

LONG-HISTORY PCA

UNDER DYNAMIC FACTOR MODEL WITH WEAKER LOADINGS

Baeho Kim

BAEHOKIM@KOREA.AC.KR

PROFESSOR OF FINANCE & BUSINESS ANALYTICS

- KOREA UNIVERSITY BUSINESS SCHOOL

VISITING SCHOLAR

- CDAR, UC BERKELEY

JOINT WORK WITH

Robert M. Anderson (UC BERKELEY) AND **Donghan Ryu** (UNIV. OF OXFORD)

WHAT DO WE MEAN BY RISK?

■ Decomposition of financial risk

▶ **First order risk** vs. **Second order risk**

- Shepard (2009, WP) [▶ Details](#)
- Bernardi, Leippold & Lohre (2019, RISK)

▶ **Volatility** vs. **Uncertainty**

- Ait-Sahalia, Matthys, Osambela & Sircar (2024, JoE) [▶ Details](#)
- Anderson, Ghysels & Juergens (2009, JFE)

▶ **Known risk** vs. **Unknown risk**

- Ellsberg (1961, QJE): Ambiguity (= unquantifiable risk)
- Brenner & Izhakian (2018, JFE)

■ Statistical perspectives

▶ **Standard deviation** vs. **Standard error**

[▶ Details](#)

▶ **Variability** (from a distribution) vs. **Model risk & Calibration risk**

- Markowitz (1952) model for portfolio optimization
 - ▶ Our primary focus lies within the estimation of Σ and Σ^{-1}
 - ▶ We don't examine whether the risk factors are priced

$$\begin{aligned} & \text{minimize} && \frac{1}{2} \omega^\top \Sigma \omega \\ & \text{subject to} && \cancel{\mu^\top \mathbf{1} = \mu} \\ & && \omega^\top \mathbf{1} = 1 \end{aligned}$$

■ Second Order Risk (SOR) issues

- ▶ Known to be (extremely) sensitive to parameter estimates $(\hat{\mu}, \hat{\Sigma})$
 - ⇒ Small errors in the estimates of these values may substantially misstate efficient allocations (i.e., error maximization)
 - ⇒ The significance of estimation error on μ diminishes in short-horizon optimization (e.g., by simply assuming $\mu \approx 0$)
- ▶ Mitigating the estimation error of Σ & Σ^{-1} is essential for determining admissible portfolio compositions within a given risk budget
 - ⇒ When the dimension becomes large, the estimation based on historical observations is generally challenging

▶ PCA

▶ Finite-sample error

- The main objective is to mitigate the **second order risk (SOR)** originating from a dynamic factor model at the portfolio level **by extending T** (i.e., $T_M \rightarrow T_L$)
- Specifically, this study aims to ... ► Motivation
 - (i) accurately forecast the **population covariance matrix** (Σ) and the **precision matrix** (Σ^{-1}) of many stock returns (i.e., $N \geq 2000$)
 - (ii) on a **short-term** (i.e., daily) basis
 - (iii) within the **Long-History PCA (LH-PCA)** framework with $T_L (> T_M \sim 250)$ days to mitigate the finite-sample error

- Recognizing (and addressing) SOR bias is crucial but difficult
 - ▶ The **true** covariance matrix is unobservable in reality
 - ▶ We need a proxy measure that is observable in practice
- Extending covariance estimation to the temporal domain introduces nontrivial challenges
 - ▶ Extending T (from T_M to T_L) requires a **Dynamic Factor Model (DFM)** approach
 - ▶ The consistency of the principal component (PC) estimates needs to be verified
- Broad (strong) + narrow (weak) factor issue
 - ▶ Narrow (weaker) factors impact only a portion of the underlying assets and pose greater challenges in detection
 - ▶ PCA blends narrow (weaker) factors with broad (stronger) factors

- The estimation error (e.g., $\|\Sigma - \hat{\Sigma}\|$) is not directly observable in reality
 - ▶ Realized Volatilities (RV) vs. Bias Statistics (BS) of MV portfolios
 - ⇒ BS can serve as a more definitive SOR measure than RV of MV portfolios
- The risk factor structure may not be static over (longer) time
 - ▶ T_M ($\approx 1 \sim 2$ years) with Static Factor Model (SFM) vs. T_L ($\approx 5 \sim 6$ years) with a Dynamic Factor Model (DFM)
 - ⇒ Under DFM, PCA can consistently estimate Σ and Σ^{-1} under the ‘**large- N & large- T** ’ framework (subject to mild regularity conditions)
- The factor strengths may not be homogeneous (strong + weak factors)
 - ▶ Homogeneous (Broad only) vs. Heterogeneous (Broad + Narrow) factors
 - ⇒ PCA with longer history can significantly reduce SOR bias in portfolio optimization with **simulated** and **empirical** data with heterogeneous factor strengths

