Instrumented Principal Component Analysis *

Bryan Kelly — Yale, AQR, NBER
Seth Pruitt — Arizona State University
Yinan Su — Johns Hopkins University

December 8, 2019

Abstract

We propose a new latent factor model with dynamic factor loadings. Additional data to each panel unit of observation instruments for its factor loadings according to a common and constant mapping. Instrumented Principal Components Analysis (IPCA) estimates the linearly parameterized mapping together with the latent factors by least squares similar to PCA. The estimation is consistent and asymptotically normal under general identification normalizations and flexible data generating processes for large panels. Compared to existing methods, IPCA is statistically more parsimonious while utilizing a wealth of additional instrumenting information, improving the identification of the factors, factor loadings, and their economic relationships with the instruments. The advantages are demonstrated with simulated data and applications to equity returns and international macroeconomics.

***PRELIMINARY AND INCOMPLETE.** 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 seminar and conference participants at ASU and Duke, Lars Hansen, Dacheng Xiu, Jonathan Wright, Yuehao Bai, and Federico Bandi. We thank Rongchen Li for excellent research assistance.

1 Introduction

Factor models are empirical workhorses in economics and finance as well as other social, biological, and physical sciences. The essence of a factor model is that variation in an individual unit of observation is well described with two ingredients – a loading on common factors, and an "idiosyncrasy" that drives the remaining individual-level variation. The factor model emphasizes parsimony: the number of factors is typically very small, while the panel data dimensions are large.

Our main idea is that the relationship between a panel unit and common factor can be directly analyzed using additional relevant data. The reality of "big data" implies that a wealth of additional information is available for this purpose – using this data to instrument for factor loadings helps us better understand the model's economic content.[1] By imposing a linear structure on how this additional information enters the problem, we arrive at a model that is easily estimated by least squares and describes panel data whose relationship to underlying latent factors varies in multiple dimensions.

For concreteness, think of $t = 1, \ldots, T$ representing time and $i = 1, \ldots, N$ representing individuals. This paper studies a "dynamic" model in which loadings are allowed to vary in both the i and t dimensions:

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

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

Equation (1) is the factor structure, where $x_{i,t}$ is scalar panel data, the target of factor analysis, f_t and $\beta_{i,t}$ are K factors and factor loadings, $\mu_{i,t}$ is the idiosyncratic error. Equation (2) links information from an L-vector of "instrumental" variables $c_{i,t}$ to the dynamic factor loading $\beta_{i,t}$. The key restriction giving content to this model is that the mapping from instruments to loadings, parameterized by the $L \times K$ matrix Γ, applies universally across i and t. This restriction allows identification about the relationship between panel units and latent factors, while simultaneously estimating the latent factors themselves. Loading 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 instrument and errors, similar to the exclusion restriction in IV, guarantees consistency.[2] In the following analysis, we often pool

[1] We use the terminology "instrument" recognizing that IPCA is to PCA as IV regression is to regression, as well as that in the generalized method of moments literature (Hansen 1982), instrumenting information is employed to represent conditional moment conditions.

[2] See Assumption 1.

the two errors together as $e_{i,t}$, yielding an equivalent representation:

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

Instrumented Principal Components Analysis (IPCA) is to estimate the dynamic factor model (1)-(2). It recovers the small K-factor structure parameterized by true Γ^0 and f_t^0 from a large N, T panel data consists of $x_{i,t}$ augmented by instrument data $c_{i,t}$ of a finite length L.

To understand why studying (1)-(2) is useful, first consider instead a "static" factor model of the form

$$x_{i,t} = \beta_i^\top f_t + \mu_{i,t}. \tag{4}$$

This model is widely used, one reason being its ease of estimation via principal components analysis (PCA) .[3] PCA puts no restriction on β_i but relying on it being static, not varying over t.

Yet factor models with variable loadings are common in many settings. For example, option "Greeks" are sensitivities with respect to underlying risks calculated from observables such as moneyness and maturity. For stock returns, Fama French (1993) reads "[observables] such as size and book-to-market equity, must proxy for sensitivity to ... risk factors in returns." These observables are all time-varying, resulting in dynamic loadings. As another example, some macroeconomic models describe inter-firm trade as interaction of nodes on networks, and 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 et al 2012). The essence of these circumstances embedd in (2) is that each individual's (option/stock/firm's) constant identity looses its relevance. Instead, a set of observable time-varying characteristics dictates the individual's behavior – its exposure to common factors.

As another comparison, one could estimate (1) without using (2) by making $\beta_{i,t}$ into a latent dynamic process in a nonlinear state-space framework (e.g. Primiceri, 2005; Pruitt, 2012; Del Negro and Otrok, forthcoming).

So what really are the advantages of our main idea embodied in (2)? Most prominently, economic theory may suggest relevant variables that IPCA can utilize to more accurately estimate the factors and factor loadings. On the other hand, IPCA can make direct eco-

[3]Stock and Watson (2002) discusses the computational advantages of estimating (4) via PCA instead of a linear state space model using MLE. The model we discuss and solve by least squares methods could also be cast in a linear state space model with particular structure on the observation equation (see below), but we leave that for future work. PCA is often derived from the eigenvalue decomposition (see Anderson 2002) of $YY^\top$, but the SVD of Y is directly related and often computationally quicker.

nomics statements about the importance of any instrumenting variable, paving the way for an economic narrative of why the panel unit has the relationship to the factors, and why the relationship varies over i and t.

The above point says IPCA is more flexible due to its time-variability and increased economic content from instrumental information. However, from a statistical point of view, IPCA is in fact often much more parsimonious. To see this, notice PCA's β's dimensionality increases with the size of the cross-section N, while IPCA's β is parameterized by the mapping matrix Γ, whose size is controlled by L. And, it tends to be the case that $L \ll N$. This allows the analysis of very large cross-sectional panels, where PCA suffers over-fitting, and latent β models also run into computational issues.

A third important advantage is that incorporating instrumenting information *gives economic interpretability* to the estimates of latent factors.[4] This interpretability follows because the k^{th} column of Γ tells us which instrumenting variables are important to the loading on factor k, and hence those instruments describe the factor's economic character.[5]

Fundamentally, (1)-(2) are designed to take a step forward in economic factor modeling. A large literature in economics has moved beyond using factor analysis merely as dimension-reduction, instead reaching for a better economic understanding of how hundreds or thousands or millions of individual variables are related to underlying aggregate forces. Factor relationships are implied by many economic theories; hence why Geweke (1977) and Sargent and Sims (1977) prominently advocated their use. "In many instances economic theory suggests that relationships will change over time" (Cooley and Prescott 1976); hence why recent advances in state-space models with time-varying parameters are shedding new light. *We would like to know why these relationships vary across time and individuals.* To understand why, our main idea is to *bind the relationships themselves to data that varies across time and individuals.* To be sure, it is a constraint to impose (2) instead of estimating each static relationship as a free parameter or each dynamic relationship as a latent process. But with that constraint, and the tests it affords, comes a broad avenue of economic

[4] A fundamental lack of factor interpretability has always been well-understood and dealt with in various ways. For instance, Stock and Watson (2002b) adopted the approach of showing how a factor's estimated static loadings related to an *a priori* grouping of the variables comprising y (in their application, a collection of macroeconomic aggregates from output, price, interest rate, trade, and money-supply categories). This approach is harder to implement when y is full of the same type of variable, for instance if y were a cross-section of stock returns. Then the natural way to understand factors would be to try and distinguish companies from another – exactly what IPCA, using companies' distinguishing features, is doing.

[5] Here is a simple analogy for what we mean. Fama and French (1993) constructed the factor-mimicking portfolios SMB and HML by sorting stock returns based on market capitalization and book-to-market, respectively. The SMB factor is a *size* factor and HML is a *value* factor: their economic character is given by the manner in which stocks were sorted, and how that sorting variable has some economic content. IPCA essentially adds this insight into a latent factor model.

discovery.

That concludes motivating the model and we start summarizing the main methodological findings and contributions of the paper.

Estimation: IPCA is a least squares estimation of the dynamic factor model (1)-(2). It minimizes the sample sum of squared errors ($e_{i,t}$ as in equation 3) over parameters $\{\Gamma, f_t\}$. This optimization is similar to PCA which minimizes over $\{\beta_i, f_t\}$. The minimization is numerically solved by an Alternating Least Squares method, by iterating on the objective function's first-order conditions with respect to Γ and f_t respectively, both are simply linear regressions. This procedure converges fast and requires no sophisticated numerical algorithms – only the ability to run regressions and recognize when regression estimates have converged. An additional benefit is dealing with unbalanced panels as easy as OLS does, which is critical for large panel applications.

The least squares optimization can be flexibly expanded with nested constraints, yielding additional estimations under different economic hypotheses. For example, restricting the lth row of Γ to be all zeros reveals the marginal contribution of the lth instrument to loading dynamics. Restricting part of the K factors to be some observable time-series tests the relevance of those series to the comovement of the panel. As a special case, restricting a constant factor tests the no-arbitrage condition in equity return applications. Kelly et. al. (2019) demonstrates these extension.

In the special case that $c_{i,t}$ is constant over t, IPCA estimation reduces to the SVD of the interaction of $x_{i,t}$ and c_i. It is an instance of Fan, Liao, and Wang's (2016) projected principal component analysis. Moreover, we show that even with $c_{i,t}$ varies over time, if $C_t^\top C_t$ is constant, then IPCA is still estimated by this SVD.[6]

Rotational Identification: This paper proposes a general method to deal with the rotational unidentification issue in factor analysis, leading to a structural understanding of normalization choice on estimation and its asymptotic errors.

Rotational unidentification is a known issue in factor analysis: loadings and factors cannot be identified together – simultaneously "rotating" the two does nothing to the data generating process or the sample fit. Hence, parameter identification proceeds in two steps. First is narrowing down to a set of parameters of best fits. All parameters in this set are rotations of each other, all have equally the best sample fits (least squares), and all parameterize the same data generating process. In the second step, econometricians pick a particular parameter in the set to "represent" the set. The rule to pick, called normalization, is often based on economic interpretability or computational elegance depending on the application.

[6]The special cases of constant $C_t^\top C_t$ is studied in a previous version of the paper. Here, C_t is the $N \times L$ matrix of stacked $c_{i,t}$.

We recognize that the choice of normalization rule is not unique, and develop general results irrelevant of the specific normalization. We still provide several specific examples. Given the general results, deriving the asymptotics for the specific cases is only a matter of algebraic calculation.

A prominent benefit of the framework is a unified understanding of the sources of parameter estimation error – a clear linear decomposition into the aforementioned two steps is attained at the large panel limit (Theorem 2). A peculiar scenario arises if the normalization depends on the random sample, of which the conventional PCA-like normalization is a case. In this scenario, the errors inherited from the second normalization step can be greater by a magnitude than the first optimization step, reflecting a strong contamination to the parameter estimation from random sample-based normalizations. To a large extend, this contamination is not a problem however. If we also normalize the true parameter *in the same way as the estimator*, and measure the estimation error against this normalized true, only the error from the first step exists.[7] Estimation error measured in this fashion, net of the normalization contamination, is more revealing about the accuracy of IPCA in recovering the true model, and hence should be based on when constructing tests of hypotheses about the true.

This framework of dealing with the rotational identification is general to other factor analysis methods.

Asymptotics: We derive the rate of convergence and the limiting distributions of the estimated factors and the mapping matrix Γ. The asymptotics are developed within the framework of large cross sections (N) and a large time dimension (T).

Blessed by the parsimonious parameterization, Γ estimation can utilize the wealth of cross-sectional information, and the estimation error convergences at the rate of $\sqrt{NT}$, net of the normalization contamination. I.e., this is the error relative to the normalized true (normalized in the same way as the estimator). This is faster by an order of magnitude than PCA loading's rate of $\sqrt{T}$, since it relies on individual-by-individual time-series regressions.[8] This convergence rate does not depend on the specific normalization choice, but the asymptotic distribution does.

A PCA-like normalization, involving diagonalizing factors' the second moment matrix, would introduce a rotational contamination with convergence rate $\sqrt{T}$. The intuitive reason is the diagonalizing procedure relies only on time-series sample averaging. This means the

[7] Bai (2003) is effectively doing this when deriving the PCA's asymptotic error. As discussed, knowing the normalized true is often as good as knowing the true itself. Our contribution is to explicitly express how to conduct the normalization to both the true and estimation, and how the normalization choice affects asymptotics for any normalization. See details in Section 5.

[8] This and the following PCA results for comparison are all from Bai (2003)

dominant term of Γ estimator's randomness is from the normalization contamination, a peculiar situation as mentioned above.

Factor's convergence rate is $\sqrt{N}$ (net of contamination), the same as PCA. This is because even if loadings are observed, estimating f_t relies on cross-sectional linear regression, whose accuracy is bounded at $\sqrt{N}$. IPCA's improved factor estimation is not revealed form this comparison. Again, the normalization contamination can add a term of rate $\sqrt{T}$ to f_t estimation as well.

2 The Data Generating Process

We provide a fundamental construction by generalizing the stochastic process to have large cross-section that we call *stochastic panels*. This is helpful to be precise about the meaning of a random variable with i, t subscripts, the common information denoted by random variables with only a t subscript, the various convergence results, and the idea that each individual (i) is i.i.d. conditional on common information. Alternatively, one could avoid the construction by stating these results as high-level assumptions. But we believe this construction is more rigorous and valuable to other problems.

2.1 Stochastic Panel

To define a stochastic panel, let there be a probability space $\{\Omega, \mathfrak{F}, Pr\}$, a random variable (vector) X, and two transformations $\mathbb{S}_{[d]} : \Omega \to \Omega$ on two directions $d \in \{[\text{cs}], [\text{ts}]\}$, satisfying the following conditions.

1. **measurable** Both the transformations are measurable against $\mathfrak{F}$. I.e., $\forall \Lambda \in \mathfrak{F}$, $\mathbb{S}_{[d]}^{-1}(\Lambda) \in \mathfrak{F}$.
2. **commutative:** $\mathbb{S}_{[d]} \circ \mathbb{S}_{[-d]} = \mathbb{S}_{[-d]} \circ \mathbb{S}_{[d]}$
3. **measure-preserving:** $\forall \Lambda \in \mathfrak{F}$, $Pr\left(\mathbb{S}_{[d]}^{-1}(\Lambda)\right) = Pr(\Lambda)$, for either d.
4. **jointly ergodic:** The two transformations are "jointly ergodic" in the sense that any event Λ s.t. $\Lambda = \mathbb{S}_{[d]}^{-1}(\Lambda), \forall d$ has $Pr(\Lambda) = 0$ or 1.
5. **cross-sectional exchangeable:** The sequence of random variables $X \circ \mathbb{S}_{[\text{cs}]}^{i-1}$ is exchangeable.

