \mathrm{veca}\left(\Gamma^\top\Gamma - \mathbb{I}_K\right) \\ \mathrm{vecb}\left(\mathbb{E}\left[\widetilde{f}_t(\Gamma)\widetilde{f}_t^\top(\Gamma)\right] - V^{ff}\right) \end{bmatrix}.^{48} \tag{66}$$

Next, we need to verify the two parts of the Condition.

**Verify IF.1:**

The top $\frac{1}{2}K(K+1)$ rows of $I_{\mathrm{Y}}$ and $I^0_{\mathrm{Y}}$ are the same. Next, we mostly need to work on the lower part of $I_{\mathrm{Y}}$ and $I^0_{\mathrm{Y}}$. Define

$$I_{[2]}(\Gamma) := \frac{1}{T}\sum_t \widehat{f}_t(\Gamma)\widehat{f}_t^\top(\Gamma) - V^{ff} \tag{67}$$

$$I^0_{[2]}(\Gamma) := \mathbb{E}\left[\widetilde{f}_t(\Gamma)\widetilde{f}_t^\top(\Gamma)\right] - V^{ff} \tag{68}$$

Given $\widehat{f}_t(\Gamma) = \widetilde{f}_t(\Gamma) + \left(\Gamma^\top C_t^\top C_t\Gamma\right)^{-1}\Gamma^\top C_t^\top \widetilde{e}_t(\Gamma)$, we have

$$\begin{aligned} I_{[2]}(\Gamma) - I^0_{[2]}(\Gamma) = \frac{1}{T}\sum_t \Big( & \widetilde{f}_t(\Gamma)\widetilde{f}_t^\top(\Gamma) - \mathbb{E}\left[\widetilde{f}_t(\Gamma)\widetilde{f}_t^\top(\Gamma)\right] \\ & + \widetilde{f}_t(\Gamma)\widetilde{e}_t^\top(\Gamma) C_t\Gamma\left(\Gamma^\top C_t^\top C_t\Gamma\right)^{-1} \\ & + \left(\Gamma^\top C_t^\top C_t\Gamma\right)^{-1}\Gamma^\top C_t^\top \widetilde{e}_t(\Gamma)\widetilde{f}_t^\top(\Gamma) \\ & + \left(\Gamma^\top C_t^\top C_t\Gamma\right)^{-1}\Gamma^\top C_t^\top \widetilde{e}_t(\Gamma)\widetilde{e}_t^\top(\Gamma) C_t\Gamma\left(\Gamma^\top C_t^\top C_t\Gamma\right)^{-1}\Big). \end{aligned} \tag{69}$$

The analysis from here is very similar to the proof of Proposition 1.

For the first term, mimic Lemma 11 to draw the conclusion that

$$\sup_{\Gamma\in\Psi}\left\|\frac{1}{T}\sum_t\left(\widetilde{f}_t(\Gamma)\widetilde{f}_t^\top(\Gamma)\right) - \mathbb{E}\left[\widetilde{f}_t(\Gamma)\widetilde{f}_t^\top(\Gamma)\right]\right\| \xrightarrow{p} 0, \qquad T\to\infty. \tag{70}$$

[48] We are constructing $I^0_{\mathrm{Y}}$ that will serve in IA to pick the true $\Gamma^0(\Theta^0_{\mathrm{Y}})$. So in defining $I^0_{\mathrm{Y}}$, function $\widetilde{f}_t(\Gamma)$ relies on a $\Gamma^0$ that is not pinned down yet. The factors needs to be accordingly rotated by "re-defining" $f^0_t = \widetilde{f}_t\left(\Gamma^0(\Theta^0_{\mathrm{Y}})\right)$.

Electronic copy available at: https://ssrn.com/abstract=2983919

Define the sum of the second, third, and forth term inside the summation as $\Xi_t(\Gamma)$:

$$\Xi_t(\Gamma) = \widetilde{f}_{t(\Gamma)}\widetilde{e}_t^\top{}_{(\Gamma)} C_t\Gamma \left(\Gamma^\top C_t^\top C_t\Gamma\right)^{-1} \tag{71}$$

$$+ \left(\Gamma^\top C_t^\top C_t\Gamma\right)^{-1} \Gamma^\top C_t^\top \widetilde{e}_{t(\Gamma)}\widetilde{f}_t^\top{}_{(\Gamma)} \tag{72}$$

$$+ \left(\Gamma^\top C_t^\top C_t\Gamma\right)^{-1} \Gamma^\top C_t^\top \widetilde{e}_{t(\Gamma)}\widetilde{e}_t^\top{}_{(\Gamma)} C_t\Gamma \left(\Gamma^\top C_t^\top C_t\Gamma\right)^{-1} \tag{73}$$

Repeat the arguments in Lemma 10 and we arrive at the conclusion that $\Xi_t(\Gamma)$ is U-mean converging.

Combining these two conclusions, following the same arguments in the proof of Proposition 1, IF.1 is verified.

**Verify IF.2:**

For any parameter $\Gamma$ and its rotation $\Gamma' = \Gamma R$, we find the relationship that mirrors Eq. 16 in Lemma 6.

$$\mathbb{E}\left[\widetilde{f}_t\left(\Gamma R\right)\widetilde{f}_t^\top\left(\Gamma R\right)\right] = R^{-1\top}\mathbb{E}\left[\widetilde{f}_{t(\Gamma)}\widetilde{f}_t^\top{}_{(\Gamma)}\right] R^{-1} \tag{74}$$

Given this property, we find $I_{\mathrm{Y}}^0$ and $I_{\mathrm{Y}}$ behave in the same way when $\Gamma$ is rotated. Therefore, the procedure of the normalization is the same as that in that Lemma 6, which in turn proves the normalization is unique, that verifies IF.2.

