EMPIRICAL CRYPTO ASSET PRICING USING FACTOR MODELS WITH HIGH-DIMENSIONAL CHARACTERISTICS

Adam Baybutt

November 9, 2023

http://www.adambaybutt.org/research.html

PREVIEW: SETUP

Consider a dynamic latent factor model with linear loadings

$$r_{i,t+1} = \underbrace{z_{i,t}^\top \Gamma_\beta}_{\beta_{i,t}^\top} f_{t+1} + \epsilon_{i,t+1}, \quad \mathbb{E}[\epsilon_{i,t+1} | z_{i,t}] = 0,$$

where we observe, for assets i and time periods t,

- asset excess returns $r_{i,t+1} \in \mathbb{R}$ and
- asset characteristics $z_{i,t} \in \mathbb{R}^p$.

PREVIEW: MAIN THEORY CONTRIBUTIONS

In this setup, under the novel asymptotics of $p, T, N \to \infty$, contribute a new estimation procedure for

- latent loadings $\Gamma_\beta \in \mathbb{R}^{p \times k}$ and
- latent factors $f_{t+1} \in \mathbb{R}^k$, for all t;

and, prove the consistency of these estimators.

Also, I extend to this setting a classic asset pricing test and provide an asymptotically valid inference procedure.

MOTIVATION

Static observable factor model:

$$r_{i,t+1} = \beta_i^\top f_{t+1} + \epsilon_{i,t+1}$$

$(NT + Tk)$ data $\gtrsim$ (Nk) params.

Static latent factor model:

$$r_{i,t+1} = \beta_i^\top f_{t+1} + \epsilon_{i,t+1}$$

$$NT \gtrsim Nk + Tk$$

Dynamic latent factor model:

$$r_{i,t+1} = z_{i,t}^\top \Gamma_\beta f_{t+1} + \epsilon_{i,t+1}$$

$$NT(1 + p) \gtrsim pk + Tk$$

$\forall t \in \{1, \ldots, T\} \wedge i \in \{1, \ldots, N\}$:

Observed:

$r_{i,t+1} \in \mathbb{R}$ asset excess returns

$z_{i,t} \in \mathbb{R}^p$ asset characteristics

Unobserved:

$\epsilon_{i,t+1} \in \mathbb{R}$ idiosyncratic error

$f_{t+1} \in \mathbb{R}^k$ low-dim. factors

$\Gamma_\beta \in \mathbb{R}^{p \times k}$ loading mapping

$H \in \mathbb{R}^{k \times k}$ rotation matrix

SETUP

Assume for time periods $t = 1, \ldots, T$ and assets $i = 1, \ldots, N$, we observe

- asset excess returns $r_{i,t+1} \in \mathbb{R}$ and asset characteristics $z_{i,t} \in \mathbb{R}^p$.

Assume the model:

$$r_{i,t+1} = \underbrace{z_{i,t}^\top \Gamma_\beta}_{\beta_{i,t}^\top} f_{t+1} + \epsilon_{i,t+1}, \quad \mathbb{E}[\epsilon_{i,t+1} | z_{i,t}] = 0,$$

where

- $f_{t+1} \in \mathbb{R}^k$ are low-dimensional latent factors and
- $\Gamma_\beta \in \mathbb{R}^{p \times k}$ are unknown factor loading parameters.
  - Key assump.: Γ_β is exactly row sparse, i.e. most rows exactly zero.

EXTENDED SETUP (1/2)

Within this framework, we address an asset pricing research question.

> What is the risk premium of an observable nontradable factor $g_{t+1} \in \mathbb{R}$?

Asset pricing context:

- Risk premium: return for exposure to the factor, ceteris paribus.
- If tradable, the risk premium is the time-series average of the factor.
- If nontradable, form factor mimicking-portfolio.
- Following Giglio, Xiu, and Zhang (2021),
  - assume latent factor model recovers true factor model and
  - project observable nontradable factor onto latent factors.