Next, we analyze the properties of a stochastic panel implied by these defining conditions.

Condition 1 is familiar in defining a stochastic process, though a stochastic process would need only one transformation.

Starting from an ω, the interpretation of $\mathbb{S}_{[\text{ts}]}(\omega)$ is "the next period"; the interpretation of $\mathbb{S}_{[\text{cs}]}(\omega)$ is "the next item". So condition 2 guarantees the recombining property that the next day's second observation is the same as second item's observation in the next day. This condition yields rich structures: Inductively, starting from a $\omega \in \Omega$, following the two transformations N and T times respectively, one can trace out a rectangle lattice of $N \times T$ sample points, as seen in Figure 1. This provides a definition of the sample panel conventionally represented with double subscripts:

$$X_{i,t} := X \circ \mathbb{S}_{[\text{cs}]}^{i-1} \circ \mathbb{S}_{[\text{ts}]}^{t-1}, \qquad \forall i = 1, \cdots, N,\ t = 1, \cdots, T. \tag{5}$$

Furthermore let $\mathfrak{F}_{[d]} \subset \mathfrak{F}$ be the collection of invariant events under transformations $\mathbb{S}_{[-d]}$. Easy to show both $\mathfrak{F}_{[d]}$ are sub-σ algebras of $\mathfrak{F}$. Let $X_{[d]}$ be $\mathfrak{F}_{[d]}$ measurable random variables. Easy to show $X_{[d]} = X_{[d]} \circ \mathbb{S}_{[-d]}$. That is to say $X_{[d]}$ is constant in the other direction. This provides a definition of random variables that are conventionally represented with a single subscript.

$$\left(X_{[\text{ts}]}\right)_t := X_{[\text{ts}]} \mathbb{S}_{[\text{ts}]}^{t-1}, \qquad t = 1, \cdots, T, \tag{6}$$

$$\left(X_{[\text{cs}]}\right)_i := X_{[\text{cs}]} \mathbb{S}_{[\text{cs}]}^{i-1}, \qquad i = 1, \cdots, N. \tag{7}$$

For example, the factor process $\{f_t\}$ is $\mathfrak{F}^{[\text{ts}]}$ measurable, i.e. invariant to $\mathbb{S}_{[\text{cs}]}$. Hence, the i subscript is redundant and dropped.

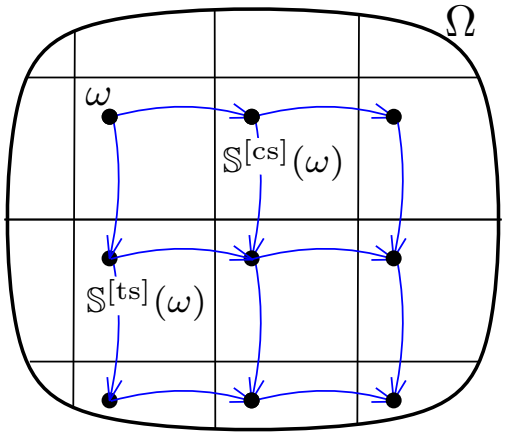

Figure 1: Stochastic panel from a lattice of sample points

The horizontal blue arrows are $\mathbb{S}_{[\text{cs}]}$, vertical blue arrows are $\mathbb{S}_{[\text{ts}]}$. One can think the square blocks are partitions in $\mathfrak{F}$. The columns are in $\mathfrak{F}^{[\text{ts}]}$, The rows are in $\mathfrak{F}^{[\text{cs}]}$.

Condition 3 implies each direction itself defines a stationary stochastic process in the

traditional sense. Stationarity implies a one-direction law of large number:

Lemma 1 (One directional LLN)**.** *Given conditions 1, 2, 3, if the moments exist,*

$$\operatorname*{plim}_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} X_{i,t} = \mathbb{E}\left[X_{\cdot,t}\middle|\mathfrak{F}^{[ts]}\right], \qquad w.p.\ 1, \forall t, \tag{8}$$

$$\operatorname*{plim}_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} X_{i,t} = \mathbb{E}\left[X_{i,\cdot}\middle|\mathfrak{F}^{[cs]}\right], \qquad w.p.\ 1, \forall i. \tag{9}$$

Proof. □

Notice the right-hand sides are $\mathfrak{F}_{[d]}$ measurable. It means, for example, the cross-sectional average convergences to some time-specific common information, usually represented with a single-t subscript.

Condition 4 implies a single-subscript random variables converge to the its unconditional expectation. These convergence results are the foundation of frequentist inference for stochastic panels.

Lemma 2 (Single-subscript LLN)**.** *Given conditions 1 to 4, if the moments exist,*

$$\operatorname*{plim}_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \left(X_{[cs]}\right)_i = \mathbb{E}\left[X_{[cs]}\right], \qquad w.p.\ 1, \tag{10}$$

$$\operatorname*{plim}_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \left(X_{[ts]}\right)_t = \mathbb{E}\left[X_{[ts]}\right]. \qquad w.p.\ 1. \tag{11}$$

Proof. □

Condition 5 is the only one not symmetric in the two directions. The general meaning is the sequence of the i in a sample does not matter for the joint distribution. The implications help clarify the various i.i.d. concepts. First, an $\mathfrak{F}^{[cs]}$-measurable process, e.g the fixed loadings β_i in a PCA setup, is i.i.d. cross-sectionally. Second, the exchangeable cross-section can be seen as mixtures of i.i.d. cross-sections conditional on $\mathfrak{F}^{[ts]}$ (??Hewitt and Savage 1955??). That is to say, for example, $c_{i,t}$ is cross-sectionally i.i.d. conditional on $\mathfrak{F}^{[ts]}$, the "rather precise" time-series information, but unconditionally not independent (only exchangeable) in the cross-section.

2.2 Specific Assumptions

With the construction above, we can be precise about the IPCA model constructions and assumptions.

Let primitive random variables c, μ, η, f mentioned in model (1)-(2) or (3) be generated from a stochastic panel. Among them, f is $\mathfrak{F}^{[\text{ts}]}$-measurable (no i subscript). Those equations, not repeated here, hold state-by-state, defining e, β, and x accordingly.

We refer to the true $\{\Gamma, f_t\}$ with the zero superscript as $\{\Gamma^0, f_t^0\}$, and use the letters without the zero superscript for generic parameters.

We put the following assumptions on the primitive random variables.

Assumption 1 (Instrument orthogonal to error)**.**

$$\mathbb{E}\left[c_{i,t}^\top e_{i,t} \middle| \mathfrak{F}^{[ts]}\right] = \mathbf{0}_{L\times 1} \tag{12}$$

This assumption can be implied by assuming the two primitive errors are orthogonal:

$$\mathbb{E}\left[c_{i,t}^\top \eta_{i,t} \middle| \mathfrak{F}^{[\text{ts}]}\right] = \mathbf{0}, \qquad \mathbb{E}\left[c_{i,t}^\top \mu_{i,t} \middle| \mathfrak{F}^{[\text{ts}]}\right] = \mathbf{0} \tag{13}$$

Assumption 2 (Moments)**.** (1) $\mathbb{E}\left\|f_t^0 f_t^{0\top}\right\|^2$ *exists.* (2) $\mathbb{E}\left[\left\|C_{i,t}e_{i,t}\right\|^2 \middle| \mathfrak{F}^{[ts]}\right]$, $\mathbb{E}\left\|\Omega_t^{ce}\right\|$ *exist.* (3) $\mathbb{E}\left[\left\|c_{i,t}^\top c_{i,t}\right\|^2 \middle| \mathfrak{F}^{[ts]}\right]$, $\mathbb{E}\left\|c_{i,t}^\top c_{i,t}\right\|^2$ *exist.* (4) $\mathbb{E}\left[\left\|c_{i,t}^\top c_{i,t}\right\|^2 \left\|f_t^0\right\|^2 \middle| \mathfrak{F}^{[ts]}\right]$, $\mathbb{E}\left[\left\|c_{i,t}^\top c_{i,t}\right\|^2 \left\|f_t^0\right\|^2\right]$ *exist.*

The following two assumption guarantees that matrices like $\Gamma^\top C_t^\top C_t \Gamma$, which appears frequently as part of the cross-sectional projection, is uniformly bounded and stays away from singularity.

Assumption 3. *The parameter space* Ψ *of* Γ *is compact and away from rank deficient, i.e.:* $\det \Gamma^\top \Gamma > \epsilon$ *for some* $\epsilon > 0$*.*

Assumption 4. *Almost surely,* $C_{i,t}$ *is bounded and* $\det \Omega_t^{cc} > \epsilon$ *for some* $\epsilon > 0$*.*

Assumption 5 (Central Limit Theorem)**.**

$$\frac{1}{\sqrt{NT}} \sum_{i,t} \text{vect}\left(C_{i,t}^\top e_{i,t} f_t^{0\top}\right) \xrightarrow{d} \text{Normal}\left(0, \Omega^{cef}\right), \tag{14}$$

where $\Omega^{cef} := \mathbb{V}\text{ar}\left[\text{vect}\left(c_{i,t}e_{i,t}f_t^{0\top}\right)\right]$.

3 Estimation and Normalization

3.1 Estimation as Optimization

Estimation is to solve an optimization problem

$$\min_{\Gamma,\{f_t\}} \sum_t \|x_t - C_t\Gamma f_t\|^2 . \tag{15}$$

It says, given a sample of $\{x_t, C_t\}$, look for $\Gamma, \{f_t\}$ to minimize the sum of squared errors over all $\{i,t\}$ in the sample. This inherits PCA's sample sum of squared errors optimization problem.

Define target function

$$G(\Gamma) = \frac{1}{2NT} \sum_t \left\| x_t - C_t\Gamma \widehat{f}_{t(\Gamma)} \right\|^2 , \tag{16}$$

where $\widehat{f}_{t(\Gamma)}$ is the optimal f_t given a Γ:[9]

$$\widehat{f}_{t(\Gamma)} = \left(\Gamma^\top C_t^\top C_t \Gamma\right)^{-1} \Gamma^\top C_t^\top x_t. \tag{17}$$

This way, the f_t is concentrated out, and the optimization is over Γ only. We work on the asymptotics Γ estimation first, by analyzing the asymptotics of G, until section 5.4 comes back to f_t.

3.2 Numerical Method and Practical Issues

This subsection digresses to the numerical method to solve the optimization problem. Unlike PCA, IPAC optimization does not have a (quasi-)analytical solutions as an eigen-decomposition.[10]

We use an Alternating Least Squares (ALS) method for the numerical solution. Notice the joint target function (equation 15) is quadratic in either Γ or F when the other is given, allowing for analytical optimization solutions of Γ and F one at a time. Given Γ, we already

[9] $\frac{1}{NT}$ looks redundant now. As we will see later, dividing by NT is to make $G(\Gamma) = \mathcal{O}_p(1)$. The $1/2$ is to be canceled out later when taking derivatives.

[10] In a previous version of this paper, we work out a situation in which such eigen-decomposition is the solution when characteristics are standardized over time. Now we work with the general case. The special case is an approximated solution and can be used as the ALS's initial guess, as documented in the Internet Appendix B of ??FinancePaper??.

have f_t solved as $\widehat{f}_{t(\Gamma)}$. Symmetrically, given F, the optimizing Γ is

$$\arg\min_{\gamma} \sum_t \|x_t - C_t \Gamma f_t\|^2 = \left(\sum_{i,t} \left(C_{i,t}^\top \otimes f_t\right) \left(C_{i,t} \otimes f_t^\top\right) \right)^{-1} \left(\sum_{i,t} \left(C_{i,t}^\top \otimes f_t\right) x_{i,t} \right), \quad (18)$$

which is a panel OLS of $x_{i,t}$ onto $C_{i,t} \otimes f_t^\top$.[11] The ALS algorithm starts with an initial guess, alternates between updating Γ and F given the other, and stops when the first order condition is satisfied up to a tolerance. In practice, it converges after a few iterations in a matter of seconds.

Unbalanced panel

3.3 Normalization and Identification Condition

The rotational unidentification is a known issue in latent factor models. Specifically for IPCA, in the population, a class of models with true parameters $\{\Gamma^0 = \Gamma^* R, f_t^0 = R^{-1} f_t^*\}$ generate exactly the same data $\{x_{i,t}, c_{i,t}\}$, given any full rank $K \times K$ matrix R. The corresponding issue in sample is that for any Γ, $G(\Gamma) = G(\Gamma R)$. That means if $\widehat{\Gamma}^*$ solves $\min_\Gamma G(\Gamma)$, then any of its rotation $\widehat{\Gamma}^* R$ is also a minimizer.

We deal with the issue by following the convention of studying the asymptotics of the estimator after some unique normalizations. The idea is that it is impossible and of no use to distinguish among a class of models that are rotations of each other. While, one can picks a particular normalization to "represent" each class, and such pick is often of economics meanings or some nice properties depending on the application. See ??BaiNgRotation??.

Compared with the convention however, we provide a general method to characterize the asymptotics that does not depend on the specific normalization. This leads to a structural understanding on the effect of normalization choice on estimation error's asymptotic distribution. After the general result, we calculate the asymptotics under a few specific normalization examples.

Let parameter space Ψ be the set of all possible parameters Γ. A Γ is called (rotational) *unidentified* with Γ' if there exists a full rank matrix R s.t. $\Gamma = \Gamma' R$. 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 unidentified with Γ. Given an identification assumption, the mapping from any Γ to such a (unique) Γ' is called a *normalization*, denoted $\Gamma' = \mathbb{N}(\Gamma; \Theta)$. Notice, there is not a unique identification condition. As in Figure 2, there can be different Θ's that "cut through" all the dashed lines (representing sets of unidentified parameters) each only once. Here, we use Θ to represent a generic one.

[11] The arg min expression is for γ, the vectorized optimizing Γ, as $\gamma = \text{vect}(\Gamma)$.

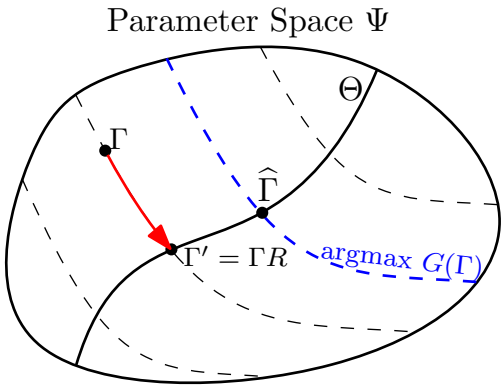

Figure 2: Identification Condition, Normalization, and Estimation

An identification condition Θ is a subset of the parameter space Ψ. The red arrow represents a normalization mapping. The dashed lines represent sets of unidentified parameters. One particular dashed line (in blue) represents set $\arg\max_\Gamma G(\Gamma)$, the parameters that minimize the target function. The intersection of sets Θ and $\arg\max_\Gamma G(\Gamma)$ defines the estimator $\widehat{\Gamma}$.