A difference to point out is that here the input of the normalization procedure $\mathbb{E}\left[\widetilde{f}_{t(\Gamma)}\widetilde{f}_t^\top{}_{(\Gamma)}\right]$ is an object about the underlying true and unknown to econometrian. □

## C.9 Proof of Theorem 1

*Proof.* We stack the score and identification functions together and treat the pair as a single function. By their definitions, $\widehat{\Gamma}$ solves $\left[S; I\right](\Gamma) = \mathbf{0}$ and $\Gamma^0(\Theta)$ solves $\left[S^0; I\right](\Gamma) = \mathbf{0}$. Proposition 1 and Condition IF.1 imply that the stacked function $[S; I]$ is uniformly converging:

$$\sup_{\Gamma\in\Psi}\left\|\left[S; I\right](\Gamma) - \left[S^0; I^0\right](\Gamma)\right\| \xrightarrow{p} 0, \qquad\qquad N, T \to \infty$$

Based on the previous analysis, Proposition 2 and IF.1-2 combined imply $\Gamma^0$ is the unique solution to the function limits' equation system: $\left[S^0; I^0\right](\Gamma) = \mathbf{0}$. We know the solution of the uniformly converging function converges to the limit's *unique* solution,

Electronic copy available at: https://ssrn.com/abstract=2983919

according to Newey and McFadden (1994). Therefore, $\widehat{\Gamma} \xrightarrow{p} \Gamma^0$. Similarly, $[S^0; I]$ uniformly converges to $[S^0; I^0]$, implying $\Gamma^0(\Theta) \xrightarrow{p} \Gamma^0$. These two convergence results combined imply $\widehat{\Gamma} - \Gamma^0(\Theta)$ converges to zero. □

Corollary 1 is a direct application of Theorem 1.

## C.10 Proof of Theorem 2

*Proof.* We first consider $\widehat{\gamma} - \gamma^0$. Later, $\widehat{\gamma} - \gamma^0(\Theta)$ has almost the same derivation, except for one important difference.

Per Lagrange's Mean Value Theorem, at each realization, there exists $\bar{\gamma}$ in-between $\widehat{\gamma}$ and $\gamma^0$ such that

$$S\left(\widehat{\gamma}\right) = S(\gamma^0) + \left.\frac{\partial S\left(\gamma\right)}{\partial \gamma^\top}\right|_{\gamma=\bar{\gamma}} \left(\widehat{\gamma} - \gamma^0\right). \tag{75}$$

Notice $S\left(\widehat{\gamma}\right) = \mathbf{0}$. That means

$$\bar{H}\left(\widehat{\gamma} - \gamma^0\right) = -S(\gamma^0), \tag{76}$$

where $\bar{H} := \left.\frac{\partial S(\gamma)}{\partial \gamma^\top}\right|_{\gamma=\bar{\gamma}}$ is shorthand for the sample Hessian matrix at $\bar{\gamma}$.[49]

Usually, without the unidentification problem, one would divide by the Hessian to get $\widehat{\gamma} - \gamma = -\bar{H}^{-1} S(\gamma^0)$, and then take the limit of $\bar{H}$ to express the asymptotic distribution. But that is not possible here when $S$ has non-unique solutions.

The unidentification problem manifests as a singular $\bar{H}$. Specifically, the score should have zero gradients on the $K^2$ directions of unidentification. Intuitively, think in the IPCA case, where the rotation matrix $R$ is $K \times K$. So there are $K^2$ directions to marginally perturb $\Gamma^0$ without changing the score — it would be constantly $\mathbf{0}$. This implies the score has zero gradient on those $K^2$ directions, meaning $\bar{H}$, although an $LK$-square matrix, has a rank of only $LK - K^2$.

A rank-deficient $\bar{H}$ leads to the non-uniqueness of the $\widehat{\gamma}$ solutions in equation (76). Given one $\widehat{\gamma} - \gamma^0$ that solves the linear equation, there are a family of other vectors

[49] Lagrange's Mean Value Theorem only applies to function of multiple inputs and scalar output, but not to vector-valued functions. So each row of $\bar{H}$ are evaluated at a different $\bar{\gamma}$. Same for the additional $K^2$ rows of $\bar{J}$ below. But importantly, the different $\bar{\gamma}$'s are all in between $\gamma^0$ and $\widehat{\gamma}$ (in a linear-combination sense). This guarantees later we can take the limit of the Hessian (see Newey and McFadden, 1994, footnote 25).

Electronic copy available at: https://ssrn.com/abstract=2983919

that all solve the equation, forming a $K^2$-dimensional sub-linear space of solutions. This is the unidentification issue manifested as the same non-unique solution problem but now for a *linearized* equation.

To solve the problem, we resort to the identification function to pin down the indeterminacy. We linearize $I$ similarly and then append it to the linearized score (remember $I\left(\widehat{\gamma}\right) = \mathbf{0}$ too):

$$I\left(\widehat{\gamma}\right) = I(\gamma^0) + \left.\frac{\partial I\left(\gamma\right)}{\partial \gamma^\top}\right|_{\gamma=\bar{\gamma}} \left(\widehat{\gamma} - \gamma^0\right). \tag{77}$$

$$\bar{J}\left(\widehat{\gamma} - \gamma^0\right) = -I(\gamma^0). \tag{78}$$

where $\bar{J} := \left.\frac{\partial I(\gamma)}{\partial \gamma^\top}\right|_{\gamma=\bar{\gamma}}$ is $I$'s counterpart of the Hessian. Notice $I$ is $K^2 \times 1$ vector and $J$ is $K^2 \times LK$ Jacobian. So, these additional $K^2$ equation pins down the addition $K^2$ degrees of freedom. Append the $K^2$ equations in (78) below (76) and form a single linearization of the estimator,

$$\left[\bar{H}; \bar{J}\right]\left(\widehat{\gamma} - \gamma^0\right) = -\left[S(\gamma^0); I(\gamma^0)\right].$$