EXTENDED SETUP (2/2)

What is the risk premium of an observable nontradable factor $g_{t+1} \in \mathbb{R}$?

Assume for true factors f_{t+1} :

$$f_{t+1} := \gamma + v_{t+1}, \qquad \mathbb{E}[v_{t+1}] = 0$$
$$g_{t+1} = \delta + \eta^\top v_{t+1} + \epsilon^g_{t+1}, \quad \mathbb{E}[v_{t+1}\epsilon^g_{t+1}] = 0.$$

where

- $\eta \in \mathbb{R}^k$ is an unknown parameter mapping and
- ϵ^g_{t+1} is measurement error in g_{t+1}.

Our target parameter is $\gamma_g = \eta^\top \gamma$.

THEORETICAL CONTRIBUTIONS

The model:

$$r_{i,t+1} = z_{i,t}^{\top}\Gamma_{\beta}(\gamma + v_{t+1}) + \epsilon_{i,t+1}, \quad \mathbb{E}[\epsilon_{i,t+1}|z_{i,t}] = 0, \; \mathbb{E}[v_{t+1}\epsilon_{i,t+1}] = 0,$$

$$g_{t+1} = \delta + \eta^{\top} v_{t+1} + \epsilon_{t+1}^{g}, \qquad \mathbb{E}[v_{t+1}\epsilon_{t+1}^{g}] = 0.$$

Two contributions, under novel asymptotics of $p, T, N \to \infty$:

1. consistently estimate latent loadings Γ_{β} and factors f_{t+1} and
2. conduct inference on $\gamma_g = \eta^{\top}\gamma$
   - under novel use of a dynamic latent factor model.

OUTLINE

1. ~~Preview~~
2. ~~Motivation~~
3. ~~Setup~~
4. ~~Theoretical Contributions~~
5. Theory Literature Review
6. Estimation
7. Key Assumptions
8. Asymptotic Results
9. Proof Outlines
10. Monte Carlo Evidence

THEORY LITERATURE REVIEW

The scope of the relevant literature is enormous. To name a few:

- *Dynamic latent factor models*: Connor and Linton (2007), Fan, Liao, and Wang (2016), Kelly, Pruitt, and Su (2019) IPCA, Kelly, Pruitt, and Su (2020), etc.
- *Tests of observable factors*: Fama and MacBeth (1973) Fama-MacBeth, Feng, Giglio, and Xiu (2020) Factor Zoo, Giglio and Xiu (2021), etc.
- *DML*: Belloni, Chernozhukov, and Hansen (2014), Chernozhukov et al. (2018), Semenova and Chernozhukov (2021), etc.

ESTIMATION (1/4)

Rewrite the model:

$$
\begin{aligned}
r_{i,t+1} &= z_{i,t}^{\top} \Gamma_{\beta} f_{t+1} + \epsilon_{i,t+1}, \\
&= z_{i,t,j} c_{t+1,j} + z_{i,t,-j}^{\top} c_{t+1,-j} + \epsilon_{i,t+1}, \qquad E[\epsilon_{i,t+1} | z_{i,t}] = 0, \\
c_{t+1,j} &:= \Gamma_{\beta,j}^{\top} f_{t+1}.
\end{aligned}
$$

To estimate $c_{t+1,j}$ $\forall t, j$

- ~~run Lasso to account for $p \sim N$, but then biased inference for γ_g;~~
- instead run Double Selection Lasso (DSL).