The purpose of constructing the identification condition is each element in it serves as a unique representation of a class of unidentified Γ's. There are two aspects, unique representation of the sample estimate as well as the true model, which are discussed below and in the next subsection respectively.

The estimator Γ^0 is defined as the minimizer of the target function $G(\Gamma)$ within an identification condition Θ:

$$\widehat{\Gamma} = \arg\min_{\Gamma \in \Theta} G(\Gamma) \tag{19}$$

To make a practical estimation, the Θ here must be known to econometricians, i.e. it can be sample-dependent (random), but cannot dependent on the underlying true parameters. Given one estimation, one can easily find another under a different identification condition following the associated normalization function. Next, we define two specific identification conditions and discuss their interpretations.

Define $\Theta^I = \{\Gamma \in \Psi \mid Block_{1:K}(\Gamma) = \mathbb{I}_K\}$, where $Block_{1:K}(\cdot)$ is an function of a matrix that cuts out its first K rows. The corresponding normalization $\mathbb{N}(\Gamma; \Theta^I) = \Gamma Block_{1:K}(\Gamma)^{-1}$.

In general, Γ has LK degrees of freedom in total, with K^2 degrees unidentified, since the rotation R is $K \times K$. So the identification conditions should restrict K^2 degrees of freedom. This identification conditions pins down the first $K \times K$ entries, and leaves the lower $L - K$ rows free.

The interpretation is a correspondence between the K factors and the first K characteristics. It says, the loading on the first factor is (one-for-one) driven by the first characteristic,

and irrelevant of the second to the K'th characteristics. Same for the rest of the loadings. Then, the estimation asks how each of the next $L-K$ characteristics can additionally contribute to the first K "pure" loadings. The condition imposes the first K characteristics to be independently useful, and be open about whether the additional $L-K$ can add power, and to which factor the power is added. The correspondence gives each factor a interpretable meaning.

This condition is similar to 2.3 in ??Bai and Ng 2008??, but using characteristics makes the meaning of factors clearer. Variations are easily achieved by changing the block of the first K rows to a sequence of any K rows, which assigns new characteristics correspondence to the factors. The estimation error of this case will be specifically calculated in subsection 5.3.1.

Define $\Theta^V_{N,T}$ as the set of $\Gamma \in \Psi$ such that

[1] Γ is ortho-normal: $\Gamma^\top \Gamma = \mathbb{I}_K$,

[2] $\widehat{f}_{t(\Gamma)}$ is orthogonal: $\frac{1}{T}\sum_t \widehat{f}_{t(\Gamma)}\widehat{f}^\top_{t}(\Gamma)$ is diagonal, with distinct and descending entries.[12]

This is the normalization used in ??FinancePaper??. It follows the PCA convention, e.g. ??Stock and Watson??, to pick a set of orthogonal f_t to represent the factor space.

Notice $\Theta^V_{N,T}$ depends on the sample (hence the N,T subscripts), because $\widehat{f}_{t(\Gamma)}$ depends on the sample. This is contrary to Θ^I.

3.4 Identification Condition on True Parameter: Γ^0 and $\Gamma^0_{N,T}$

This subsection explains the two different true parameter constructions due to the sample-dependency in normalization: the random $\Gamma^0_{N,T}$ used to measure estimation error against, and the deterministic Γ^0 used as a fixed reference point in asymptotic analysis. Figure 3 is a helpful illustration.

A set of unidentified true parameters all represent the same model. Against which one should we measure the estimation error, given we have made an identification choice on the estimate? Naturally, we measure against the true under the same identification condition as the estimate. For example, given estimation $\widehat{\Gamma} \in \Theta^V_{N,T}$, we make an identification assumption to pick the true $\Gamma^0_{N,T} \in \Theta^V_{N,T}$ too, and study the asymptotics of $\widehat{\Gamma} - \Gamma^0_{N,T}$ to infer about estimation accuracy (arrow B in Figure 3).

This leads to the somewhat peculiar situation that the true parameter $\Gamma^0_{N,T}$ can be sample-dependent (i.e. random). The data generating process is not sample-dependent of

[12] To pin down the signs indeterminacy, we restrict the sample mean of each factors to be positive. There are other ways to do this, depending on application. See (**?**) for related discussion.

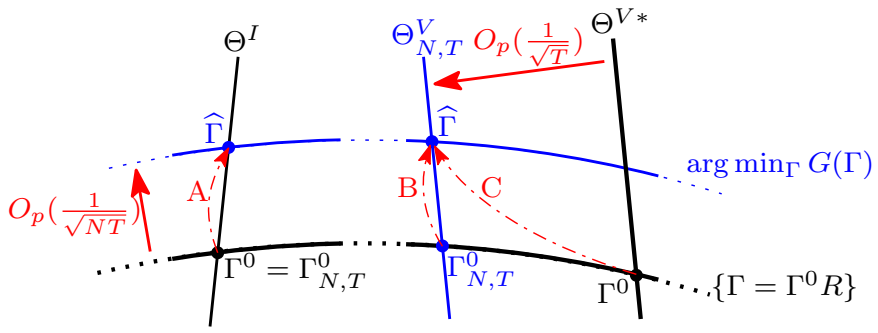

Figure 3: Estimators and True Parameters under Different Identification Conditions

The top horizontal curve is the set of unidentified estimates. The bottom curve is the set of unidentified true parameters. The three vertical curves are identification conditions. Black items are deterministic, blue are sample-dependent. The two solid red arrows points to the the random sets from their deterministic limits, marked with the stochastic order of the difference. The three dashed red arrows mark three pairs of estimation errors.

course. Just its parametric representation ($\Gamma^0_{N,T}$) inherits the randomness of the normalization method Θ.

However, we need a deterministic Γ^0 to define the data generating process (in section **??**) and to conduct asymptotic analysis (in next sections). For this purpose, assume there is a deterministic identification condition Θ^* which is the probability limit of Θ.[13] And we make the identification assumption that $\Gamma^0 \in \Theta^*$. Specifically, the deterministic limit of Θ^I is itself, so the two true constructions coincide: $\Gamma^0 = \Gamma^0_{N,T}$; the deterministic limit of $\Theta^V_{N,T}$ is Θ^{V*} (drawn in Figure 3 and written in Appendix **??**), yielding different Γ^0 and $\Gamma^0_{N,T}$.

To recapitulate this section, we make the following points. First, one can only estimate (as in equation 19) with $\Theta^V_{N,T}$ but not Θ^{V*}, because the later involves quantities about the true which are unavailable to the one making the estimation.

Second, we are mostly interested in estimation error $\widehat{\Gamma} - \Gamma^0_{N,T}$ (e.g. arrow B) rather than $\widehat{\Gamma} - \Gamma^0$ (arrow C). Because the first references against the true that is within the same identification condition, and hence informative about the distance between the estimated and true model. The second is "contaminated" by the randomness of the identification condition $\Theta^V_{N,T}$. Referencing against $\Gamma^0_{N,T}$ rather than Γ^0 is re-drawing the bull's-eye after the bullet hits the target — of course the accuracy is good. It avoids the inaccuracy of the random identification assumption, which is in fact the stochastic dominant term for the case of Θ^V (shown in Theorem **??**).

This construction follows ??Bai and Ng??, but is more explicit. The true factor f^0_t,

[13]The exact meaning of "probability limit of the set" is given in Condition **??** after defining the identification function.

which is orthogonal in population, would not be exactly orthogonal in a finite sample (a.s.). Each sample orthogonality normalization gives a different rotation of f_t^0, which they wrote as $R_{N,T}^{-1} f_t^0$. They focus on estimation errors like $f_t^0 - R_{N,T}^{-1} f_t^0$ and $\hat{\beta}_i - \beta_i^0 R_{N,T}$ (the second corresponds to our arrow B, just change β to Γ).

Third, the deterministic Γ^0 is the underlying reference point important for the analysis, although it is not explicit in the end results which are about $\widehat{\Gamma}$ and $\Gamma_{N,T}^0$. Both $\widehat{\Gamma}$ and $\Gamma_{N,T}^0$'s asymptotics are constructed with respects to Γ^0 next.

4 Consistency

The goal is to show $\text{plim}_{N,T\to\infty} \widehat{\Gamma} - \Gamma_{N,T}^0 = 0$. The strategy is, recognizing they are solutions to the first order condition and identification condition, studying the convergence of the functions in those conditions. This follows ??Newey McFadden??'s strategy analyzing a canonical M-estimator. The new challenges are simultaneous N, T convergence and the identification issues.

4.1 Score Function Expression

Define score function as the derivative of the target function,

$$S(\Gamma) = \frac{\partial G(\Gamma)}{\partial \gamma}.^{14} \tag{20}$$

The optimizer $\widehat{\Gamma}$ satisfies the first order condition: $S(\widehat{\Gamma}) = \mathbf{0}_{LK\times 1}$.

This subsection manipulates the score function to a form consists of primitives, in order to study its asymptotics. First, we have:

$$S(\Gamma) = \frac{1}{NT} \sum_t \left(C_t^\top \otimes \widehat{f}_{t(\Gamma)} \right) \left(x_t - C_t \Gamma \widehat{f}_{t(\Gamma)} \right) = \frac{1}{NT} \sum_t \text{vect} \left(C_t^\top \widehat{e}_{t(\Gamma)} \widehat{f}_t^\top{}_{(\Gamma)} \right). \tag{21}$$

where $\widehat{e}_{t(\Gamma)} = x_t - C_t \Gamma \widehat{f}_{t(\Gamma)}$. Here $\widehat{f}_{t(\Gamma)}$ and $\widehat{e}_{t(\Gamma)}$ are OLS coefficient and error of the cross-sectional regression x_t onto $C_t\Gamma$.

Define $\tilde{f}_{t(\Gamma)}, \tilde{e}_{t(\Gamma)}$ as the population counterparts of such cross-sectional regression. Per stochastic panel construction, conditional on $\mathfrak{F}^{[\text{ts}]}$ (knowing all common information), the

[14] Here, $\gamma = \text{vect}(\Gamma) = \text{vec}(\Gamma^\top)$ is Γ vectorized by going rows first. The Jacobian puts the derivatives in a column following the format of γ.

cross-sectional distribution is i.i.d. So, $\tilde{f}_{t(\Gamma)}$ is the OLS coefficient of this i.i.d. population:

$$\tilde{f}_{t(\Gamma)} = \mathbb{E}\left[\Gamma^\top c_{i,t}^\top c_{i,t}\Gamma \middle| \mathfrak{F}^{[\text{ts}]}\right]^{-1} \mathbb{E}\left[\Gamma^\top c_{i,t}^\top x_{i,t} \middle| \mathfrak{F}^{[\text{ts}]}\right] \tag{22}$$

$$= \left(\Gamma^\top \Omega_t^{cc} \Gamma\right)^{-1} \Gamma^\top \Omega_t^{cc} \Gamma^0 f_t^0 \tag{23}$$

where we defined shorthand $\Omega_t^{cc} := \mathbb{E}\left[c_{i,t}^\top c_{i,t} \middle| \mathfrak{F}^{[\text{ts}]}\right]$.[15] And

$$\tilde{e}_{t(\Gamma)} = x_t - C_t \Gamma \tilde{f}_{t(\Gamma)} = e_t + C_t \Pi_{t(\Gamma)} f_t^0 \tag{24}$$

where shorthand $\Pi_{t(\Gamma)} := \left(\mathbb{I}_L - \Gamma\left(\Gamma^\top \Omega_t^{cc}\Gamma\right)^{-1}\Gamma^\top \Omega_t^{cc}\right)\Gamma^0$. Notice $\Pi_t(\Gamma^0) = 0$, $\tilde{f}_t(\Gamma^0) = f_t^0$ and $\tilde{e}_t(\Gamma^0) = e_t$.

The above connects $\tilde{f}_{t(\Gamma)}, \tilde{e}_{t(\Gamma)}$ to f_t^0, e_t. The relationship between $\widehat{f}_{t(\Gamma)}, \widehat{e}_{t(\Gamma)}$ and $\tilde{f}_{t(\Gamma)}, \tilde{e}_{t(\Gamma)}$ is from standard OLS. Put together, we break down $\tilde{f}_{t(\Gamma)}, \tilde{e}_{t(\Gamma)}$ to primitives as:

$$\widehat{e}_{t(\Gamma)} = e_t + C_t \Pi_{t(\Gamma)} f_t^0 - C_t \Gamma \left(\Gamma^\top C_t^\top C_t \Gamma\right)^{-1} \Gamma^\top C_t^\top \tilde{e}_{t(\Gamma)} \tag{25}$$

$$\widehat{f}_{t(\Gamma)} = \tilde{f}_{t(\Gamma)} + \left(\Gamma^\top C_t^\top C_t \Gamma\right)^{-1} \Gamma^\top C_t^\top \tilde{e}_{t(\Gamma)} \tag{26}$$

Plug those back to the score. Each summand in equation (21) yields $3 \times 2 = 6$ terms by the distribution rule.

$$C_t^\top \widehat{e}_{t(\Gamma)} \widehat{f}_t^\top{}_{(\Gamma)} \tag{27}$$

$$= C_t^\top e_t \qquad \tilde{f}_t^\top{}_{(\Gamma)} \tag{28}$$

$$+ C_t^\top C_t \Pi_{t(\Gamma)} f_t^0 \qquad \tilde{f}_t^\top{}_{(\Gamma)} \tag{29}$$

$$- C_t^\top C_t \Gamma \left(\Gamma^\top C_t^\top C_t \Gamma\right)^{-1} \Gamma^\top C_t^\top \tilde{e}_{t(\Gamma)} \qquad \tilde{f}_t^\top{}_{(\Gamma)} \tag{30}$$

$$+ C_t^\top e_t \qquad \tilde{e}_t^\top{}_{(\Gamma)} C_t \Gamma \left(\Gamma^\top C_t^\top C_t \Gamma\right)^{-1} \tag{31}$$

$$+ C_t^\top C_t \Pi_{t(\Gamma)} f_t^0 \qquad \tilde{e}_t^\top{}_{(\Gamma)} C_t \Gamma \left(\Gamma^\top C_t^\top C_t \Gamma\right)^{-1} \tag{32}$$

$$- C_t^\top C_t \Gamma \left(\Gamma^\top C_t^\top C_t \Gamma\right)^{-1} \Gamma^\top C_t^\top \tilde{e}_{t(\Gamma)} \qquad \tilde{e}_t^\top{}_{(\Gamma)} C_t \Gamma \left(\Gamma^\top C_t^\top C_t \Gamma\right)^{-1}. \tag{33}$$

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

$$S(\Gamma) = \frac{1}{NT} \sum_t \text{vect}\left(S_t^{[1]}(\Gamma) + \cdots + S_t^{[6]}(\Gamma)\right). \tag{34}$$

[15] Othogonality condition Assumption **??** and **??** are used.