Now this linear equation system has a unique solution at $(\widehat{\gamma} - \gamma^0)$, because the stacked $\left[\bar{H}; \bar{J}\right]$ matrix of size $(LK + K^2) \times LK$ is now full rank $LK$. To solve it, left multiply both sides by the pseudoinverse $\left(\left[\bar{H}; \bar{J}\right]^\top \left[\bar{H}; \bar{J}\right]\right)^{-1} \left[\bar{H}; \bar{J}\right]^\top$:

$$\widehat{\gamma} - \gamma^0 = -\left(\bar{H}^\top \bar{H} + \bar{J}^\top \bar{J}\right)^{-1} \left(\bar{H}^\top S(\gamma^0) + \bar{J}^\top I(\gamma^0)\right). \tag{79}$$

With the estimator thus linearized, the rest of asymptotic derivation follows canonical $M$-estimator analysis. Given, $\widehat{\gamma} \xrightarrow{p} \gamma^0$, then $\bar{\gamma}$, the mean value between $\widehat{\gamma}$ and $\gamma^0$, goes to $\gamma^0$ as well. By the continuous mapping theorem, we find the limits of the Hessians:

$$\underset{N,T\to\infty}{\text{plim}}\ \bar{H} = H^0 := \left.\frac{\partial S^0\left(\gamma\right)}{\partial \gamma^\top}\right|_{\gamma=\gamma^0}, \quad \underset{N,T\to\infty}{\text{plim}}\ \bar{J} = J^0 := \left.\frac{\partial I^0\left(\gamma\right)}{\partial \gamma^\top}\right|_{\gamma=\gamma^0}.^{50} \tag{80}$$

Then, taking the probability limit of equation (79) leads to line 2.a in Theorem 2.

The derivation of line 2.b is similar, save for an important difference. Let us redo this proof, but now start by applying the Lagrange's Mean Value Theorem to $(\widehat{\gamma} - \gamma^0(\Theta))$ instead of $(\widehat{\gamma} - \gamma^0)$. Denote the requisite mean value as $\bar{\bar{\gamma}}$ and denote

Electronic copy available at: https://ssrn.com/abstract=2983919

the Hessians evaluated at $\bar{\bar{\gamma}}$ as $\bar{\bar{H}}, \bar{\bar{J}}$. Then, the same logic applies till we find the counterpart of equation (79) as

$$\widehat{\gamma} - \gamma^0(\Theta) = -\left(\bar{\bar{H}}^\top \bar{\bar{H}} + \bar{\bar{J}}^\top \bar{\bar{J}}\right)^{-1} \left(\bar{\bar{H}}^\top S\left(\gamma^0(\Theta)\right) + \bar{\bar{J}}^\top I\left(\gamma^0(\Theta)\right)\right). \tag{81}$$

Not only $\widehat{\gamma}$ but also $\gamma^0(\Theta)$ converge to $\gamma^0$. Hence the new mean value $\bar{\bar{\gamma}}$ between them converges to $\gamma^0$ as well. So, the limits of $\bar{\bar{H}}$ and $\bar{\bar{J}}$ are still $H^0$ and $J^0$, which are evaluated at the same deterministic $\gamma^0$.[51]

The important difference in this case is that $I\left(\gamma^0(\Theta)\right) = \mathbf{0}$ by construction, which eliminates the $I$ term in (81) and leads to line 2.b in the Theorem. This difference carries an important intuition that is discussed below.

The remainder of the proof only needs to shows $S\left(\gamma^0(\Theta)\right) = S(\gamma^0) + o_p\left(S(\gamma^0)\right)$. Apply Mean Value for another time in between $\gamma^0(\Theta)$ to $\gamma^0$:

$$S\left(\Gamma^0(\Theta)\right) - S(\Gamma^0) = \left.\frac{\partial S(\Gamma)}{\partial \gamma^\top}\right|_{\gamma=\gamma^{[3]}} \left(\gamma^0(\Theta) - \gamma^0\right) \tag{82}$$

$$= \left(\left.\frac{\partial S(\Gamma)}{\partial \gamma^\top}\right|_{\gamma=\gamma^{[3]}} - \left.\frac{\partial S^0(\Gamma)}{\partial \gamma^\top}\right|_{\gamma=\gamma^{[3]}}\right)\left(\gamma^0(\Theta) - \gamma^0\right) + \left.\frac{\partial S^0(\Gamma)}{\partial \gamma^\top}\right|_{\gamma=\gamma^{[3]}} \left(\gamma^0(\Theta) - \gamma^0\right) \tag{83}$$

Notice the first term is $\mathcal{O}_p\left(S^0(\Gamma)\right) o_p(1) = o_p\left(S^0(\Gamma)\right)$. The second term is $\mathbf{0}$ because $S^0$ is constant at $\mathbf{0}$ from $\gamma^0(\Theta)$ to $\gamma^0$. □

## C.11 Lemma 3

We state and prove a general version of Lemma 3 that can be evaluated at two specific cases.

**Lemma 13** (Asymptotic Distribution of the Score evaluated at $\Gamma^0$ — Generic Identification)**.** *Under Assumptions A–F, if the identification condition $\Theta$ has an associated identification function $I(\Gamma)$ that satisfies IF.1–2, and $\gamma^0$ is under IA, then as $N, T \to \infty$ such that $T/N \to 0$,*

$$\sqrt{NT} S(\Gamma^0) \xrightarrow{d} \text{Normal}\left(\mathbf{0}, \mathbb{V}^{[1]}\right).$$

[51] This shows the importance of keeping the deterministic $\gamma^0$ as the limiting reference point in the linearizion, even though we are ultimately interested in the result about $\gamma^0(\Theta)$ and $\widehat{\gamma}$.