ESTIMATION (2/4)

Model:

$$
\begin{aligned}
r_{i,t+1} &= z_{i,t,j} c_{t+1,j} + z_{i,t,-j}^{\top} c_{t+1,-j} + \epsilon_{i,t+1}, \qquad E[\epsilon_{i,t+1} | z_{i,t}] = 0, \\
c_{t+1,j} &:= \Gamma_{\beta,j}^{\top} f_{t+1}.
\end{aligned} \tag{1}
$$

Procedure:

1. To estimate $\widehat{c}_{t+1,j}$, run $T \times p$ cross sectional DSL regressions. DSL
2. To estimate $\widehat{\Gamma}_{\beta} \in \mathbb{R}^{p \times k}$ and $\widehat{F} \in \mathbb{R}^{T \times k}$, run PCA on $\widehat{C} := \widehat{F}\widehat{\Gamma}_{\beta}^{\top} \in \mathbb{R}^{T \times p}$.
3. Given exact row sparsity, soft-threshold $\widehat{\Gamma}_{\beta}$ to set most rows to zero for $\check{\Gamma}_{\beta}$.

ESTIMATION (3/4)

Model for risk premia of nontradable observable factors:

$$r_{i,t+1} = z_{i,t}^{\top}\Gamma_{\beta}(\gamma + v_{t+1}) + \epsilon_{i,t+1}, \quad \mathbb{E}[\epsilon_{i,t+1}] = 0,\ \mathbb{E}[v_{t+1}\epsilon_{i,t+1}] = 0,$$

$$g_{t+1} = \eta^{\top} v_{t+1} + \epsilon^{g}_{t+1}, \quad \mathbb{E}[\epsilon^{g}_{t+1}] = 0,\ \mathbb{E}[v_{t+1}\epsilon^{g}_{t+1}] = 0.$$

Identification:

- Cannot jointly estimate η and v_{t+1} (Γ_{β} and f_{t+1}) without further restrictions. E.g., three classic approaches of Bai and Ng (2013).
- So parameters are identified up to rotation matrix $H \in \mathbb{R}^{k\times k}$. That is, $\eta = H^{-1}\eta_0$ and $\gamma = H\gamma_0$ ($\Gamma_{\beta} = \Gamma_b^0 H^{-1}$ and $f_{t+1} = H f^0_{t+1}$).
- Utilize rotation invariant result of Giglio and Xiu (2021):

$$\gamma_g = \eta_0^{\top} H^{-1\top} H\gamma_0 = \eta^{\top}\gamma$$

ESTIMATION (4/4)

Model for risk premia of nontradable observable factors:

$$\begin{aligned} r_{i,t+1} &= z_{i,t}^{\top}\Gamma_{\beta}(\gamma + v_{t+1}) + \epsilon_{i,t+1}, \quad & \mathbb{E}[\epsilon_{i,t+1}] = 0,\ \mathbb{E}[v_{t+1}\epsilon_{i,t+1}] = 0, \\ g_{t+1} &= \eta^{\top} v_{t+1} + \epsilon^{g}_{t+1}, & \mathbb{E}[\epsilon^{g}_{t+1}] = 0,\ \mathbb{E}[v_{t+1}\epsilon^{g}_{t+1}] = 0. \end{aligned} \tag{2}$$

Procedure: $\widehat{\gamma}_g = \widehat{\eta}^{\top}\widehat{\gamma}$

- Estimate factor innovations $\widehat{v}_{t+1}$ and loadings $\check{\Gamma}_{\beta}$ as before but with demeaned returns.
- Estimate latent factor risk premia $\widehat{\gamma}$ via CS OLS of average returns $\bar{r} \in \mathbb{R}^N$ on estimated latent factor loadings $\bar{\widehat{\beta}} := T^{-1}\sum_t Z_t\widehat{\Gamma}_{\beta} \in \mathbb{R}^N$.
- Estimate latent to observable factor mapping $\widehat{\eta}$ via TS OLS of demeaned g_{t+1} on estimated latent factor innovations $\widehat{v}_{t+1}$.