4.2 Score Function Convergence

We analyze the uniform probability limit of $S(\Gamma)$. Here, we provide a loose description of the rationale of the results. First, taking $N \to \infty$ at a fixed t, we have three modular results, $\frac{1}{N} C_t^\top e_t \to \mathbf{0}$, $\frac{1}{N} \Gamma C_t^\top \tilde{e}_{t(\Gamma)} \to \mathbf{0}$, and $\frac{1}{N} C_t^\top C_t = \mathcal{O}_p(1)$, by the LLN at the conditional i.i.d. cross-section. Plug these into the score function (equation 34), we find each term for $p = 1, 3 \sim 6$ have $\frac{1}{N} S_t^{[p]}(\Gamma) \to \mathbf{0}$, except for $\frac{1}{N} S_t^{[2]}(\Gamma) \to \Omega_t^{cc} \Pi_{t(\Gamma)} f_t^0 \tilde{f}_{t(\Gamma)}^\top$, which is $\mathfrak{F}^{[\text{ts}]}$ measurable as expected (Lemma 9) . Then, taking the time-series average yields the finite-T large-N convergence of score (line 35) in Proposition 1. Finally, taking $T \to \infty$ brings the score to the unconditional expectation, following an ergodic time-series LLN (line 36).

Proposition 1 (Uniform Convergence of the Score Function)**.**
Under Assumptions 2, 3, 4, the score function converges uniformly in probability:

$$\sup_{\Gamma \in \Psi} \|S(\Gamma) - S_T(\Gamma)\| \xrightarrow{p} 0, \qquad N \to \infty, \forall T, \tag{35}$$

$$\sup_{\Gamma \in \Psi} \left\|S(\Gamma) - S^0(\Gamma)\right\| \xrightarrow{p} 0, \qquad N, T \to \infty, \tag{36}$$

where $S_T(\Gamma) = \frac{1}{T} \sum_t \text{vect}\left(\Omega_t^{cc} \Pi_{t(\Gamma)} f_t^0 \tilde{f}_{t(\Gamma)}^\top\right), S^0(\Gamma) = \mathbb{E}\ \text{vect}\left(\Omega_t^{cc} \Pi_{t(\Gamma)} f_t^0 \tilde{f}_{t(\Gamma)}^\top\right).$

The rigorous proof is in Appendix A.2, which is much more involved. The complications include uniform convergence across Γ (which is necessary for solution convergence) and T-sums of large N limits. The general strategy is to use convergence in expectation rather than convergence in probability.

The next proposition says the limit of the score $S^0(\Gamma) = \mathbf{0}$ only when $\Gamma = \Gamma^0$ or rotations of Γ^0. This means the first order condition of a large sample can uniquely recover the true parameter, only up to the indeterminacy of the rotation. The proof is in Appendix A.4.

Proposition 2 ("Only" True Solves Score's Limit)**.**
Under Assumption 4, $S^0(\Gamma) = \mathbf{0}$ *if and only if* Γ *is unidentified with* Γ^0*.*

4.3 Identification Function and Estimator Convergence

Identification function is a construction to quantify and operationalize the identification condition. Define identification function $I(\Gamma)$ such that its solution is the identification condition, that is $\Theta = \{\Gamma \mid I(\Gamma) = \mathbf{0}\}$.

For Θ^I, a way to specify an identification function is

$$I(\Gamma) = \text{vect}\left(Block_{1:K}(\Gamma) - \mathbb{I}_K\right). \tag{37}$$

For $\Theta^V_{N,T}$, specify I as stacked by two parts corresponding to [1] and [2] in $\Theta^V_{N,T}$ definition with lengths $\frac{1}{2}K(K+1)$ and $\frac{1}{2}K(K-1)$ respectively:

$$I(\Gamma) = \begin{bmatrix} \text{veca}\left(\Gamma^\top\Gamma - \mathbb{I}_K\right) \\ \text{vecb}\left(I^{[2]}(\Gamma)\right) \end{bmatrix}, \tag{38}$$

where $I^{[2]}(\Gamma) = \frac{1}{T}\sum_t \widehat{f}_{t(\Gamma)}\widehat{f}^\top_{t(\Gamma)} - \widehat{V}^{ff}$.[16] Notice the second part is of sample data. Hence this I is a random function.

We provide the necessary conditions on a general identification function to imply Γ consistency. In Lemma **??**, we verify the two specific ones above satisfy the conditions.

Condition 1 (Necessary Conditions on Identification Function of Consistency)**.**

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

$$\sup_{\Gamma\in\Psi}\left\|I(\Gamma) - I^0(\Gamma)\right\| \xrightarrow{p} 0, \qquad N,T\to\infty, \tag{39}$$

2. *$I^0(\Gamma^0) = \mathbf{0}$.*

3. *I^0 uniquely identifies: If both $\Gamma^{[1]} \neq \Gamma^{[2]}$ solve $I^0(\Gamma) = \mathbf{0}$, then $\Gamma^{[1]}$ is not unidentified with $\Gamma^{[2]}$.*

Theorem 1 (Γ Estimation Consistency)**.** *If the preconditions of Propositions 1 and 2 are met, and the identification condition Θ has a proper identification function as in Condition 1, then $\widehat{\Gamma} \xrightarrow{p} \Gamma^0, \Gamma^0_{N,T} \xrightarrow{p} \Gamma^0$, which implies, $\widehat{\Gamma} - \Gamma^0_{N,T} \xrightarrow{p} \mathbf{0}$. as $N,T\to\infty$.*

Proof. We stack pairs of score and identification functions together and treat the stacked as a single function. We have by definitions, $\widehat{\Gamma}$ solves $[S;I](\Gamma) = \mathbf{0}$, Γ^0 solves $[S^0;I^0](\Gamma) = \mathbf{0}$, $\Gamma^0_{N,T}$ solves $[S^0;I](\Gamma) = \mathbf{0}$. Proposition 1 and Condition 1(1) implies the stacked function $[S;I]$ is uniformly convergent:

$$\sup_{\Gamma\in\Psi}\left\|[S;I](\Gamma) - \left[S^0;I^0\right](\Gamma)\right\| \xrightarrow{p} 0, \qquad N,T\to\infty \tag{40}$$

Proposition 2 and Condition 1(2) imply that $[S^0;I^0]$ has unique solution Γ^0. Then, by standard arguments , the solution of the uniformly converging function convergences to the limiting function's solution. We have $\widehat{\Gamma} \xrightarrow{p} \Gamma^0$.

[16] veca (A) stacks A's upper triangle entries, including the diagonal, in a vector by going right first, then down. vecb (A) stacks A's upper triangle entries, NOT including the diagonal, in a vector by going right first, then down. $\widehat{V}^{ff}$ is the sample factor second moment matrix of the econometrician's normalization discretion. As specified in $\Theta^V_{N,T}$, it is any diagonal matrix.

Of course, $[S^0; I]$ also uniformly converges to $[S^0; I^0]$. That give $\Gamma^0_{N,T} \xrightarrow{p} \Gamma^0$.
The two results combined imply $\widehat{\Gamma} - \Gamma^0_{N,T}$ converges to zero. □

5 Asymptotic Distributions

We derive the asymptotic distribution of two forms of estimation error: $\widehat{\gamma} - \gamma^0$ and $\widehat{\gamma} - \gamma^0_{N,T}$. The difference is against which true parameter representation the error is calculated.

The first form of "error", $\widehat{\gamma} - \gamma^0$, tells about the asymptotic distribution of the estimator $\widehat{\gamma}$, since γ^0 is given and deterministic. An interesting discovery is that, for sample-dependent identification conditions, $\widehat{\gamma} - \gamma^0$ depends on not just the random sample target function (its first order condition), but also the random sample identification conditions. More interestingly, the asymptotic randomness from the identification condition can sometimes be the dominate term in driving the asymptotic distribution of $\widehat{\gamma} - \gamma^0$. This case reflects a strong contamination of estimation accuracy by an improper choice of identification condition.

However, in some sense, such contamination is not problematic. The second form of error, $\widehat{\gamma} - \gamma^0_{N,T}$, informs about the accuracy of the estimation of the model, rather than a specific parametric representation of the model. When the goal is to distinguish the estimated and the true *models*, rather than to recover the deterministic representation γ^0, one would care about the asymptotic distribution of error in the this form. As we will see, $\widehat{\gamma} - \gamma^0_{N,T}$ does not depend on the asymptotic randomness of the identification condition, but only the randomness of the first order condition. The intuitive reason is both $\widehat{\gamma}$ and $\gamma^0_{N,T}$ are in the same (potentially sample-dependent) identification condition, so the contamination from which identification condition to use does not show up.

5.1 The Results for Generic Identification

The strategy follows ??Newey McFadden??'s analysis of a canonical M-estimator, linearizing the score function around γ^0. The innovations are about dealing with the identification conditions. We consider first the mean-value in between $\widehat{\Gamma} - \Gamma^0$. Later, the case of $\widehat{\Gamma} - \Gamma^0_{N,T}$ has the same derivation but one important difference. The two results are respectively summarized in Theorem 2.

According to Lagrange's Mean Value Theorem, there exist a $\bar{\gamma}$ in between $\widehat{\gamma}$ and γ^0 at every sample, such that

$$S(\widehat{\Gamma}) = S(\Gamma^0) + \left.\frac{\partial S(\Gamma)}{\partial \gamma^\top}\right|_{\gamma=\bar{\gamma}} \left(\widehat{\gamma} - \gamma^0\right). \tag{41}$$

Notice $S(\widehat{\Gamma}) = \mathbf{0}$. That means

$$\bar{H}\left(\widehat{\gamma} - \gamma^0\right) = -S(\Gamma^0), \tag{42}$$

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

Normally without the unidentification problem, one would divide through the Hessian to get $\widehat{\gamma} - \gamma = -\bar{H}^{-1}S(\Gamma^0)$ to represent the asymptotic distribution. But that is not possible here.

On the surface, the unidentification problem manifests as $\bar{H}$ being singular. To see why, notice the score should have zero gradient on the directions of unidentification. Since the rotation matrix R is $K \times K$, there are K^2 directions to marginally perturb Γ around a reference point without changing the score. That implies the score has zero gradient on those K^2 directions, meaning $\bar{H}$, although an LK square matrix, has a rank of only $LK - K^2$.

From another angle, the problem is the same non-unique solution situation, but now happens on the linearized equation (42) rather than the original non-linear equation $S(\widehat{\Gamma}) = \mathbf{0}$. There are a family of $\widehat{\gamma}$ on a K^2-dimension sub-linear space, that all solve equation (42).

Next, we complement the linearized "score = 0" equation with the "identification function = 0" equation linearized in the same fashion to pin down the indeterminacy. Conduct the same linearization on I, remember $I(\widehat{\Gamma}) = \mathbf{0}$ too:

$$I(\widehat{\Gamma}) = I(\Gamma^0) + \frac{\partial I(\Gamma)}{\partial \gamma^\top}\Big|_{\gamma=\bar{\gamma}}\left(\widehat{\gamma} - \gamma^0\right). \tag{43}$$

$$\bar{J}\left(\widehat{\gamma} - \gamma^0\right) = -I(\Gamma^0). \tag{44}$$

where $\bar{J}$ is defined as or the Hessian counterpart of the identification function. Notice I is $K^2 \times 1$ and J is $K^2 \times LK$, so this equation pins down K^2 more degrees of freedom.

Stack equations (42) and (44) vertically to form a single linear equation about the estimator,[17]

$$\left[\bar{H}; \bar{J}\right]\left(\widehat{\gamma} - \gamma^0\right) = -\left[S(\Gamma^0); I(\Gamma^0)\right]. \tag{45}$$

Now this linear equation set has a unique solution, because the stacked $\left[\bar{H}; \bar{J}\right]$ with size $(LK + K^2) \times LK$ is full rank LK. To solve it, left multiply $\left(\left[\bar{H}; \bar{J}\right]^\top \left[\bar{H}; \bar{J}\right]\right)^{-1}\left[\bar{H}; \bar{J}\right]^\top$ on both sides:

$$\widehat{\gamma} - \gamma^0 = -\left(\bar{H}^\top \bar{H} + \bar{J}^\top \bar{J}\right)^{-1}\left(\bar{H}^\top S(\Gamma^0) + \bar{J}^\top I(\Gamma^0)\right). \tag{46}$$

[17] [A; B] means vertical stack the two matrices.

With the stacking tricks above, the rest of asymptotics derivation is back to the canonical M-estimator case. According to Theorem 1, $\widehat{\gamma} \to \gamma^0$, at the $N, T \to \infty$ probability limit. That means $\bar{\gamma}$, the mean value between $\widehat{\gamma}$ and γ^0, goes to γ^0 as well. Define the probability limits of the Hessians, by the continuous mapping theorem:

$$\operatorname*{plim}_{N,T\to\infty} \bar{H} = H^0 := \left.\frac{\partial S^0(\Gamma)}{\partial \gamma^\top}\right|_{\gamma=\gamma^0}, \quad \operatorname*{plim}_{N,T\to\infty} \bar{J} = J^0 := \left.\frac{\partial I^0(\Gamma)}{\partial \gamma^\top}\right|_{\gamma=\gamma^0}.^{18} \tag{47}$$

Then, taking probability limit on equation (46) leads to the first line in Theorem 2.