Electronic copy available at: https://ssrn.com/abstract=2983919

*where* $\mathbb{V}^{[1]} = \left(Q^0 \otimes \mathbb{I}_K\right)\Omega^{cef}\left(Q^{0\top} \otimes \mathbb{I}_K\right)$ *and* $Q^0 := Q_{t(\Gamma^0)}$ *given that* $Q_{t(\Gamma)} := \mathbb{I}_L - \Omega_t^{cc}\Gamma\left(\Gamma^\top\Omega_t^{cc}\Gamma\right)^{-1}\Gamma^\top$ *is constant over* $t$ *under Assumption F.*

Lemma 3 simply provides two special cases under Lemma 13. In particular, there are two differences between $\mathbb{V}_{\mathrm{X}}^{[1]}$ and $\mathbb{V}_{\mathrm{Y}}^{[1]}$. First, $Q^0$ is evaluated under either $\Gamma^0(\Theta_{\mathrm{X}})$ or $\Gamma^0(\Theta_{\mathrm{Y}}^0)$. Second, the correspond true $f_t^0$ also need to be rotated, resulting in rotated values of the asymptotic variance $\Omega^{cef}$.

Below is a proof of the general Lemma 13.

*Proof.* We evaluate score at $\Gamma^0$ by breaking it into the six parts in Eq. 41. When evaluated at $\Gamma^0$, $S_t^{[2]}(\Gamma^0) = S_t^{[4]}(\Gamma^0) = \mathbf{0}$, because they contain $Q_{t(\Gamma)}{}^\top\Gamma^0$ which is zero at $\Gamma = \Gamma^0$. For the rest four terms, we have them in two pairs, and only need to show they have the following results: as $N, T \to \infty$,

$$\frac{1}{\sqrt{NT}}\sum_t \operatorname{vect}\left(S_t^{[1]}(\Gamma^0) + S_t^{[3]}(\Gamma^0)\right) \xrightarrow{d} \text{Normal}\left(0, \mathbb{V}^{[1]}\right), \tag{84}$$

$$\frac{1}{\sqrt{NT}}\sum_t \operatorname{vect}\left(S_t^{[4]}(\Gamma^0) + S_t^{[6]}(\Gamma^0)\right) \xrightarrow{p} \mathbf{0}. \tag{85}$$

**Equation 84:** We have

$$S_t^{[1]}(\Gamma^0) + S_t^{[3]}(\Gamma^0) = \left[\mathbb{I}_L - C_t^\top C_t\Gamma^0\left(\Gamma^{0\top}C_t^\top C_t\Gamma^0\right)^{-1}\Gamma^{0\top}\right]C_t^\top e_t f_t^{0\top} \tag{86}$$

$$= \left(Q_{t(\Gamma^0)} - \epsilon_{N,t}\right)C_t^\top e_t f_t^{0\top} \tag{87}$$

where $\epsilon_{N,t} = C_t^\top C_t\Gamma^0\left(\Gamma^{0\top}C_t^\top C_t\Gamma^0\right)^{-1}\Gamma^{0\top} - \Omega_t^{cc}\Gamma^0\left(\Gamma^{0\top}\Omega_t^{cc}\Gamma^0\right)^{-1}\Gamma^{0\top}$. We break out the two terms, put them into $\frac{1}{\sqrt{NT}}\sum_t \operatorname{vect}(\cdot)$, and respectively show their asymptotics are:

$$\frac{1}{\sqrt{NT}}\sum_t \operatorname{vect}\left(Q_{t(\Gamma^0)}C_t^\top e_t f_t^{0\top}\right) \xrightarrow{d} \text{Normal}\left(0, \mathbb{V}^{[1]}\right), \tag{88}$$

$$\frac{1}{\sqrt{NT}}\sum_t \operatorname{vect}\left(\epsilon_{N,t}C_t^\top e_t f_t^{0\top}\right) \xrightarrow{p} \mathbf{0} \qquad N, T \to \infty. \tag{89}$$

For the first one, notice $\operatorname{vect}\left(Q^0 C_t^\top e_t f_t^{0\top}\right) = \left(Q^0 \otimes \mathbb{I}_K\right)\operatorname{vect}\left(C_t^\top e_t f_t^{0\top}\right)$. Then it is obvious applying a CMT to the CLT in Assumption D(1).

Electronic copy available at: https://ssrn.com/abstract=2983919

For the second one, we limit its second moment:

$$\left\|\frac{1}{\sqrt{NT}}\sum_t \mathrm{vect}\left(\epsilon_{N,t}C_t^\top e_t f_t^{0\top}\right)\right\|^2 \tag{90}$$

$$= \frac{1}{NT}\sum_{i,j,t,s} \mathrm{vect}\left(\epsilon_{N,t}c_{i,t}^\top e_{i,t} f_t^{0\top}\right)^\top \mathrm{vect}\left(\epsilon_{N,s}c_{j,s}^\top e_{i,t} {f^0}_s{}^\top\right) \tag{91}$$

$$= \frac{1}{NT}\sum_{i,j,t,s} \mathrm{vect}\left(\epsilon_{N,t}\otimes {f^0}_t\right) c_{i,t}^\top e_{i,t} e_{j,s} c_{j,s} \mathrm{vect}\left(\epsilon_{N,s}\otimes {f^0}_s\right)^\top \tag{92}$$

Take expectation:

$$\mathbb{E}\left\|\frac{1}{\sqrt{NT}}\sum_t \mathrm{vect}\left(\epsilon_{N,t}C_t^\top e_t f_t^{0\top}\right)\right\|^2 \tag{93}$$

$$\leq \frac{1}{NT}\sum_{i,j,t,s} \mathbb{E}\left\|\mathrm{vect}\left(\epsilon_{N,t}\otimes {f^0}_t\right)\right\| \left\|\tau_{ij,ts}\right\| \mathbb{E}\left\|\mathrm{vect}\left(\epsilon_{N,s}\otimes {f^0}_s\right)\right\| \tag{94}$$

$$\leq \sup_t \mathbb{E}\left\|\mathrm{vect}\left(\epsilon_{N,t}\otimes {f^0}_t\right)\right\|^2 \left(\frac{1}{NT}\sum_{i,j,t,s}\left\|\tau_{ij,ts}\right\|\right) \tag{95}$$

Notice the first term → 0, as $N, T \to \infty$, second term is bounded according to Assumption E. Therefore, this second moment → 0.