OUTLINE

1. ~~Motivation~~
2. ~~Setup~~
3. ~~Theory Research Questions~~
4. ~~Theory Literature Review~~
5. ~~Estimation~~
6. Key Assumptions
7. Asymptotic Results
8. Proof Outlines
9. Monte Carlo Evidence

KEY ASSUMPTIONS (1/2)

Assumption (Consistency of DSL)

1. *Sparse Loading: Loading matrix* Γ_β *admits an exactly sparse form. That is, for* $\exists s \in \mathbb{N}_+$, *i.e.* $p > s \geq 1$, Γ_β *has at most s nonzero rows:* $\sum_{j=1}^{p} \mathbb{1}\left\{ \left\| \Gamma_{\beta,j} \right\|_1 > 0 \right\} \leq s.$ *Additional DSL Assumptions*

KEY ASSUMPTIONS (2/2)

Assumption (Consistency of Latent Factor Model)

2. *Nonzero and distinct eigenvalues: from the infeasible eigendecomposition of* $(T\,p)^{-1}CC^\top$*, the k largest eigenvalues* λ_i *for* $i \in \{1, \ldots, k\}$ *are bounded away from zero and distinct,*

$$\min_{i:i\neq\kappa} |\lambda_\kappa - \lambda_i| > 0.$$

Additional Latent Factor Model Consistency Assumptions

ASYMPTOTIC RESULTS (1/3)

Proposition (Consistency of Latent Factors)

Under the DSLFM model (1) *and aforementioned Assumptions 1 and 2, with additional Appendix Assumptions 1-6, where* $T, N, p \to \infty$, *then for all t the latent factor estimator has the property that*

$$\widehat{f}_{t+1} - H^{\top} f^{0}_{t+1} = O_p\left(\sqrt{\frac{s \log(T p)}{N}}\right).$$

PROOF OUTLINE: CONSISTENT LATENT FACTORS

Recall $C = F\Gamma_\beta^\top$, thus $(T\,p)^{-1}CC^\top = (T\,p)^{-1}F\Gamma_\beta^\top\Gamma_\beta F^\top$.

Key rate: $\max_{t,j}\left|\widehat{c}_{t+1,j} - c_{t+1,j}\right| = O_p\left(\sqrt{\frac{\log(T\,p)}{N}}\right)$.

Gives control over the distance between feasible and infeasible matrix:

$$\left\|(T\,p)^{-1}\widehat{C}\widehat{C}^\top - (T\,p)^{-1}CC^\top\right\| = O_p\left(\frac{\log T\,p}{N}\right).$$

Davis Kahan Theorem bounds distance between eigenvectors by distance between matrices.

Finally, use Wely inequality to bound distance between eigenvalues.

ASYMPTOTIC RESULTS (2/3)

Proposition (Consistency of Latent Factor Loadings)

Under the DSLFM model (1) *and aforementioned Assumptions 1 and 2, with additional Appendix Assumptions 1-6, where* $T, N, p \to \infty$, *then the latent loading estimator has the property that*

$$\check{\Gamma}_\beta - \Gamma_\beta^0 H^{-1} = O_p\left(\sqrt{\frac{s\log(Tp)}{N}}\right).$$

PROOF OUTLINE: CONSISTENT LOADINGS

Aforementioned results yield:

$$\left\| \widehat{\Gamma}_\beta - \Gamma_\beta^0 (H^\top)^{-1} \right\|_\infty = O_p \left(\sqrt{\frac{\log(T\,p)}{N}} \right).$$

Utilizing Theorem 2.10 from Belloni et al. (2018) under exact sparsity of Γ_β^0, s.t.

$$\lambda \geq (1 - \alpha) - \text{quantile of } \left\| \widehat{\Gamma}_\beta - \Gamma_\beta^0 (H^\top)^{-1} \right\|_\infty,$$

then given $\alpha \to 0$ and $\lambda \lesssim \sqrt{\log(T\,p)/N}$, we have for all $q \geq 1$

$$\left\| \check{\Gamma}_{\beta,l} - \Gamma_\beta^0 (H^\top)_l^{-1} \right\|_q \lesssim_P s^{1/q} \sqrt{\frac{\log(T\,p)}{N}}.$$

ASYMPTOTIC RESULTS (3/3)

Theorem (Normality of Observable Factor Risk Premium)