The derivation of the second line is the same except an important difference. From the beginning of this subsection, apply the Lagrange's Mean Value Theorem between $\widehat{\gamma}$ and $\gamma^0_{N,T}$ instead of $\widehat{\gamma}$ and γ^0. Denote the new mean value as $\bar{\bar{\gamma}}$ and denote the Hessians evaluated at $\bar{\bar{\gamma}}$ as $\bar{\bar{H}}, \bar{\bar{J}}$, at each sample. The same steps lead to the counterpart of equation (46):

$$\widehat{\gamma} - \gamma^0_{N,T} = -\left(\bar{\bar{H}}^\top \bar{\bar{H}} + \bar{\bar{J}}^\top \bar{\bar{J}}\right)^{-1} \left(\bar{\bar{H}}^\top S(\Gamma^0_{N,T}) + \bar{\bar{J}}^\top I(\Gamma^0_{N,T})\right). \tag{48}$$

Also according to Theorem 1 not just $\widehat{\gamma}$ but also $\gamma^0_{N,T} \to \gamma^0$, hence the mean value $\bar{\bar{\gamma}}$ in between them converges to γ^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 γ^0.[19]

Importantly, the difference in this case is that $I(\Gamma^0_{N,T}) = \mathbf{0}$ by construction, which would eliminate a term in the second line of Theorem 2. The implication is $\widehat{\gamma} - \gamma^0_{N,T}$ does not depend on the asymptotic randomness of the identification condition. This can be seen intuitively from Figure 3. Looking at the small neighborhood around Γ^0 on the right. The randomness of the first order condition (Score $= 0$) line is given by the asymptotic distribution of $S(\Gamma^0)$, as in equation (42). Symmetrically, the randomness of the identification condition $\Theta^V_{N,T}$, is given by $I(\Gamma^0)$. Then it is straightforward to see why $\widehat{\gamma} - \gamma^0$ loads on both $S(\Gamma^0)$ and $I(\Gamma^0)$, while $\widehat{\gamma} - \gamma^0_{N,T}$ loads only on $S(\Gamma^0)$. Even though the position of $\Theta^V_{N,T}$ is "wobbling" left-to-right around Θ^{V*}, the difference of $\widehat{\gamma} - \gamma^0_{N,T}$ does not depend on the wobbling position of $\Theta^V_{N,T}$. But, it does depend on the "slope" of $\Theta^V_{N,T}$, which has a deterministic plim: J^0.

Theorem 2 (Asymptotic Distribution of Estimation Error - General)**.** *Under the preconditions of Theorem 1,*

$$\widehat{\gamma} - \gamma^0 = -\left(H^{0\top} H^0 + J^{0\top} J^0\right)^{-1} \left(H^{0\top} S(\Gamma^0) + J^{0\top} I(\Gamma^0)\right) + o_p\left(S(\Gamma^0) + I(\Gamma^0)\right) \tag{49}$$

$$\widehat{\gamma} - \gamma^0_{N,T} = -\left(H^{0\top} H^0 + J^{0\top} J^0\right)^{-1} H^{0\top} S(\Gamma^0) + o_p\left(S(\Gamma^0)\right) \tag{50}$$

[19] This shows the importance of keeping the deterministic γ^0 as the limiting reference point in the linearizion, even though in case the result is not explicit about γ^0.

This theorem is general to not just IPCA, and the analysis method can be applied to other estimations involving identification conditions.

Next, we specify to the IPCA case by calculating the four right hand side inputs in Theorem 2. The next subsection works out H^0 and $S(\Gamma^0)$, which are invariant to identification choices. Then we break out into three specific identification cases (corresponding to the three arrows in Figure 3), calculate the corresponding J^0 and $I(\Gamma^0)$, and discuss the implications on the asymptotic distribution respectively.

5.2 H^0 and Asymptotic Distribution of $S(\Gamma^0)$

So far, we have been as general as possible about the data generating process in analyzing the asymptotics. From this point, we sometimes impose some practical assumptions, which are required for calculating the analytical expressions. Assumption 6 shuts down the randomness of cross-sectional second moment of $c_{i,t}$, and is imposed in the next three lemmas.[20]

Assumption 6 (Practical Assumptions)**.** *Ω_t^{cc} is constant at Ω^{cc}.*

Lemma 3 (Calculate H^0)**.** *Under Assumption 6,*

$$H^0 = \left(\Omega^{cc} \otimes V^{ff}\right) \left.\frac{\partial \text{vect}\left(\Pi_{(\Gamma)}\right)}{\partial \gamma}\right|_{\gamma=\gamma^0}, \tag{51}$$

where $V^{ff} = \mathbb{E}\left[f_t^0 f_t^{0\top}\right]$.[21]

Proof in Appendix B.1.

The point-wise probability limit of S at Γ^0 is zero, as in the consistency section, while this subsection concerns about the asymptotic distribution.

When evaluated at Γ^0, $S_t^{[2]}$ and $S_t^{[4]}$ are zero, because they contain ${Q_{t(\Gamma)}}^\top \Gamma^0$ which is zero at $\Gamma = \Gamma^0$. The rest four terms, assigned into two pairs, are analyzed by the following two lemmas respectively.

Lemma 4 (Asymptotic Distribution of $S_t^{[1]} + S_t^{[3]}$)**.** *Under Assumptions 6, 5*

$$\frac{1}{\sqrt{NT}} \sum_t \text{vect}\left(S_t^{[1]}(\Gamma^0) + S_t^{[3]}(\Gamma^0)\right) \xrightarrow{d} \text{Normal}\left(0, \mathbb{V}^{[1]}\right), \qquad N, T \to \infty \tag{52}$$

[20] The results without imposing the assumption are still provided as intermediate steps in the respective proofs, which can be calculated with simulation.

[21] Under Assumption 6, $\Pi_{t(\Gamma)}$ is also time-constant and deterministic, so we drop its t subscript. So $\frac{\partial \text{vect}(\Pi(\Gamma))}{\partial \gamma}$ is just a deterministic derivatives that we do not write out but easily calculated with symbolic or numerical computation in simulation exercises.

where $\mathbb{V}^{[1]} = (Q^0 \otimes \mathbb{I}_L)\,\Omega^{cef}\left(Q^{0\top} \otimes \mathbb{I}_L\right)$ *and* $Q^0 := Q_{t(\Gamma^0)}$ *given that* Q_t *is constant over* t *under Assumption 6.*

Proof.

$$S_t^{[1]}(\Gamma^0) + S_t^{[3]}(\Gamma^0) = C_t^\top M_t(\Gamma^0) e_t f_t^{0\top} \tag{53}$$

$$= \left[\mathbb{I}_L - C_t^\top C_t \Gamma \left(\Gamma^\top C_t^\top C_t \Gamma\right)^{-1} \Gamma^\top\right] C_t^\top e_t f_t^{0\top} \tag{54}$$

$$= \left(Q_{t(\Gamma^0)} - \epsilon_{N,t}\right) C_t^\top e_t f_t^{0\top} \tag{55}$$

where

$$\epsilon_{N,t} = C_t^\top C_t \Gamma^0 \left(\Gamma^{0\top} C_t^\top C_t \Gamma^0\right)^{-1} \Gamma^{0\top} - \Omega_t^{cc} \Gamma^0 \left(\Gamma^{0\top} \Omega_t^{cc} \Gamma^0\right)^{-1} \Gamma^{0\top}. \tag{56}$$

We break out the two terms, put them into $\frac{1}{\sqrt{NT}} \sum_t \text{vect}\,(\cdot)$, and respectively show their asymptotics are:

$$\frac{1}{\sqrt{NT}} \sum_t \text{vect}\left(Q_{t(\Gamma^0)} C_t^\top e_t f_t^{0\top}\right) \xrightarrow{d} \text{Normal}\left(0, \mathbb{V}^{[1]}\right), \tag{57}$$

$$\frac{1}{\sqrt{NT}} \sum_t \text{vect}\left(\epsilon_{N,t} C_t^\top e_t f_t^{0\top}\right) \xrightarrow{p} \mathbf{0} \qquad N, T \to \infty. \tag{58}$$

For the first one, notice $\text{vect}\left(Q^0 C_t^\top e_t f_t^{0\top}\right) = (Q^0 \otimes \mathbb{I}_L)\,\text{vect}\left(C_t^\top e_t f_t^{0\top}\right)$. Then it is obvious applying a CMT to the CLT in Assumption 5.

The second one has two steps. First, fixing N, as T increases, we can apply the time-series CLT:

$$\frac{1}{\sqrt{T}} \sum_t \text{vect}\left(\epsilon_{N,t} \left(\frac{1}{\sqrt{N}} C_t^\top {e_t}^\top\right) f_t^{0\top}\right) \xrightarrow{d} \text{Normal}\left(\mathbf{0}, Var^{[N]}\right), \qquad T \to \infty \tag{59}$$

$$Var^{[N]} = \mathbb{V}\text{ar}\left[\text{vect}\left(\epsilon_{N,t} \left(\frac{1}{\sqrt{N}} C_t^\top {e_t}^\top\right) f_t^{0\top}\right) \middle| \mathfrak{F}^{[\text{cs}]}\right] \tag{60}$$

Then, as N increases, $\epsilon_{N,t} = o_p(1)$, $\frac{1}{\sqrt{N}} \sum_i c_{i,t} e_{i,t} = \mathcal{O}_p(1)$, $f_t^0 = \mathcal{O}_p(1)$ uniformly across t. So the product of three has a converging variance.

$$\operatorname*{plim}_{N\to\infty} \mathbb{V}\text{ar}\left[\text{vect}\left(\epsilon_{N,t} \left(\frac{1}{\sqrt{N}} C_t^\top {e_t}^\top\right) f_t^{0\top}\right) \middle| \mathfrak{F}^{[\text{cs}]}\right] = \mathbf{0} \tag{61}$$

The two directions combined:

$$\frac{1}{\sqrt{T}}\sum_t \text{vect}\left(\epsilon_{N,t}\left(\frac{1}{\sqrt{N}}C_t^\top {e_t}^\top\right) f_t^{0\top}\right) \xrightarrow{p} \mathbf{0}, \qquad N,T\to\infty. \tag{62}$$

. □

Lemma 5.

$$\frac{1}{T}\sum_t \text{vect}\left(S_t^{[4]}(\Gamma^0)+S_t^{[6]}(\Gamma^0)\right) \xrightarrow{p} \mathbb{E}^{[1]} \qquad N,T\to\infty \tag{63}$$

and under practical assumption 7, $\mathbb{E}^{[1]}=\mathbf{0}$.

Proof. Define

$$Q_t^N(\Gamma)=\left(\mathbb{I}_L - C_t^\top C_t\Gamma\left(\Gamma^\top C_t^\top C_t\Gamma\right)^{-1}\Gamma^\top\right) \tag{64}$$

Break down the summand into four terms as below,

$$\begin{aligned}
S_t^{[4]}(\Gamma^0)+S_t^{[6]}(\Gamma^0) &= C_t^\top M_t(\Gamma^0) e_t {e_t}^\top C_t\Gamma^0\left(\Gamma^{0\top}C_t^\top C_t\Gamma^0\right)^{-1} && (65)\\
&= \frac{1}{N}Q_t^N(\Gamma^0)C_t^\top e_t {e_t}^\top C_t\Gamma^0\left(\frac{1}{N}\Gamma^{0\top}C_t^\top C_t\Gamma^0\right)^{-1} && (66)\\
&= \frac{1}{N}Q_{t(\Gamma^0)}C_t^\top e_t {e_t}^\top C_t\Gamma^0\left(\Gamma^{0\top}\Omega_t^{cc}\Gamma^0\right)^{-1} && (67)\\
&+ \frac{1}{N}\left(Q_t^N(\Gamma^0)-Q_{t(\Gamma^0)}\right)C_t^\top e_t {e_t}^\top C_t\Gamma^0\left(\Gamma^{0\top}\Omega_t^{cc}\Gamma^0\right)^{-1} && (68)\\
&+ \frac{1}{N}Q_{t(\Gamma^0)}C_t^\top e_t {e_t}^\top C_t\Gamma^0\left(\left(\Gamma^{0\top}\Omega_t^{cc}\Gamma^0\right)^{-1}-\left(\Gamma^{0\top}\frac{1}{N}C_t^\top C_t\Gamma^0\right)^{-1}\right) && (69)\\
&+ \mathcal{O}_p\left(\Omega_t^{cc}-\frac{1}{N}C_t^\top C_t\right)C_t^\top e_t {e_t}^\top C_t\mathcal{O}_p\left(\Omega_t^{cc}-\frac{1}{N}C_t^\top C_t\right) && (70)
\end{aligned}$$

The last three terms (lines 68 - 70) all hinge on the difference of $\Omega_t^{cc}-\frac{1}{N}C_t^\top C_t$, which converges to zero uniformly at large N (assumption **??**). Then, it is easy to use Lemma **??** to show their time-series averages all have zero probability limit at large N,T. Next, we need to analyze the first term (line 67).

For the first term (line 67), write down its times-series average and break it down the by

self- and cross-interaction terms:

$$\frac{1}{T}\sum_{t}\text{vect}\left(\frac{1}{N}Q_{t(\Gamma^0)}C_t^\top e_t {e_t}^\top C_t\Gamma^0\left(\Gamma^{0\top}\Omega_t^{cc}\Gamma^0\right)^{-1}\right) \tag{71}$$

$$=\frac{1}{NT}\sum_{i,t}\text{vect}\left(Q_{t(\Gamma^0)}C_{i,t}^\top {e_{i,t}}^2 \quad C_{i,t}\Gamma^0\left(\Gamma^{0\top}\Omega_t^{cc}\Gamma^0\right)^{-1}\right) \tag{72}$$

$$+\frac{1}{NT}\sum_{i\neq j,t}\text{vect}\left(Q_{t(\Gamma^0)}C_{i,t}^\top e_{i,t}e_{j,t}C_{j,t}\Gamma^0\left(\Gamma^{0\top}\Omega_t^{cc}\Gamma^0\right)^{-1}\right) \tag{73}$$

The self-interaction term (line 72) admits the form of N,T average as in Lemma (**??**). It converges to the unconditional expectation in the large-N,T probability limit, which we define as $\mathbb{E}^{[1]}$:

$$\mathbb{E}^{[1]} := \mathbb{E}\left[\text{vect}\left(Q_{t(\Gamma^0)}C_{i,t}^\top {e_{i,t}}^2 C_{i,t}\Gamma^0\left(\Gamma^{0\top}\Omega_t^{cc}\Gamma^0\right)^{-1}\right)\right] \tag{74}$$

Assumption 7 (Cross-sectional Homoscedastic Error)**.**

$$\mathbb{E}\left[(e_{i,t})^2\middle|\mathfrak{F}^{[ts]}, C_{i,t}\right] = \mathbb{E}\left[(e_{i,t})^2\middle|\mathfrak{F}^{[ts]}\right] \tag{75}$$

When practical assumption 7 is imposed, we can further calculate that $\mathbb{E}^{[1]} = \mathbf{0}$:

$$\mathbb{E}^{[1]} := \mathbb{E}\left[\mathbb{E}\left[\text{vect}\left(Q_{t(\Gamma^0)}C_{i,t}^\top {e_{i,t}}^2 C_{i,t}\Gamma^0\left(\Gamma^{0\top}\Omega_t^{cc}\Gamma^0\right)^{-1}\right)\middle|\mathfrak{F}^{[\text{ts}]}, C_{i,t}\right]\right] \tag{76}$$

$$= \mathbb{E}\left[\text{vect}\left(Q_{t(\Gamma^0)}C_{i,t}^\top \mathbb{E}\left[{e_{i,t}}^2\middle|\mathfrak{F}^{[\text{ts}]}, C_{i,t}\right] C_{i,t}\Gamma^0\left(\Gamma^{0\top}\Omega_t^{cc}\Gamma^0\right)^{-1}\right)\right] \tag{77}$$

$$= \mathbb{E}\left[\text{vect}\left(\mathbb{E}\left[{e_{i,t}}^2\middle|\mathfrak{F}^{[\text{ts}]}\right] Q_{t(\Gamma^0)}\mathbb{E}\left[C_{i,t}^\top C_{i,t}\middle|\mathfrak{F}^{[\text{ts}]}\right]\Gamma^0\left(\Gamma^{0\top}\Omega_t^{cc}\Gamma^0\right)^{-1}\right)\right] \tag{78}$$

$$= \mathbb{E}\left[\text{vect}\left(\mathbb{E}\left[{e_{i,t}}^2\middle|\mathfrak{F}^{[\text{ts}]}\right] Q_{t(\Gamma^0)}\mathbb{E}\left[C_{i,t}^\top C_{i,t}\middle|\mathfrak{F}^{[\text{ts}]}\right]\Gamma^0\left(\Gamma^{0\top}\Omega_t^{cc}\Gamma^0\right)^{-1}\right)\right] \tag{79}$$

$$= \mathbf{0}_{LK\times 1} \tag{80}$$

in which, the last step holds because $Q_{t(\Gamma^0)}\Omega_t^{cc}\Gamma^0 = \mathbf{0}_{L\times K}$ period by period.