**Equation 85:** The summand:

$$S_t^{[4]}(\Gamma^0) + S_t^{[6]}(\Gamma^0) \tag{96}$$

$$= C_t^\top M_t(\Gamma^0) e_t e_t{}^\top C_t \Gamma^0 \left(\Gamma^{0\top} C_t^\top C_t \Gamma^0\right)^{-1} \tag{97}$$

$$= Q_t^N(\Gamma^0)\frac{1}{N}\left(C_t^\top e_t\right)\left(e_t{}^\top C_t\right)\Gamma^0\left(\frac{1}{N}\Gamma^{0\top}C_t^\top C_t\Gamma^0\right)^{-1} \tag{98}$$

where $Q_t^N(\Gamma) := \left(\mathbb{I}_L - \frac{1}{N}C_t^\top C_t \Gamma\left(\frac{1}{N}\Gamma^\top C_t^\top C_t \Gamma\right)^{-1}\Gamma^\top\right)$.

Notice the first term in (98) $Q_t^N(\Gamma)$ has a constant large-$N$ probability limit $Q^0$, which is defined in the statement of Lamma 13. It is the similar situation for the last terms, $\Gamma^0\left(\frac{1}{N}\Gamma^{0\top}C_t^\top C_t\Gamma^0\right)^{-1}$. Since both of the two parts are only of $C_t$, their large-$N$ convergence is bounded according to Assumption C.2. The constant large-$N$ limits can be analyzed outside of the $t$-summation. Therefore, all we need to show is about

Electronic copy available at: https://ssrn.com/abstract=2983919

the $t$-sum of the middle terms: $\frac{1}{\sqrt{NT}} \sum_t \text{vect}\left(\frac{1}{N}\left(C_t^\top e_t\right)\left(e_t{}^\top C_t\right)\right) \xrightarrow{p} \mathbf{0}$.

$$\begin{aligned}
&\frac{1}{\sqrt{NT}} \sum_t \text{vect}\left(\frac{1}{N}\left(C_t^\top e_t\right)\left(e_t{}^\top C_t\right)\right) && (99)\\
&= \frac{1}{T}\frac{\sqrt{T}}{\sqrt{N}} \sum_t \text{vect}\left(\frac{1}{N}\left(\sum_i c_{i,t}^\top e_{i,t}\right)\left(\sum_i c_{i,t}^\top e_{i,t}\right)^\top\right) && (100)\\
&= \frac{1}{T} \sum_t \text{vect}\left(\frac{\sqrt{T}}{\sqrt{N}}\frac{1}{N}\sum_{i,j} c_{i,t}^\top e_{i,t} e_{j,t} c_{j,t}\right) && (101)
\end{aligned}$$

Take unconditional expectation:

$$\begin{aligned}
&\mathbb{E}\frac{1}{\sqrt{NT}} \sum_t \text{vect}\left(\frac{1}{N}\left(C_t^\top e_t\right)\left(e_t{}^\top C_t\right)\right) && (102)\\
&= \mathbb{E}\text{vect}\left(\frac{\sqrt{T}}{\sqrt{N}}\frac{1}{N}\sum_{i,j} c_{i,t}^\top e_{i,t} e_{j,t} c_{j,t}\right) && (103)\\
&= \mathbb{E}\text{vect}\left(\frac{\sqrt{T}}{\sqrt{N}}\frac{1}{N}\sum_{i} c_{i,t}^\top c_{i,t} e_{i,t}^2\right) && (104)\\
&\to \mathbf{0}. && (105)
\end{aligned}$$

The first equation is due to time series stationarity, the second equation is due to cross-sectional i.i.d. SP.5, and the final limit is due to the cross-sectional LLN and the condition that $T/N \to 0$. Notice the random variable is non-negative, and its unconditional expectation is converging, thereby we have shown it probability limit is converging as well. □

Additional comment: based on this analysis, we conjecture that if $T/N$ converges to a fixed positive number, then Eq. 85 has a non-zero probability limit. As a result, the score's asymptotic distribution is still normal but with a bias. That means $\Gamma$ estimation would by asymptotically biased, but still consistent, as the "incidental parameter problem" would have suggested. We leave this scenario for future research.

## C.12 Calculate $H^0_{\text{X}}, H^0_{\text{Y}}$

Similarly to the situation in C.11, we state and prove and general result, about $H^0$ calculation, and discuss how to evaluate the general expression at $\Theta_{\text{X}}$ and $\Theta_{\text{Y}}$.

Electronic copy available at: https://ssrn.com/abstract=2983919

**Lemma 14** (Calculate $H^0$ - General)**.** *Under Assumption F,*

$$H^0 = \left(\Omega^{cc} \otimes V^{ff}\right) \frac{\partial}{\partial \gamma} \text{vect}\left(\Pi_{(\Gamma)}\right)\Big|_{\gamma=\gamma^0}.$$ [52]

Similar to the evaluation of $\mathbb{V}^{[1]}$ in Subsection C.11, there are two differences between $H^0_{\text{X}}$ and $H^0_{\text{Y}}$. First, the expression of $\Pi_{(\Gamma)}$ and its derivative is only about $\Omega^{cc}$ which does not depend on the rotation. But the derivative needs to be evaluated at either $\Gamma^0(\Theta_{\text{X}})$ or $\Gamma^0(\Theta^0_{\text{Y}})$. Second, the correspond true $f^0_t$ also need to be rotated. This will result in difference in the assumed asymptotic variance $V^{ff}$.

Next is the proof of the general Lemma 14.