Under the models (1) *and* (2)*; Assumptions 1 and 2; Appendix Assumptions 1-10, and, if* $Ts^2 \log(Tp)/N \to 0$*, then as* $T, N, p \to \infty$ *the estimator* $\hat{\gamma}_g$ *obeys*

$$\sqrt{T}\frac{(\hat{\gamma}_g - \gamma_g)}{\sigma_g} \xrightarrow{d} \mathcal{N}(0,1).$$

PROOF OUTLINE: NORMALITY

$$\sqrt{T}\left(\widehat{\gamma}_g - \gamma_g\right) = \sqrt{T}\left(\widehat{\eta}^\top\widehat{\gamma} - \eta^\top\gamma\right)$$
$$= \underbrace{\sqrt{T}\gamma^\top\left(\widehat{\eta} - \eta\right)}_{\to_d \mathcal{N}(0,\sigma^2)} + \sqrt{T}\eta^\top\left(\widehat{\gamma} - \gamma\right) + o_p(1).$$

$$\sqrt{T}\eta^\top\left(\widehat{\gamma} - \gamma\right) = \sqrt{T}\left(\widehat{\gamma} - \widetilde{\gamma}\right) + \sqrt{T}\left(\widetilde{\gamma} - H\gamma_0\right)$$
$$= o_p(1) + \underbrace{\sqrt{T}\left(\frac{\bar{\beta}^\top\bar{\beta}}{N}\right)^{-1}\frac{\bar{\beta}^\top}{N}\frac{1}{T}\sum_t Z_t\Gamma_\beta^0 v_{t+1}^0}_{\to N(0,\sigma^2)}$$
$$+ \underbrace{\sqrt{T}H^\top\left(\frac{\Gamma_\beta^{0\top}\bar{Z}^\top\bar{Z}\Gamma_\beta^0}{N}\right)^{-1}\frac{\bar{\beta}^\top}{N}\frac{1}{T}\sum_t \epsilon_{t+1}}_{o_p(1)}$$

MONTE CARLO EVIDENCE (1/2)

Goal: study the finite-sample estimation error of our latent loading and factor estimators and the coverage properties of our risk premium estimator compared to relevant benchmarks.

DGP: for $S = 200, T = 100, N = 500, k = 3,\ p \in \{10, 50\}, s = p/10$

- Latent loadings: fit IPCA to empirical panel; set $p - s$ rows to zero.
- Latent factors: fit IPCA to empirical panel; fit VAR(1) to fitted latent factors; simulate from fitted VAR(1) with normal innovations.
- Characteristics: fit panel VAR(1) to demeaned empirical panel of $\{Z_t\}_{t=1}^{T}$ and simulate from VAR(1) with normal innovations. Set means to bs.
- Returns and observable factor are generated according to the model where errors are calibrated to empirical R^2.

MONTE CARLO EVIDENCE (2/2)

Low-Dimensional: $p = 10$ Simulation Results Low-Dim.

- Factor of ~ 3 superior estimation error for Γ_β.
- Order of magnitude inferior estimation error for f_{t+1}.
- DSLFM under-covers (6-9%) while Giglio over-covers (2-4%) γ_g.

High-Dimensional: $p = 50$ Simulation Results High-Dim.

- Factor of >3 superior estimation error for Γ_β.
- Inferior ($\times 4$) estimation error for f_{t+1}.
- DSLFM degrades 1% while Giglio degrades > 3% γ_g.

REFERENCES (1/2)

Bai, Jushan, and Serena Ng. 2013. "Principal components estimation and identification of static factors." *Journal of econometrics* 176 (1): 18–29.

Belloni, Alexandre, Victor Chernozhukov, Denis Chetverikov, Christian Hansen, and Kengo Kato. 2018. "High-dimensional econometrics and regularized GMM." *arXiv preprint arXiv:1806.01888*.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. "Inference on treatment effects after selection among high-dimensional controls." *The Review of Economic Studies* 81 (2): 608–650.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. "Double/debiased machine learning for treatment and structural parameters." *The Econometrics Journal*.