The cross-interaction terms (line 73) have zero expectations conditional on $\mathfrak{F}^{[\text{ts}]}$, for each

$i \neq j$. In addition, the conditional second moment is bounded as N increases:

$$\mathbb{E}\left[\frac{1}{N}\sum_{i\neq j}\text{vect}\left(Q_{t(\Gamma^0)}C_{i,t}^{\top}e_{i,t}e_{j,t}C_{j,t}\Gamma^0\left(\Gamma^{0\top}\Omega_t^{cc}\Gamma^0\right)^{-1}\right)\middle|\mathfrak{F}^{[\text{ts}]}\right] = \mathbf{0} \tag{81}$$

$$\mathbb{V}\text{ar}\left[\frac{1}{N}\sum_{i\neq j}\text{vect}\left(Q_{t(\Gamma^0)}C_{i,t}^{\top}e_{i,t}e_{j,t}C_{j,t}\Gamma^0\left(\Gamma^{0\top}\Omega_t^{cc}\Gamma^0\right)^{-1}\right)\middle|\mathfrak{F}^{[\text{ts}]}\right] \tag{82}$$

$$= \frac{1}{N^2}(N^2-N)\mathbb{V}\text{ar}\left[\text{vect}\left(Q_{t(\Gamma^0)}C_{i,t}^{\top}e_{i,t}e_{j,t}C_{j,t}\Gamma^0\left(\Gamma^{0\top}\Omega_t^{cc}\Gamma^0\right)^{-1}\right)\middle|\mathfrak{F}^{[\text{ts}]}\right] \tag{83}$$

$$= \mathcal{O}_p(1) \tag{84}$$

for some $i \neq j$ due to exchangeability. Combined, the time series average converges to zero in the large-N,T probability limit.

The analyses above prove Lemma 5. □

Proposition 3 (Asymptotic Distribution of the Score at the Deterministic True)**.** *If the ratio of N and T converges to a fixed number: $T/N \to \left(r^{[1]}\right)^2$,*

$$\sqrt{NT}S(\Gamma^0) \xrightarrow{d} \text{Normal}\left(r^{[1]}\mathbb{E}^{[1]}, \mathbb{V}^{[1]}\right). \tag{85}$$

Alternatively, under practical assumption 7 and if T/N is bounded,

$$\sqrt{NT}S(\Gamma^0) \xrightarrow{d} \text{Normal}\left(\mathbf{0}, \mathbb{V}^{[1]}\right). \tag{86}$$

The proof along with the expressions of $\mathbb{E}^{[1]}, \mathbb{V}^{[1]}$ are in appendix **??**.

5.3 Asymptotic Distribution - Specific Cases

5.3.1 Specific Case: Both $\widehat{\Gamma}, \Gamma^0 \in \Theta^I$

In the case of deterministic identification condition

$$J^0 = \left.\frac{\partial I(\Gamma)}{\partial \gamma^{\top}}\right|_{\gamma=\gamma^0} \tag{87}$$

and $I(\Gamma_{N,T}^0) = \mathbf{0}_{K^2\times 1}$,

Blow up (46) by $\sqrt{NT}$, take plims,

$$\underset{N,T\to\infty}{\text{plim}}\ \sqrt{NT}\left(\widehat{\gamma}-\gamma^0\right) + \left(H^{0\top}H^0 + J^{0\top}J^0\right)^{-1}H^{0\top}\sqrt{NT}S(\Gamma_{N,T}^0) = \mathbf{0} \tag{88}$$

$$\sqrt{NT}\left(\widehat{\gamma}-\gamma^0\right) \xrightarrow{d} -\left(H^{0\top}H^0 + J^{0\top}J^0\right)^{-1}H^{0\top}\text{Normal}\left(\mathbf{0}, \mathbb{V}^{[1]}\right) \tag{89}$$

In the special case of both $\widehat{\Gamma}, \Gamma^0 \in \Theta^I$, one way to specify an I is

$$I(\Gamma) = \text{vect}\left(Block_{1:K}(\Gamma) - \mathbb{I}_K\right) \tag{90}$$

Then

$$J^0 = [\mathbb{I}_{K^2 \times K^2}, \mathbf{0}_{K^2 \times (L-K)K}] \tag{91}$$

5.3.2 Specific Case: $\widehat{\Gamma} \in \Theta^V_{N,T}$, $\Gamma^0 \in \Theta^V$

Define the identification function I as stacked by two parts corresponding to [1] and [2] in $\Theta^V_{N,T}$ with lengths $\frac{1}{2}K(K+1)$ and $\frac{1}{2}K(K-1)$ respectively:[22]

$$I(\Gamma) = \begin{bmatrix} \text{veca}\left(\Gamma^\top \Gamma - \mathbb{I}_K\right) \\ \text{vecb}\left(I^{[2]}(\Gamma)\right) \end{bmatrix} \tag{92}$$

where $I^{[2]}(\Gamma) = \frac{1}{T}\sum_t \widehat{f}_{t(\Gamma)} \widehat{f}_t^\top{}_{(\Gamma)} - V^{ff}$.

Notice the second part is of sample data. Hence I is a random function now.

The first part of I is deterministic, so the first part of J^0 is easy:

$$J^0{}_{\left[1:\left(\frac{1}{2}K(K+1)\right),\ :\ \right]} = \left.\frac{\partial \text{veca}\left(\Gamma^\top \Gamma\right)}{\partial \gamma^\top}\right|_{\gamma=\gamma^0} \tag{93}$$

To analyzed the second part, calculate:

$$I^{[2]}(\Gamma) = \frac{1}{T}\sum_t \widehat{f}_{t(\Gamma)} \widehat{f}_t^\top{}_{(\Gamma)} - V^{ff} \tag{94}$$

Given,

$$\widehat{f}_{t(\Gamma)} = \tilde{f}_{t(\Gamma)} + \left(\Gamma^\top C_t^\top C_t \Gamma\right)^{-1} \Gamma^\top C_t^\top \tilde{e}_{t(\Gamma)} \tag{95}$$

[22] veca (A) stacks A's upper triangle entries, including the diagonal, in a vector by going right first, then down. vecb (A) stacks A's upper triangle entries, NOT including the diagonal, in a vector by going right first, then down.

$$I^{[2]}(\Gamma) = \frac{1}{T}\sum_t \Big(\tilde{f}_{t(\Gamma)} \qquad\qquad \tilde{f}_{t\,(\Gamma)}^{\top} - V^{ff}$$
$$\tilde{f}_{t(\Gamma)} \qquad\qquad \tilde{e}_{t\,(\Gamma)}^{\top} C_t \Gamma \left(\Gamma^{\top} C_t^{\top} C_t \Gamma\right)^{-1}$$
$$\left(\Gamma^{\top} C_t^{\top} C_t \Gamma\right)^{-1} \Gamma^{\top} C_t^{\top} \tilde{e}_{t(\Gamma)} \qquad \tilde{f}_{t\,(\Gamma)}^{\top}$$
$$\left(\Gamma^{\top} C_t^{\top} C_t \Gamma\right)^{-1} \Gamma^{\top} C_t^{\top} \tilde{e}_{t(\Gamma)} \qquad \tilde{e}_{t\,(\Gamma)}^{\top} C_t \Gamma \left(\Gamma^{\top} C_t^{\top} C_t \Gamma\right)^{-1} \Big) \tag{96}$$

First term has a plim in general, which in general is non-zero. When evaluated at Γ^0, since $\tilde{f}_t(\Gamma^0) = f_t^0$, it is $\mathcal{O}_p\left(1/\sqrt{T}\right)$. Second and third are $\mathcal{O}_p\left(1/\sqrt{NT}\right)$. Fourth is $\mathcal{O}_p\left(1/N\right)$.

$$\operatorname*{plim}_{N,T\to\infty} \sqrt{T} I^{[2]}(\Gamma^0) - \frac{1}{\sqrt{T}} \sum_t \left(f_t^0 f_t^{0\top} - V^{ff} \right) = \mathbf{0}_{K\times K} \tag{97}$$

$$\sqrt{T}\text{vecb}\left(I^{[2]}(\Gamma^0)\right) \xrightarrow{d} \text{Normal}\left(\mathbf{0}_{K\times K}, \mathbb{V}^{[2]}\right) \tag{98}$$

$$\mathbb{V}^{[2]} = \mathbb{V}\text{ar}\left[\text{vecb}\left(f_t^0 f_t^{0\top} - V^{ff}\right)\right] \tag{99}$$

Second part of J^0:

$$J^0{}_{\left[\left(\frac{1}{2}K(K+1)+1:K^2\right),\ :\ \right]} = \operatorname*{plim}_{N,T\to\infty} \frac{\partial}{\partial \gamma^{\top}} \text{vecb}\left(I^{[2]}\right)\Big|_{\gamma^0} \tag{100}$$

$$= \operatorname*{plim}_{N,T\to\infty} \frac{\partial}{\partial \gamma^{\top}} \text{vecb}\left(\frac{1}{T}\sum_t \left(\tilde{f}_{t(\Gamma)} \tilde{f}_{t\,(\Gamma)}^{\top}\right)\right)\Big|_{\gamma^0} \tag{101}$$

$$\frac{\partial}{\partial \gamma_p} \text{vecb}\left(\tilde{f}_{t(\Gamma)} \tilde{f}_{t\,(\Gamma)}^{\top}\right)\Big|_{\gamma^0} = \text{vecb}\left(\frac{\partial}{\partial \gamma_p}\tilde{f}_{t(\Gamma)}\Big|_{\gamma^0} f_t^{0\top} + f_t^0 \frac{\partial}{\partial \gamma_p}\tilde{f}_{t\,(\Gamma)}^{\top}\Big|_{\gamma^0}\right) \tag{102}$$

$$\tilde{f}_{t(\Gamma)} = \left(\Gamma^{\top}\Omega_t^{cc}\Gamma\right)^{-1}\Gamma^{\top}\Omega_t^{cc}\Gamma^0 f_t^0 \tag{103}$$

$$\frac{\partial}{\partial \gamma_p}\tilde{f}_{t(\Gamma)}\Big|_{\gamma^0} = \frac{\partial}{\partial \gamma_p}\left(\Gamma^{\top}\Omega_t^{cc}\Gamma\right)^{-1}\Gamma^{\top}\Omega_t^{cc}\Gamma^0\Big|_{\gamma^0} f_t^0 \tag{104}$$

$$:= D_p\left(\Gamma^0, \Omega_t^{cc}\right) f_t^0 \tag{105}$$

$$\frac{\partial}{\partial \gamma_p} \text{vecb}\left(\tilde{f}_{t(\Gamma)} \tilde{f}_{t\,(\Gamma)}^{\top}\right)\Big|_{\gamma^0} = \text{vecb}\left(D_p\left(\Gamma^0, \Omega_t^{cc}\right) f_t^0 f_t^{0\top} + f_t^0 f_t^{0\top} D_p^{\top}\left(\Gamma^0, \Omega_t^{cc}\right)\right) \tag{106}$$

$$J^0{}_{\left[\left(\frac{1}{2}K(K+1)+1:K^2\right),\ p\ \right]} = \mathbb{E}\left[\text{vecb}\left(D_p\left(\Gamma^0, \Omega_t^{cc}\right) f_t^0 f_t^{0\top} + f_t^0 f_t^{0\top} D_p^{\top}\left(\Gamma^0, \Omega_t^{cc}\right)\right)\right] \tag{107}$$

Under *Practical Assumptions*:

$$J^0{}_{\left[\left(\frac{1}{2}K(K+1)+1:K^2\right),\ p\ \right]} = \text{vecb}\left(D_p\left(\Gamma^0, \Omega^{cc}\right) V^{ff} + V^{ff} D_p^{\top}\left(\Gamma^0, \Omega^{cc}\right)\right) \tag{108}$$

Also the first part of $I(\Gamma^0)$ is deterministically zero:

$$I_{\left[1:\left(\frac{1}{2}K(K+1)\right),\ :\ \right]}(\Gamma^0) = \mathbf{0}_{\frac{1}{2}K(K+1)\times 1} \tag{109}$$

Now we have J^0 and adist of $I(\Gamma^0)$, put them back into (46). Blow (46) up by $\sqrt{T}$ and take plims:

$$\underset{N,T\to\infty}{\text{plim}}\ \sqrt{T}\left(\widehat{\gamma}-\gamma^0\right)+\left(H^{0\top}H^0+J^{0\top}J^0\right)^{-1}\left(H^{0\top}\sqrt{T}S(\Gamma^0)+J^{0\top}\sqrt{T}I(\Gamma^0)\right)=\mathbf{0} \tag{110}$$

Notice:

$$\underset{N,T\to\infty}{\text{plim}}\ \sqrt{T}S(\Gamma^0)=\mathbf{0} \tag{111}$$

So the $\mathcal{O}_p(1)$ randomness in adist comes from $\sqrt{T}I(\Gamma^0)$ in this case.

$$\sqrt{T}\left(\widehat{\gamma}-\gamma^0\right)\xrightarrow{d}-\left(H^{0\top}H^0+J^{0\top}J^0\right)^{-1}J^{0\top}\text{Normal}\left(\mathbf{0},\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbb{V}^{[2]}\end{bmatrix}\right) \tag{112}$$

5.3.3 Specific Case: Both $\widehat{\Gamma},\Gamma^0\in\Theta^V_{N,T}$

Call the deterministic Γ^0 in previous case Γ^0_V. Should first prove in the consistency part that $\text{plim}_{N,T\to\infty}\widehat{\Gamma}=\Gamma^0_V, \text{plim}_{N,T\to\infty}\Gamma^0=\Gamma^0_V$. Then the plims of $\bar{H},\bar{J}$ are all evaluated at Γ^0_V, so is the dlim of $S(\Gamma^0)$. Need to write proof in op notation.