*Proof.* Start from the definition

$$H^0 := \left.\frac{\partial S^0(\Gamma)}{\partial \gamma^\top}\right|_{\gamma=\gamma^0} = \mathbb{E}\left[\left.\frac{\partial \text{vect}\left(\Omega^{cc}_t \Pi_{t(\Gamma)} f^0_t \widetilde{f}^\top_t(\Gamma)\right)}{\partial \gamma^\top}\right|_{\gamma=\gamma^0}\right] \tag{106}$$

Notice $\Pi_t(\Gamma^0) = \mathbf{0}$. Therefore, the terms involving $\nabla \widetilde{f}_t$ after taking derivative will drop out, only $\nabla \Pi_t(\Gamma^0)$ terms survive. We write the $H^0$ column by column. The $p$'th column, or the derivative w.r.t. the $p$'th entry $\gamma_p$ simplifies as

$$H^0{}_p = \mathbb{E}\left[\text{vect}\left(\Omega^{cc}_t \left.\frac{\partial \Pi_{t(\Gamma)}}{\partial \gamma_p}\right|_{\gamma=\gamma^0} f^0_t f^{0\top}_t\right)\right] \tag{107}$$

This result above does not require Assumption F, and can be calculated by LLN simulation if one is interested in the general case. Now, we impose the constant $\Omega^{cc}_t$ assumption F:

$$H^0{}_p = \text{vect}\left(\Omega^{cc} \left.\frac{\partial \Pi_{(\Gamma)}}{\partial \gamma_p}\right|_{\gamma=\gamma^0} V^{ff}\right) = \left(\Omega^{cc} \otimes V^{ff}\right) \left.\frac{\partial \text{vect}\left(\Pi_{(\Gamma)}\right)}{\partial \gamma_p}\right|_{\gamma=\gamma^0} \tag{108}$$

Finally, append the columns together, we have $H^0 = \left(\Omega^{cc} \otimes V^{ff}\right) \left.\frac{\partial \text{vect}(\Pi(\Gamma))}{\partial \gamma^\top}\right|_{\gamma=\gamma^0}$ □

[52] Under Assumption F, $\Pi_{t(\Gamma)}$ is also time-constant and deterministic, so we drop its $t$ subscript. We choose do not write out the deterministic derivatives $\frac{\partial}{\partial \gamma} \text{vect}\left(\Pi_{(\Gamma)}\right)$ analytically for conciseness, which is easily calculated numerically in the simulated and empirical exercises.

Electronic copy available at: https://ssrn.com/abstract=2983919

## C.13 Calculate $J^0_{\mathrm{X}}, J^0_{\mathrm{Y}}$

$J^0_{\mathrm{X}}$ is much easier. The identification function $I_{\mathrm{X}}$ is deterministic, i.e. $I_{\mathrm{X}} = I^0_{\mathrm{X}}$. According to $J^0$'s definition (Eq. 80), we calculating $I^0_{\mathrm{X}}$'s derivative and get

$$J^0_{\mathrm{X}} = [\mathbb{I}_{K^2\times K^2}, \mathbf{0}_{K^2\times(L-K)K}]. \tag{109}$$

$J^0_{\mathrm{Y}}$ is much more involved and we summarize the calculation in the following lemma.

**Lemma 15** (Calculate $J^0_{\mathrm{Y}}$)**.** *Under Assumption F: $J^0_{\mathrm{Y}}$ is stacked up by two parts, the top $\frac{1}{2}K(K+1)$ rows are $J^0_{\mathrm{Y}\left[1:\left(\frac{1}{2}K(K+1)\right),\ :\ \right]} = \frac{\partial \mathrm{veca}\left(\Gamma^\top\Gamma\right)}{\partial\gamma^\top}\Big|_{\gamma=\gamma^0}$. For the bottom $\frac{1}{2}K(K-1)$ rows, the p'th column is*

$$J^0_{\mathrm{Y}\left[\left(\frac{1}{2}K(K+1)+1:K^2\right),\ p\ \right]} = \mathrm{vecb}\left(D_p\left(\Gamma^0, \Omega^{cc}\right)V^{ff} + V^{ff}D_p^\top\left(\Gamma^0, \Omega^{cc}\right)\right) \tag{110}$$

*Proof.* We omit the "S" subscript in this proof. The first part of $I$ is deterministic, so the first part of $J^0$ is easy. Lower part of $J^0$:

$$\begin{aligned}
J^0{}_{\left[\left(\frac{1}{2}K(K+1)+1:K^2\right),\ :\ \right]} &= \operatorname*{plim}_{N,T\to\infty} \frac{\partial}{\partial\gamma^\top}\mathrm{vecb}\left(I_{[2]}\right)\Big|_{\gamma^0} \\
&= \operatorname*{plim}_{N,T\to\infty} \frac{\partial}{\partial\gamma^\top}\mathrm{vecb}\left(\frac{1}{T}\sum_t\left(\widetilde{f}_{t(\Gamma)}\widetilde{f}_t^\top(\Gamma)\right)\right)\Bigg|_{\gamma^0} \\
\frac{\partial}{\partial\gamma_p}\mathrm{vecb}\left(\widetilde{f}_{t(\Gamma)}\widetilde{f}_t^\top(\Gamma)\right)\Big|_{\gamma^0} &= \mathrm{vecb}\left(\frac{\partial}{\partial\gamma_p}\widetilde{f}_{t(\Gamma)}\Big|_{\gamma^0} f_t^{0\top} + f_t^0\frac{\partial}{\partial\gamma_p}\widetilde{f}_t^\top(\Gamma)\Big|_{\gamma^0}\right) \\
\frac{\partial}{\partial\gamma_p}\widetilde{f}_{t(\Gamma)}\Big|_{\gamma^0} &= \frac{\partial}{\partial\gamma_p}\left(\Gamma^\top\Omega_t^{cc}\Gamma\right)^{-1}\Gamma^\top\Omega_t^{cc}\Gamma^0\Big|_{\gamma^0} f_t^0 := D_p\left(\Gamma^0, \Omega_t^{cc}\right)f_t^0 \\
\frac{\partial}{\partial\gamma_p}\mathrm{vecb}\left(\widetilde{f}_{t(\Gamma)}\widetilde{f}_t^\top(\Gamma)\right)\Big|_{\gamma^0} &= \mathrm{vecb}\left(D_p\left(\Gamma^0, \Omega_t^{cc}\right)f_t^0 f_t^{0\top} + f_t^0 f_t^{0\top}D_p^\top\left(\Gamma^0, \Omega_t^{cc}\right)\right) \\
J^0{}_{\left[\left(\frac{1}{2}K(K+1)+1:K^2\right),\ p\ \right]} &= \mathbb{E}\left[\mathrm{vecb}\left(D_p\left(\Gamma^0, \Omega_t^{cc}\right)f_t^0 f_t^{0\top} + f_t^0 f_t^{0\top}D_p^\top\left(\Gamma^0, \Omega_t^{cc}\right)\right)\right]
\end{aligned}$$