Connor, Gregory, and Oliver Linton. 2007. "Semiparametric estimation of a characteristic-based factor model of common stock returns." *Journal of Empirical Finance* 14 (5): 694–717.

Fama, Eugene F, and James D MacBeth. 1973. "Risk, return, and equilibrium: Empirical tests." *Journal of political economy* 81 (3): 607–636.

Fan, Jianqing, Yuan Liao, and Weichen Wang. 2016. "Projected principal component analysis in factor models." *Annals of statistics* 44 (1): 219.

Feng, Guanhao, Stefano Giglio, and Dacheng Xiu. 2020. "Taming the factor zoo: A test of new factors." *The Journal of Finance* 75 (3): 1327–1370.

Giglio, Stefano, and Dacheng Xiu. 2021. "Asset pricing with omitted factors." *Journal of Political Economy* 129 (7): 1947–1990.

Giglio, Stefano, Dacheng Xiu, and Dake Zhang. 2021. "Test assets and weak factors."Technical report, National Bureau of Economic Research.

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

Kelly, Bryan T, Seth Pruitt, and Yinan Su. 2020. "Instrumented principal component analysis." *Available at SSRN 2983919*.

Semenova, Vira, and Victor Chernozhukov. 2021. "Debiased machine learning of conditional average treatment effects and other causal functions." *The Econometrics Journal* 24 (2): 264–289.

REFERENCES (2/2)

See the paper for the rest of the long list...

Thank You!

APPENDIX: IPCA

The model is

$$r_{i,t} = z_{i,t-1}^{\top} \Gamma_\delta \, f_t + \epsilon_{i,t}.$$

The objective function is to minimize the sum of the squared errors:

$$\min_{\Gamma_\delta, f_t} \sum_{t=1}^{T} \left(r_t - Z_{t-1} \Gamma_\delta \, f_t\right)^{\top} \left(r_t - Z_{t-1} \Gamma_\delta \, f_t\right).$$

APPENDIX: IPCA

The first-order conditions are

$$\hat{f}_t = \left(\hat{\Gamma}_\delta' Z_{t-1}' Z_{t-1} \hat{\Gamma}_\delta\right)^{-1} \hat{\Gamma}_\delta' Z_{t-1}^\top r_t,$$

$$\operatorname{vec}\left(\hat{\Gamma}_\delta'\right) = \left(\sum_{t=1}^{T-1} Z_{t-1}' Z_{t-1} \otimes \hat{f}_t \hat{f}_t'\right)^{-1} \left(\sum_{t=1}^{T-1} \left[Z_{t-1} \otimes \hat{f}_t'\right]' r_t\right).$$

Factor realizations are period-by-period cross section regression coefficients of r_t on the latent loading matrix δ_{t-1}.

Γ_δ is the coefficient of returns regressed on the factors interacted with firm-specific characteristics.

APPENDIX: IPCA

Similarities:

- (Second-stage) factor model relationship and joint fitting.
- Cross-sectional and time-series two step procedures a la Fama MacBeth.
- Efficiency gains from using asset covariates.
- Accommodate unbalanced panels.

Pro Double Lasso:

- Sparse estimation
- Convex objective functions
- Model high dimensional p
- Closed-form inference for target question

Pro IPCA:

- Conceptually simpler optimization
- Fewer assumptions for asymptotic theory
- Rapid estimation

Back to Lit Review Back to Est

APPENDIX: FAMA-MACBETH REGRESSIONS

The classic observable factor model estimation is the Fama and MacBeth (1973) procedure.

We first run N TS regressions for each asset followed by T CS regressions for each time period.

That is, we first estimate $\hat{\beta}_i$ for each asset i by running TS OLS of $\{r_{i,t+1}\}_{t=1}^{T}$ on $\{f_{t+1}\}_{t=1}^{T}$.