By assumption, not only $I(\widehat{\Gamma})=\mathbf{0}$, but also $I(\Gamma^0)=\mathbf{0}$. According to Theorem **??**,

$$\sqrt{NT}\left(\widehat{\gamma}-\gamma^0\right)\xrightarrow{d}-\left(H^{0\top}H^0+J^{0\top}J^0\right)^{-1}H^{0\top}\text{Normal}\left(\mathbf{0},\mathbb{V}^{[1]}\right) \tag{113}$$

where H^0,J^0 are all the same as the previous subsection, $\mathbb{V}^{[1]}$ is evaluated at Γ^0_V.

5.4 Factor Asymptotics

Sample factor estimate is $\widehat{f}_t(\widehat{\Gamma})$. According to equation 26 and 23,

$$\widehat{f}_{t(\Gamma)}=\tilde{f}_{t(\Gamma)}+\left(\Gamma^\top C_t^\top C_t\Gamma\right)^{-1}\Gamma^\top C_t^\top\tilde{e}_{t(\Gamma)} \tag{114}$$

$$\tilde{f}_{t(\Gamma)}=\left(\Gamma^\top\Omega_t^{cc}\Gamma\right)^{-1}\Gamma^\top\Omega_t^{cc}\Gamma^0 f_t^0 \tag{115}$$

$$\widehat{f}_t(\widehat{\Gamma})=\tilde{f}_t(\widehat{\Gamma})+\left(\widehat{\Gamma}^\top C_t^\top C_t\widehat{\Gamma}\right)^{-1}\widehat{\Gamma}^\top C_t^\top\tilde{e}_t(\widehat{\Gamma}) \tag{116}$$

First term, we have $\tilde{f}_t(\Gamma^0) = f_t^0$ and $\widehat{\Gamma}$ consistency from Theorem **??**, therefore,

$$\operatorname*{plim}_{N,T\to\infty} \tilde{f}_t(\widehat{\Gamma}) = f_t^0 \tag{117}$$

Asymptotic distribution can be derived from the Delta method, rate $= \mathcal{O}_p\left(\text{GEER}_{N,T}\right)$.[23]

Second term: first separate out

$$\left(\widehat{\Gamma}^\top C_t^\top C_t \widehat{\Gamma}\right)^{-1} \widehat{\Gamma}^\top C_t^\top \tilde{e}_t(\widehat{\Gamma}) = \left(\Gamma^\top C_t^\top C_t \Gamma\right)^{-1} \Gamma^\top C_t^\top \tilde{e}_{t(\Gamma)} + \mathcal{O}_p\left(\text{GEER}_{N,T}\right) \tag{118}$$

First part is a cross-sectional sum, it has plim of zero, at rate $\mathcal{O}_p\left(1/\sqrt{N}\right)$.

Theorem 3 (Sample Factor's Asymptotics)**.**
Sample factor estimate $\widehat{f}_t(\widehat{\Gamma})$ is consistent.

$$\operatorname*{plim}_{N,T\to\infty} \widehat{f}_t(\widehat{\Gamma}) = f_t^0 \tag{119}$$

Estimation error rate is $\mathcal{O}_p\left(\max\left\{1/\sqrt{N}, \text{GEER}_{N,T}\right\}\right)$. (i.e. for deterministic identification it is $\mathcal{O}_p\left(1/\sqrt{N}\right)$. For random??, it is $\mathcal{O}_p\left(\max\left\{1/\sqrt{N}, 1/\sqrt{T}\right\}\right)$.
Asymptotic distribution is

6 Simulations

Matlab mostly finished. Need to write.

7 Applications

[23] $\text{GEER}_{N,T}$ stands for Gamma estimation error rate, as derived in Theorem **??**. It depends on identification condition choice.

References

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

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

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

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

Incomplete

Appendix

A Proofs for Section 4

A.1 Some Modular Lemmas

Let $x_{N,t}(\Gamma)$ represent a sequence (indexed by N) of stochastic process functions of Γ with finite dimension. Define several large N limiting conditions that it can be subject to.

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

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

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

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

Lemma 6. *If $x_{N,t}(\Gamma)$ is stationary in t and mean converging, then its time-series average is converging in the large-N probability limit uniformly for any T. That is, $\forall \epsilon, \delta > 0, \exists N^{[1]}$ s.t. $\forall T$ and $\forall N > N^{[1]}$*

$$Pr\left\{\sup_{\Gamma\in\Psi} \left\| \frac{1}{T}\sum_t x_{N,t}(\Gamma) \right\| > \epsilon \right\} < \delta. \tag{120}$$

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

$$\sup_{\Gamma\in\Psi} \left\| \frac{1}{T}\sum_t x_{N,t}(\Gamma) \right\| \le \sup_{\Gamma\in\Psi} \frac{1}{T}\sum_t \|x_{N,t}(\Gamma)\| \le \frac{1}{T}\sum_t \sup_{\Gamma\in\Psi} \|x_{N,t}(\Gamma)\| . \tag{121}$$

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

$$\mathbb{E}\sup_{\Gamma\in\Psi} \left\| \frac{1}{T}\sum_t x_{N,t}(\Gamma) \right\| \le \mathbb{E}\frac{1}{T}\sum_t \sup_{\Gamma\in\Psi} \|x_{N,t}(\Gamma)\| = \mathbb{E}\sup_{\Gamma\in\Psi} \|x_{N,t}(\Gamma)\| . \tag{122}$$

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 mean converging. Hence, the first term is T-uniform

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

$$\mathbb{E} \sup_{\Gamma \in \Psi} \left\| \frac{1}{T} \sum_{t} x_{N,t}(\Gamma) \right\| < \epsilon. \tag{123}$$

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

This Lemma 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 ϵ, δ. Hence, this statement implies large N, T simultaneous convergence, a property that will be used in the proof of Proposition 1 later.

Lemma 6 also shows 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 mean converging sequence.

Lemma 7. *If $x^{[1]}_{N,t}(\Gamma)$ is mean square converging, $x^{[2]}_{N,t}(\Gamma)$ is mean square bounded, and $x^{[3]}_{N,t}(\Gamma)$ is a.s. bounded, then*

1. *$x^{[1]}_{N,t}(\Gamma) x^{[2]}_{N,t}(\Gamma)$ is mean converging, which implies*
2. *$x^{[1]}_{N,t}(\Gamma)$ by itself is also mean converging.*
3. *$x^{[1]}_{N,t}(\Gamma) x^{[3]}_{N,t}(\Gamma)$ is still mean square converging.*
4. *$x^{[2]}_{N,t}(\Gamma) x^{[3]}_{N,t}(\Gamma)$ is still mean square bounded.*

Proof. 1. For each ω, we have

$$\sup_{\Gamma} \|x_{N,t}(\Gamma)\| = \sup_{\Gamma} \left\| x^{[1]}_{N,t}(\Gamma) x^{[2]}_{N,t}(\Gamma) \right\| \leq \sup_{\Gamma} \left\| x^{[1]}_{N,t}(\Gamma) \right\| \sup_{\Gamma} \left\| x^{[2]}_{N,t}(\Gamma) \right\|. \tag{124}$$

So, put inside expectation:

$$\mathbb{E} \sup_{\Gamma} \|x_{N,t}(\Gamma)\| \leq \mathbb{E} \left[\sup_{\Gamma} \left\| x^{[1]}_{N,t}(\Gamma) \right\| \sup_{\Gamma} \left\| x^{[2]}_{N,t}(\Gamma) \right\| \right] \tag{125}$$

$$\leq \left(\mathbb{E} \left[\sup_{\Gamma} \left\| x^{[1]}_{N,t}(\Gamma) \right\|^2 \right] \mathbb{E} \left[\sup_{\Gamma} \left\| x^{[2]}_{N,t}(\Gamma) \right\|^2 \right] \right)^{1/2} \tag{126}$$

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:

$$\left\|x_{N,t}(\Gamma)\right\|^2 = \left\|x_{N,t}^{[1]}(\Gamma)x_{N,t}^{[3]}(\Gamma)\right\|^2 \leq \left\|x_{N,t}^{[1]}(\Gamma)\right\|^2 \left\|x_{N,t}^{[3]}(\Gamma)\right\|^2 \tag{127}$$

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

$$\mathbb{E}\sup_\Gamma \left\|x_{N,t}(\Gamma)\right\|^2 \leq \mathbb{E}\sup_\Gamma \left\|x_{N,t}^{[1]}(\Gamma)\right\|^2 \left\|x_{N,t}^{[3]}(\Gamma)\right\|^2 \tag{128}$$

$$\leq \mathbb{E}\left[\sup_\Gamma \left\|x_{N,t}^{[1]}(\Gamma)\right\|^2 \sup_\Gamma \left\|x_{N,t}^{[3]}(\Gamma)\right\|^2\right] \tag{129}$$

For any ω, if $\sup_{\Gamma\in\Psi}\left\|x_{N,t}^{[3]}(\Gamma)\right\| < M$, then $\sup_{\Gamma\in\Psi}\left\|x_{N,t}^{[3]}(\Gamma)\right\|^2 < M^2$. That means $\left\|x_{N,t}^{[3]}(\Gamma)\right\|^2$ is also 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:

$$\mathbb{E}\sup_\Gamma \left\|x_{N,t}(\Gamma)\right\|^2 \leq \mathbb{E}\left[\sup_\Gamma \left\|x_{N,t}^{[1]}(\Gamma)\right\|^2\right] M^2 \tag{130}$$

Take large-N limits on both sides:

$$\lim_N \mathbb{E}\sup_\Gamma \left\|x_{N,t}(\Gamma)\right\|^2 \leq \lim_N \mathbb{E}\left[\sup_\Gamma \left\|x_{N,t}^{[1]}(\Gamma)\right\|^2\right] M^2 \tag{131}$$

$$= 0. \tag{132}$$

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

$$\mathbb{E}\left[x\middle|\mathfrak{F}^{[ts]}\right] = \mathbf{0},\ \mathbb{E}\left[\|x\|^2\middle|\mathfrak{F}^{[ts]}\right] < +\infty,\ \textit{and}\ \ \mathbb{E}\|x\|^2 < +\infty, \tag{133}$$

then its cross-sectional average $x_{N,t}(\Gamma) = \frac{1}{N}\sum_i x_{i,t}$ is mean square converging. [24]

[24] notice here Γ does not enter in the random function.

Proof.

$$\|x_{N,t}(\Gamma)\|^2 = \left\| \frac{1}{N} \sum_i x_{i,t} \right\|^2 = \sum_m \left(\frac{1}{N} \sum_i x_{m,i,t} \right)^2 \tag{134}$$

where m is the element index of the matrix x. Importantly, there are only finite m's. So, just need to show for each summand m, the cross-sectional average is mean square converging.

First analyze the t-s conditional expectation. By t-s conditional i.i.d. and zero t-s conditional expectation:

$$\mathbb{E}\left[\left(\frac{1}{N} \sum_i x_{m,i,t} \right)^2 \middle| \mathfrak{F}^{[\text{ts}]} \right] = \frac{1}{N} \mathbb{E}\left[x_{m,i,t}^2 \middle| \mathfrak{F}^{[\text{ts}]} \right]. \tag{135}$$

By LIE,

$$\mathbb{E}\left[\left(\frac{1}{N} \sum_i x_{m,i,t} \right)^2 \right] = \frac{1}{N} \mathbb{E}\left[x_{m,i,t}^2 \right] \to 0, \qquad N \to +\infty, \tag{136}$$

□

Define $\Omega_t^{ce} = \mathbb{V}\text{ar}\left[C_{i,t} e_{i,t} \middle| \mathfrak{F}^{[\text{ts}]} \right]$ ($L \times L$ variance covariance matrix).

A.2 Proof of Proposition 1

First, Lemma 9 deals with the cross-section convergence. Then, Lemma 10 deals with the time-series dimension. In the third step, the results are put together for large N, T convergence.

Lemma 9 (Large N Cross-sectional Convergence at each t)**.**

$$\frac{1}{N} \sum_i \left(S_t^{[1]}(\Gamma) + \cdots + S_t^{[6]}(\Gamma) \right) - \Omega_t^{cc} \Pi_{t(\Gamma)} f_t^0 \widetilde{f}_t^{\top}{}_{(\Gamma)} \tag{137}$$

is mean converging.

Proof. The cross-section convergence is the bulk of the analysis. We proceed by analyzing the six terms one by on. 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 mean square converging.

This is by Lemma 8, treating $C_{i,t}e_{i,t}$ as the $x_{i,t}$ in the lemma. The conditions are met given assumptions 1, 2.

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

   This is because it is a continuous function w.r.t. Γ, Ω_t^{cc}, and Γ^0 whose domains are all bounded and away from singularity given assumptions 3, 4.

3. $\tilde{f}_t^{\top(\Gamma)}$ is mean square bounded.

   $\tilde{f}_{t(\Gamma)} = \left(\Gamma^\top \Omega_t^{cc} \Gamma\right)^{-1} \Gamma^\top \Omega_t^{cc} \Gamma^0 f_t^0$, in which $\left(\Gamma^\top \Omega_t^{cc} \Gamma\right)^{-1} \Gamma^\top \Omega_t^{cc} \Gamma^0$ is a.s. bounded by the previous statement, f_t^0 is mean square bounded by assumption 21. Then apply lemma 7.4.