By Assumption F, $J^0{}_{\left[\left(\frac{1}{2}K(K+1)+1:K^2\right),\ p\ \right]} = \mathrm{vecb}\left(D_p\left(\Gamma^0, \Omega^{cc}\right)V^{ff} + V^{ff}D_p^\top\left(\Gamma^0, \Omega^{cc}\right)\right)$. □

Electronic copy available at: https://ssrn.com/abstract=2983919

## C.14 Proof of Theorem 4

*Proof.* (1) Given $\widehat{\Gamma}$ consistency to the two targets (Theorem 1), and that $\widetilde{f}_{t(\Gamma)}$ has bounded derivative around $\Gamma^0$, we have $\widetilde{f}_t(\widehat{\Gamma}) - \widetilde{f}_t\left(\Gamma^0(\Theta)\right) \xrightarrow{p} \mathbf{0}$. It remains to show the second term is $o_p(1)$. We have shown that $\frac{1}{N}\Gamma^\top C_t^\top \widetilde{e}_{t(\Gamma)}$ is U-mean square converging and $N\left(\Gamma^\top C_t^\top C_t \Gamma\right)^{-1}$ is U-a.s. bounded (see intermediate steps proving Lemma 10). Combined, $\left(\Gamma^\top C_t^\top C_t \Gamma\right)^{-1}\Gamma^\top C_t^\top \widetilde{e}_{t(\Gamma)}$ is U-mean square converging. Because that is a uniform result, when evaluate at $\widehat{\Gamma}$ for the second term, it is converging (in m.s. which implies in probability).

(2) Given Theorem 2 and Lemma 3, $\widehat{\Gamma} - \Gamma^0(\Theta) = \mathcal{O}_p\left(1/\sqrt{NT}\right)$. So, the first term is $o_p\left(1/\sqrt{NT}\right)$. For the second term:

$$\sqrt{N}\left(\widehat{\Gamma}^\top C_t^\top C_t \widehat{\Gamma}\right)^{-1}\widehat{\Gamma}^\top C_t^\top \widetilde{e}_t(\widehat{\Gamma}) = \sqrt{N}\left(\Gamma^{0\top} C_t^\top C_t \Gamma^0\right)^{-1}\Gamma^{0\top} C_t^\top e_t + o_p(1) \tag{111}$$

$$\xrightarrow{d} \text{Normal}\left(\mathbf{0}, \mathbb{V}_t^{[2]}\right), \tag{112}$$

where $\mathbb{V}_t^{[2]} = \left(\Gamma^{0\top}\Omega_t^{cc}\Gamma^0\right)^{-1}\Gamma^{0\top}\Omega_t^{ce}\Gamma^0\left(\Gamma^{0\top}\Omega_t^{cc}\Gamma^0\right)^{-1}$, with $\Omega_t^{ce}$ from Assumption D(2). Notice neither $\Omega_t^{cc}$ or $\Omega_t^{ce}$ are affected by the rotation. So one only needs to plug in the value of $\Gamma^0$ as either $\Gamma^0(\Theta_{\mathrm{X}})$ or $\Gamma^0(\Theta_{\mathrm{Y}}^0)$ to evaluate $\mathbb{V}_t^{[2]}$ for the specific cases. □

## C.15 Proof of Lemma 4

*Proof.* The top $\frac{1}{2}K(K+1)$ rows of $I_{\mathrm{Y}}(\Gamma^0)$ are deterministic at $\mathbf{0}$: $I_{\mathrm{Y}_{\left[1:\frac{1}{2}K(K+1)\right]}}(\Gamma^0) = \mathbf{0}_{\frac{1}{2}K(K+1)\times 1}$. Next we analyze the four terms in Eq. 69 in the proof Lemma 12 First term has a plim, which in general is non-zero. When evaluated at $\Gamma^0$, since $\widetilde{f}_t(\Gamma^0) = f_t^0$, it is $\mathcal{O}_p\left(1/\sqrt{T}\right)$. According to the analysis in Lemma 12, second and third are $\mathcal{O}_p\left(1/\sqrt{NT}\right)$. Fourth is $\mathcal{O}_p(1/N)$. That is to say

$$\operatorname*{plim}_{N,T\to\infty} \sqrt{T} I_{[2]}(\Gamma^0) - \frac{1}{\sqrt{T}}\sum_t \left(f_t^0 f_t^{0\top} - V^{ff}\right) = \mathbf{0}_{K\times K} \tag{113}$$

By the times-series CLT in D(3): $\sqrt{T}\mathrm{vecb}\left(I_{[2]}(\Gamma^0)\right) \xrightarrow{d} \text{Normal}\left(\mathbf{0}_{\frac{1}{2}K(K-1)\times 1}, \mathbb{V}^{[3]}\right)$. In addition, the cross-covariances between the top and the bottom part are $\mathbf{0}$. □

Electronic copy available at: https://ssrn.com/abstract=2983919

### C.16 Proof of Theorem 6

*Proof.* Straightforward based on Theorems 4 and 5. □

## D The Sign Issue in Normalization

As mentioned in footnote 21, identification conditions like $\Theta_{\mathrm{Y}}$ with only [1] and [2] does not pin down the signs of each $\Gamma$ column. To be precise, for any $\Gamma \in \Theta_{\mathrm{Y}}$, let $\Gamma' = \Gamma \mathrm{diag}\left\{s\right\}$, where $\mathrm{diag}\left\{s\right\}$ is a $K \times K$ diagonal matrix with $+1$ or $-1$'s on the diagonal. Then $\Gamma$ and $\Gamma'$ are obviously rotationally equivalent, because the corresponding factors can flip the signs accordingly. Meanwhile, both $\Gamma, \Gamma'$ satisfy $\Theta_{\mathrm{Y}}$'s [1] and [2]. Without adding additional sign restrictions, this would violate the uniqueness property of an identification condition.