Next, we run $\forall t$ the CS OLS of asset excess returns $\{r_{i,t+1}\}_{i=1}^{N}$ on estimated factor loadings $\{\hat{\beta}_i\}_{i=1}^{N}$.

We recover estimates $\hat{\lambda}_t$ for the risk premium $\lambda_t = \mathbb{E}_t[f_{t+1}]$ as well as

the pricing errors from the cross-sectional residuals, $\hat{\alpha}_{i,t+1}$.

Finally, we estimate the parameters of interest: the static risk premium $\hat{\lambda}$ and the static average pricing error $\hat{\alpha}_i$ as the time-series averages of the relevant estimator, $\hat{\lambda}_t$ and $\hat{\alpha}_{i,t+1}$, respectively.

Back

APPENDIX: DSL ESTIMATION PROCEDURE

$$
\begin{aligned}
r_{i,t+1} &= z_{i,t,j} c_{t+1,j} + z_{i,t,-j}^{\top} c_{t+1,-j} + \epsilon_{i,t+1}, \quad & E[\epsilon_{i,t+1} | z_{i,t}] = 0, \\
z_{i,t,j} &= z_{i,t,-j}^{\top} \delta_{t,j} + \epsilon_{i,t,j}^{z}, & E[\epsilon_{i,t,j}^{z} | z_{i,t,-j}] = 0, \\
c_{t+1,j} &:= \Gamma_{\beta,j}^{\top} f_{t+1}.
\end{aligned}
$$

For $\widehat{c}_{t+1,j}$, run $T \times p$ Double Selection Lasso CS regressions $\forall t, j$.

Lasso $\{r_{i,t+1}\}_{i=1}^{N} \to \{z_{i,t}\}_{i=1}^{N}$ for $\widehat{I}_1$ = nonzero elements of $\widehat{c}_t$.

Lasso $\{z_{i,t,j}\}_{i=1}^{N} \to \{z_{i,t,-j}\}_{i=1}^{N}$ for $\widehat{I}_2$ = nonzero elemnts of $\widehat{\delta}_{t,j}$.

Define $\widehat{I} := \widehat{I}_1 \cup \widehat{I}_2 \cup \widehat{I}_3$ where $\widehat{I}_3$ is manually chosen.

OLS $\{r_{i,t}\}_{i=1}^{N}$ on elements of $\{z_{i,t-1}\}_{i=1}^{N}$ in $\widehat{I}$. Back

APPENDIX: ASSUMPTIONS

Assumption (DSL Uniform Consistency)

1. *Bounded Characteristic Portfolios: For a finite absolute constant M and* $\forall t, j$, $\left|c_{t+1,j}\right| = \left|\Gamma_{\beta,j}^{\top} f_{t+1}\right| < M$.
2. *Sparsity rate: The sparsity index obeys* $s^2 \log^2(p \vee N) / \left(\sqrt{N \log(T p)}\right) \leq \delta_{N,T}$. *Additionally,* $\log^3 p/N \leq \delta_{N,T}$.
3. *Weak dependence between the first- and second-stage errors: There exists a positive constant M such that* $\forall\, p, T, N$:

$$\left|\sqrt{\frac{1}{N}} \sum_{i=1}^{N} \epsilon_{i,t,j}^{z} \epsilon_{i,t+1}\right| \leq M \log(T p).$$

4. *Additional standard DSL assumptions in Appendix C.2 of the paper.*

Back

APPENDIX: ASSUMPTIONS

Assumption (Consistency of Latent Factor Model)

5. *Factors:* $\mathbb{E}\left\| f_{t+1}^0 \right\|^4 \leq M < \infty$ *and* $T^{-1}\sum_t f_{t+1}^0 f_{t+1}^{0\top} \to_p \Sigma_f$ *for some* $k \times k$ *positive definite matrix* Σ_f.
6. *Factor Loadings:* $\forall j,\ \left\|\Gamma_{\beta,j}\right\| \leq M < \infty$ *and* $\left\|\Gamma_\beta^\top \Gamma_\beta / p - \Sigma_\Gamma\right\| \to 0$ *for some* $k \times k$ *positive definite matrix* Σ_Γ.