4. $\frac{1}{N} \sum_i \left(c_{i,t}^\top c_{i,t} - \Omega_t^{cc}\right)$ is 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 8. The conditions are met given the definition of Ω_t^{cc} and assumption 2.

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

   Notice decomposition $\frac{1}{N} S_t^{[1]}(\Gamma) = \left[\frac{1}{N} C_t^\top e_t\right] \left[\tilde{f}_t^{\top(\Gamma)}\right]$, use the two previous statements about the two parts and apply lemma 7.1.

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

$$\Pi_{t(\Gamma)} f_t^0 \tilde{f}_t^{\top(\Gamma)} = \left[\left(\mathbb{I}_L - \Gamma \left(\Gamma^\top \Omega_t^{cc} \Gamma\right)^{-1} \Gamma^\top \Omega_t^{cc}\right) \Gamma^0\right] \left[f_t^0 f_t^{0\top}\right] \left[\Gamma^{0\top} \Omega_t^{cc} \Gamma \left(\Gamma^\top \Omega_t^{cc} \Gamma\right)^{-1}\right] \tag{138}$$

   The third term, according to statement number 2 above, is a.s. bounded. By the same arguments, so is the first term. The middle term is mean square bounded by assumption 2(1). Then by Lemma 7.4, the three things together is mean square bounded.

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

$$\frac{1}{N} S_t^{[2]}(\Gamma) = \frac{1}{N} C_t^\top C_t \Pi_{t(\Gamma)} f_t^0 \tilde{f}_t^{\top(\Gamma)} \tag{139}$$

$$\frac{1}{N} S_t^{[2]}(\Gamma) - \Omega_t^{cc} \Pi_{t(\Gamma)} f_t^0 \tilde{f}_t^{\top(\Gamma)} = \left[\frac{1}{N} \sum_i \left(c_{i,t}^\top c_{i,t} - \Omega_t^{cc}\right)\right] \left[\Pi_{t(\Gamma)} f_t^0 \tilde{f}_t^{\top(\Gamma)}\right] \tag{140}$$

   Then, straightforward application of the previous two statements on Lemma 7.1.

8. $C_t^\top C_t \Gamma \left(\Gamma^\top C_t^\top C_t \Gamma\right)^{-1}$ is 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 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 Ω_t^{cc} a.s., which is invertible a.s. by Assumption 4. So for large enough N, $\frac{1}{N} C_t^\top C_t$ is invertible a.s. as well. Also Γ is full rank according to Assumption 3.

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

$$\frac{1}{N} \Gamma^\top C_t^\top \tilde{e}_{t(\Gamma)} = \frac{1}{N} \Gamma^\top C_t^\top e_t + \frac{1}{N} \Gamma^\top C_t^\top C_t Q_{t\,(\Gamma)}^\top \Gamma^0 f_t^0 \tag{141}$$

$$= \frac{1}{N} \Gamma^\top C_t^\top e_t + \frac{1}{N} \Gamma^\top \sum_i \left(c_{i,t}^\top c_{i,t} - \Omega_t^{cc}\right) Q_{t\,(\Gamma)}^\top \Gamma^0 f_t^0 \tag{142}$$

   The first term is mean square converging, according to statement 1, assumption 3, and Lemma 7.3. We want to show so is the second. We put a $\text{vect}\,(\cdot)$ operator to the summand, which does not affect the norm. Rearrange it as,

$$\text{vect}\left(\left(c_{i,t}^\top c_{i,t} - \Omega_t^{cc}\right) Q_{t\,(\Gamma)}^\top \Gamma^0 f_t^0\right) = \left(\left(c_{i,t}^\top c_{i,t} - \Omega_t^{cc}\right) \otimes f_t^{0\top}\right) \text{vect}\left(Q_{t\,(\Gamma)}^\top \Gamma^0\right) \tag{143}$$

   So the second term all together equals:

$$\Gamma^\top \left[\frac{1}{N} \sum_i \left(c_{i,t}^\top c_{i,t} - \Omega_t^{cc}\right) \otimes f_t^{0\top}\right] \text{vect}\left(Q_{t\,(\Gamma)}^\top \Gamma^0\right) \tag{144}$$

   The first and third part is a.s. bounded. The middle part is mean square converging, lemma 8, given assumptions 2(4). Then, we can just apply 7.4 twice.

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

$$\frac{1}{N} S_t^{[3]}(\Gamma) = \left[C_t^\top C_t \Gamma \left(\Gamma^\top C_t^\top C_t \Gamma\right)^{-1}\right] \left[\frac{1}{N} \Gamma^\top C_t^\top \tilde{e}_{t(\Gamma)}\right] \left[\tilde{f}_{t\,(\Gamma)}^\top\right] \tag{145}$$

    The three term are a.s. bounded, mean square converging, and mean square bounded, and apply Lemma 7.

11. $\frac{1}{N} S_t^{[4]}(\Gamma)$, $\frac{1}{N} S_t^{[5]}(\Gamma)$, $\frac{1}{N} S_t^{[6]}(\Gamma)$ are all 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 7 are omitted.

$$\frac{1}{N}S_t^{[4]}(\Gamma) = \left[\frac{1}{N}C_t^\top e_t\right]\left[\frac{1}{N}\tilde{e}_t^\top{}_{(\Gamma)}C_t\Gamma\right]\left(\Gamma^\top\frac{1}{N}C_t^\top C_t\Gamma\right)^{-1} \tag{146}$$

$$\frac{1}{N}S_t^{[5]}(\Gamma) = \left[\frac{1}{N}C_t^\top C_t\Pi_{t(\Gamma)}\right]\left[f_t^0\right]\left[\frac{1}{N}\tilde{e}_t^\top{}_{(\Gamma)}C_t\Gamma\right]\left(\Gamma^\top\frac{1}{N}C_t^\top C_t\Gamma\right)^{-1} \tag{147}$$

$$\frac{1}{N}S_t^{[6]}(\Gamma) = \left[\frac{1}{N}C_t^\top C_t\Gamma\left(\Gamma^\top\frac{1}{N}C_t^\top C_t\Gamma\right)^{-1}\right]\left[\frac{1}{N}\tilde{e}_t^\top{}_{(\Gamma)}C_t\Gamma\right]\left[\frac{1}{N}\Gamma^\top C_t^\top\tilde{e}_{t(\Gamma)}\right] \tag{148}$$

$$\left(\Gamma^\top\frac{1}{N}C_t^\top C_t\Gamma\right)^{-1} \tag{149}$$

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

Lemma 10 (S_T Convergence)**.**

$$\sup_{\Gamma\in\Psi}\left\|S_T(\Gamma) - S^0(\Gamma)\right\| \xrightarrow{p} 0, \qquad\qquad T\to\infty. \tag{150}$$

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 mcfadden??. It requires to construct the Γ-irrelevant random variable d_t, and verify it dominates t and has finite expectation. Notice,

$$S_T(\Gamma) = \frac{1}{T}\sum_t \text{vect}\left(t(\Gamma)\right) \tag{151}$$

$$t(\Gamma) = \Omega_t^{cc}\Pi_{t(\Gamma)}f_t^0\tilde{f}_t^\top{}_{(\Gamma)} \tag{152}$$

$$\|\text{vect}\left(t(\Gamma)\right)\| = \|t(\Gamma)\| = \left\|\left[\Omega_t^{cc}\Pi_{t(\Gamma)}\right]\left[f_t^0 f_t^{0\top}\right]\left[\Gamma^{0\top}\Omega_t^{cc}\Gamma\left(\Gamma^\top\Omega_t^{cc}\Gamma\right)^{-1}\right]\right\| \tag{153}$$

$$\leq \|\Omega_t^{cc}\Pi_{t(\Gamma)}\|\left\|\Gamma^{0\top}\Omega_t^{cc}\Gamma\left(\Gamma^\top\Omega_t^{cc}\Gamma\right)^{-1}\right\|\left\|f_t^0 f_t^{0\top}\right\| \tag{154}$$

$$\leq M\left\|f_t^0 f_t^{0\top}\right\| \tag{155}$$

$$:= d_t \tag{156}$$

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

$$Pr\left\{\sup_{\Gamma\in\Psi}\|\Omega_t^{cc}\Pi_{t(\Gamma)}\|\left\|\Gamma^{0\top}\Omega_t^{cc}\Gamma\left(\Gamma^\top\Omega_t^{cc}\Gamma\right)^{-1}\right\| < M\right\} = 1. \tag{157}$$

A finite M exists, because the two norms within sup are continuous functions on compact

domains, given assumptions 3, 4.

We have thus constructed d_t, and shown that $\|\text{vect}\,(t(\Gamma))\| \leq \|d_t\|$. It is also straightforward that $\mathbb{E}d_t < \infty$ given Assumption 2.1. □

After preparing the lemmas above, we can finally start proving Proposition 1.

Proof. According to Lemma 9,

$$\frac{1}{N}\sum_i \left(S_t^{[1]}(\Gamma) + \cdots + S_t^{[6]}(\Gamma)\right) - \Omega_t^{cc}\Pi_{t(\Gamma)} f_t^0 \tilde{f}_t^{\top}{}_{(\Gamma)} \tag{158}$$

is mean converging. We have the time-series average:

$$S_T(\Gamma) = \frac{1}{T}\sum_t \text{vect}\left(\Omega_t^{cc}\Pi_{t(\Gamma)} f_t^0 \tilde{f}_t^{\top}{}_{(\Gamma)}\right) \tag{159}$$

Apply Lemma 6, we have

$$\sup_{\Gamma\in\Psi} \|S(\Gamma) - S_T(\Gamma)\| \xrightarrow{p} 0, \qquad N \to \infty, \forall T. \tag{160}$$

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

$$Pr\left\{\sup_{\Gamma\in\Psi} \|S(\Gamma) - S_T(\Gamma)\| < \delta\right\} > 1 - \epsilon. \tag{161}$$

By Lemma 10,

$$\sup_{\Gamma\in\Psi} \left\|S_T(\Gamma) - S^0(\Gamma)\right\| \xrightarrow{p} 0, \qquad T \to \infty. \tag{162}$$

That is to say, $\forall \epsilon, \delta > 0$, $\exists T^{[1]}$, s.t. irrelevant of N, $\forall T > T^{[1]}$

$$Pr\left\{\sup_{\Gamma\in\Psi} \left\|S_T(\Gamma) - S^0(\Gamma)\right\| < \delta\right\} > 1 - \epsilon. \tag{163}$$

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

$$Pr\left\{\sup_{\Gamma\in\Psi} \left\|S(\Gamma) - S^0(\Gamma)\right\| < 2\delta\right\} \tag{164}$$

$$\geq Pr\left\{\sup_{\Gamma\in\Psi} \|S(\Gamma) - S_T(\Gamma)\| < \delta \text{ AND } \sup_{\Gamma\in\Psi} \left\|S_T(\Gamma) - S^0(\Gamma)\right\| < \delta\right\} \tag{165}$$

$$\geq 1 - 2\epsilon. \tag{166}$$

That means the required conclusion:

$$\sup_{\Gamma\in\Psi}\left\|S(\Gamma)-S^0(\Gamma)\right\|\xrightarrow{p}0, \qquad N,T\to\infty \tag{167}$$

□

A.3 Specific Identification Functions Satisfy Condition 1

Lemma 11. *Need to make*

Lemma 12. *Need to make*

A.4 Proof of Theorem 2

Proof. "If": It is easy to verify that Γ^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 unidentified with Γ^0, $S^0(\Gamma)\neq\mathbf{0}$. The term in S^0:

$$\Omega_t^{cc}\Pi_{t(\Gamma)}f_t^0\tilde{f}_{t\,(\Gamma)}^{\top}=\Omega_t^{cc}\left(\mathbb{I}_L-\Gamma\left(\Gamma^\top\Omega_t^{cc}\Gamma\right)^{-1}\Gamma^\top\Omega_t^{cc}\right)\Gamma^0f_t^0f_t^{0\top}\Gamma^{0\top}\Omega_t^{cc}\Gamma\left(\Gamma^\top\Omega_t^{cc}\Gamma\right)^{-1} \tag{168}$$

$$:=ARf_t^0f_t^{0\top}R^\top B, \tag{169}$$

where R is rotation s.t. $\mathbb{E}Rf_t^0f_t^{0\top}R$ 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 unidentified with Γ^0, $\Pi_{t(\Gamma)}\neq\mathbf{0}$, so A, B are full rank K w.p. 1, by Assumption 4. One can construct constant p,q of length L,K s.t. the signs of each entry in $p^\top A$ and Bq are always the same. As a result, $p^\top\mathbb{E}\left[ARf_t^0f_t^{0\top}RB\right]q>0$. □

B Proofs and Intermediate Steps for Section 5

B.1 Proof of Lemma 3

$$H^0=\left.\frac{\partial S^0(\Gamma)}{\partial\gamma^\top}\right|_{\gamma=\gamma^0}=\mathbb{E}\left[\left.\frac{\partial\text{vect}\left(\Omega_t^{cc}\Pi_{t(\Gamma)}f_t^0\tilde{f}_{t\,(\Gamma)}^{\top}\right)}{\partial\gamma^\top}\right|_{\gamma=\gamma^0}\right] \tag{170}$$

Notice $\Pi_t(\Gamma^0)=\mathbf{0}$. Therefore, the terms involving $\nabla\tilde{f}_t$ will drop out. We write the H^0

column by column. The p'th column, or the derivative w.r.t. the p'th entry γ_p simplifies as

$$H^0{}_p = \mathbb{E}\left[\text{vect}\left(\Omega_t^{cc}\left.\frac{\partial \Pi_t(\Gamma)}{\partial \gamma_p}\right|_{\gamma=\gamma^0} f_t^0 f_t^{0\top}\right)\right] \tag{171}$$

This result above does not require Assumption 6, and can be calculated by LLN simulation if one is interested in the general case. Now, we impose the constant Ω_t^{cc} assumption:

$$H^0{}_p = \text{vect}\left(\Omega^{cc}\left.\frac{\partial \Pi(\Gamma)}{\partial \gamma_p}\right|_{\gamma=\gamma^0} V^{ff}\right) \tag{172}$$

$$= \left(\Omega^{cc} \otimes V^{ff}\right)\left.\frac{\partial \text{vect}\left(\Pi(\Gamma)\right)}{\partial \gamma_p}\right|_{\gamma=\gamma^0} \tag{173}$$

Then stack the columns together, we have the desired result:

$$H^0 = \left(\Omega^{cc} \otimes V^{ff}\right)\left.\frac{\partial \text{vect}\left(\Pi(\Gamma)\right)}{\partial \gamma^\top}\right|_{\gamma=\gamma^0} \tag{174}$$