It is not hard to add restrictions for the signs to pin down this remaining bit of unidentification in theory. For example, one can restrict all the sample factors means ($\frac{1}{T}\sum_t \widehat{f}_{t(\Gamma)}$) to bebbbb positive (call this [$3'$]). Alternatively, restricting the first non-zero element in each column of $\Gamma$ to be positive works as well (call this [$3''$]). Similar restrictions are seen in Stock and Watson (2002), Bai and Ng (2013).

However, we report that the sign issue is trickier in finite sample simulations. For example, if the true factors has zero (or close to zero) expectations ($\mathbb{E}\left[f_t^0\right] = \mathbf{0}$). Then, even if the true factors are observed, its finite sample averages are arbitrarily positive or negative, making [$3'$] unstable. Even if $\widehat{\Gamma}$ is estimated rather accurately, the small sign flipping of the factor mean makes a large difference in $\widehat{\Gamma} - \Gamma^0$, keeping it away from converging. Similarly, [$3''$] also runs into finite sample problems if $\Gamma^0$'s first non-zero elements in some columns are close to zero.

So how to design a sign restriction such that the signal-to-noise ratio in picking the sign is always maximized, adapting to any potential peculiar model? One way is to make $\Gamma$'s signs always align with those of $\Gamma^0$. So let [$3'''$] be $\Gamma$ s.t. $\Gamma_k^\top \Gamma_k^0 > 0, \forall k$. Then, [$3'''$]'s obvious problem is it depends on population information ($\Gamma^0$), disqualifying it as an estimation's identification condition.

Finally, we give a sign restriction that is both theoretically sound and practically easy to use in simulation exercises. Let [3] be the set of $\Gamma$ s.t. $\Gamma_k^\top \widehat{\Gamma}_k > 0, \forall k$, where $\widehat{\Gamma} = \arg\min_{\Gamma \in [1],[2],[3'']} G(\Gamma)$. The idea is for the $\Gamma$ within the target minimizing set, we use [$3''$] (or [$3'$]) to pin down its signs. That gives the unique estimate. And for

Electronic copy available at: https://ssrn.com/abstract=2983919

$\Gamma$ outside of the set, we align its signs to the estimate. This will normalize the true towards the direction of the estimate when constructing the reference point $\Gamma^0(\Theta)$.

This avoids the sign indeterminacy problem within a small neighborhood around $\widehat{\Gamma}$. The benefit is when calculating estimation error, $\left\|\widehat{\Gamma} - \Gamma^0(\Theta)\right\|$ always take the smallest possible value among all possible sign combinations. To calculate that quantity in simulation exercises, one can first ignore the sign issue and calculate $\widehat{\Gamma}$ and $\Gamma^0(\Theta)$ both up to sign unidentification. Then pick the signs to minimize that norm between the two:

$$\min_{s_1 \ldots s_K = \pm 1} \left\|\widehat{\Gamma} - \Gamma^0(\Theta)\mathrm{diag}\left\{s\right\}\right\|. \tag{114}$$

# E International Macroeconomic Instruments Data

There are fifty-nine indicators (besides GDP growth itself) that we could use as instruments. Because we require a country-year observation to have GDP growth and all its instruments non-missing, we first restrict attention to only indicators that are 80% nonmissing in the panel; we choose 80% as a cut-off so as to be able to include import and export percentage, which are very natural instruments to include. We then drop indicators that *a priori* appear unconnected to global growth: adolescent fertility rates, fertility rates, life expectancy, mortality rates, total population, and surface area. We drop gross national income and gross national income per capita because of their high correlation with GDP and GDP growth, respectively. We drop Guinea, Haiti, and Tanzania because they don't have non-missing GDP growth before 1985. Because we want to include exports/imports as instruments, we drop Congo, Ethiopia, Paraguay, and Zambia because they don't have non-missing import/export data before 1985. We want to include Germany in our analysis and it has no CO2 emissions data before 1985, so we drop that indicator. We want to include Hong Kong in our analysis and it has no domestic credit data before 1985, so we drop that indicator. We want to include Belgium in our analysis and it has no population density data before 1985, so we drop that indicator. We want to include capital formation but Luxumbourg and Cabo Verde have no data before 1985, so we drop those countries. We drop the percentage of merchandise trade given its high correlation with imports and exports which we do include. Urban population growth is quite correlated with population growth, so we exclude the former. This leaves us with the indicators

Electronic copy available at: https://ssrn.com/abstract=2983919

discussed in the text, combined with an Industrial dummy. We then restrict attention to the available countries and Kose et al. (2012) labeled as Industrial and Emerging, leaving the countries in Table 3.

Table 3: List of Countries

Argentina
Australia
Austria
Belgium
Brazil
Canada
Chile
China
Colombia
Denmark
Egypt, Arab Rep.
Finland
France
Germany
Greece
Hong Kong SAR, China
Iceland
India
Indonesia
Ireland
Israel
Italy
Japan
Jordan
Korea, Rep.
Malaysia
Mexico
Morocco
Netherlands
New Zealand
Norway
Pakistan
Peru
Philippines
Portugal
Singapore
South Africa
Spain
Sweden
Switzerland
Thailand
Turkey
United Kingdom
United States
Venezuela, RB

Electronic copy available at: https://ssrn.com/abstract=2983919