Back

APPENDIX: ASSUMPTIONS

Assumption (Inference)

$\exists$ *a generic absolute constant* $M < \infty$ *such that for all* p, T, N :

7. *Bounded idiosyncratic errors:* $\mathbb{E}[(\sum_t \epsilon_{i,t+1})^2] \leq TM$.
8. *Bounded scaled factor innovations:* $\mathbb{E}[(\sum_t z_{i,t}^\top \Gamma_\beta^0 v_{t+1}^0)^2] \leq sTM$.
9. *Bounded measurement errors:* $\mathbb{E}[(\epsilon_{t+1}^g)^2] \leq M$.

APPENDIX: ASSUMPTIONS

Assumption (Inference)

9. *Convergence of characteristics:*
   $\frac{1}{NT}\sum_i \sum_{t'} \mathbb{E}[z_{i,t,j}]z_{i,t',j'} \to_p \mathcal{Z}_{t,j,j'}$ *uniformly over* t, j, j' *for* $j, j' \in \{1, 2, \ldots, p\}$ *and a nonstochastic finite constant* $\mathcal{Z}_{t,j,j'} \in \mathbb{R}$.
10. *CLT: As* $T \to \infty$,

$$\frac{1}{\sqrt{T}} \sum_t \begin{pmatrix} v_{t+1}^0 \epsilon_{t+1}^g \\ \Pi_t v_{t+1}^0 \end{pmatrix} \xrightarrow{d} \mathcal{N}(0, \Phi)$$

for random matrix $\Pi_t \in \mathbb{R}^{k \times k}$ *and nonstochastic matrix* $\Phi \in \mathbb{R}^{2k \times 2k}$.

APPENDIX: SIMULATION LOW-DIMENSIONAL

p	Parameter	Metric	(1) IPCA	(2) Three-Pass Est.	(3) DSLFM
10	Γ	MSE	0.112526		0.040480
		$Bias^2$	0.020931		0.029007
		Var	0.091596		0.011473
	F	MSE	0.046446	1.023278	1.008919
		$Bias^2$	0.000538	0.006095	0.007407
		Var	0.041890	1.006150	0.992703
	β	MSE	1.736775	0.348060	0.336661
		$Bias^2$	0.051617	0.027838	0.027619
		Var	1.551492	0.008405	0.000433
	C	MSE	0.007724		0.034307
		$Bias^2$	0.000066		0.000184
		Var	0.012636		0.033998
	γ_g	MSE		0.000086	0.000125
		$Bias^2$		0.000003	0.000019
		Var		0.000028	0.000015
		Cov90		0.971000	0.835000
		Cov95		0.990000	0.855000

Back to Simulation Result Summary.

APPENDIX: SIMULATION HIGH-DIMENSIONAL

p	Parameter	Metric	(1) IPCA	(2) Three-Pass Est.	(3) DSLFM
50	Γ	MSE	0.024564		0.009921
		$Bias^2$	0.008984		0.008385
		Var	0.015580		0.001536
	F	MSE	0.223446	1.034021	1.011574
		$Bias^2$	0.009573	0.033910	0.033418
		Var	0.228714	0.989699	0.967504
	β	MSE	4.171191	0.430072	0.396931
		$Bias^2$	0.606915	0.161588	0.155526
		Var	4.084398	0.013159	0.000983
	C	MSE	0.013972		0.007161
		$Bias^2$	0.000751		0.000212
		Var	0.013849		0.007001
	γ_g	MSE		0.015229	0.014656
		$Bias^2$		0.015084	0.014495
		Var		0.000058	0.000069
		Cov90		1.000000	0.828571
		Cov95		1.000000	0.842857

Back to Simulation Result Summary.