- This study aims to ...
 - (i) accurately forecast the **population covariance matrix** (Σ) and the **precision matrix** (Σ^{-1}) of many stock returns (i.e., $N \geq 2000$)
 - (ii) on a **short-term** (i.e., daily) basis
 - (iii) within the **Long-History PCA (LH-PCA)** framework with $T_L (> T_M \sim 250)$ days
 - (iv) under **dynamic factor model** with **weaker** loadings
 - (v) across **simulated** and **empirical** datasets

(OBSERVABLE) SOR BIAS MEASURES

MEASUREMENT OF THE ESTIMATION ERROR

- In practice, we can never observe the **true** covariance matrix Σ
 - Instead, we obtain an **estimated** covariance matrix $\hat{\Sigma}$, which may be contaminated by finite-sample estimation errors
 - ▶ The finite-sample estimation errors produce the **excess dispersion bias** in the estimated factor loadings ($T \ll N$)
 - ▶ The excess dispersion bias becomes more pronounced in the presence of the weak (typically narrow) factors
 - ▶ The estimated minimum variance portfolio is substantially more volatile than predicted
- ▶ Sources of excess dispersion bias
- Notations
 - ▶ ω : the **true** GMVP weights based on Σ (unobservable)
 - ▶ $\hat{\omega}$: the **estimated** GMVP weights based on $\hat{\Sigma}$ (observable)
 - ▶ $\sigma(\omega, \Sigma)$: the **true** volatility of ω from Σ (unobservable)
 - ▶ $\sigma(\hat{\omega}, \hat{\Sigma})$: the **predicted** volatility of $\hat{\omega}$ from $\hat{\Sigma}$ (observable)
 - ▶ $\sigma(\hat{\omega}, \Sigma)$: the **actual** (= to-be-realized) volatility of $\hat{\omega}$ from Σ (observable)

- The **true** GMVP and its variance

$$\omega = \frac{\Sigma^{-1}\mathbf{1}}{\mathbf{1}^\top \Sigma^{-1}\mathbf{1}} \quad \Rightarrow \quad \sigma^2(\omega, \Sigma) = \omega^\top \Sigma \omega = \frac{1}{\mathbf{1}^\top \Sigma^{-1}\mathbf{1}}$$

- The **estimated** GMVP and its **predicted** variance

$$\hat{\omega} = \frac{\hat{\Sigma}^{-1}\mathbf{1}}{\mathbf{1}^\top \hat{\Sigma}^{-1}\mathbf{1}} \quad \Rightarrow \quad \sigma^2(\hat{\omega}, \hat{\Sigma}) = \hat{\omega}^\top \hat{\Sigma} \hat{\omega} = \frac{1}{\mathbf{1}^\top \hat{\Sigma}^{-1}\mathbf{1}}$$

- Define $\epsilon \triangleq \hat{\Sigma}^{-1} - \Sigma^{-1}$ as the estimation error of the precision matrix
- The **predicted** variance of the **estimated** GMVP

$$\begin{aligned} (\text{Predicted variance of } \hat{\omega}) &= \sigma^2(\hat{\omega}, \hat{\Sigma}) \\ &= \frac{1}{\mathbf{1}^\top (\Sigma^{-1} + \epsilon) \mathbf{1}} = \frac{1}{\mathbf{1}^\top \Sigma^{-1} \mathbf{1} + \mathbf{1}^\top \epsilon \mathbf{1}} \end{aligned}$$

- A stylized fact is that the finite-sample estimation error in $\hat{\Sigma}^{-1}$ (typically) underpredicts the GMVP volatility as the error increases

$$\sigma^2(\hat{\omega}, \hat{\Sigma}) = \frac{1}{\mathbf{1}^\top \Sigma^{-1} \mathbf{1} + \underbrace{\mathbf{1}^\top \epsilon \mathbf{1}}_{\uparrow \text{ as } \|\epsilon\| \uparrow}} \downarrow \text{ as } \|\epsilon\| \uparrow$$

- Define the volatility ratio (VR) as Derivation

$$(\text{VR}) \triangleq \frac{\sigma(\hat{\omega}, \Sigma)}{\sigma(\hat{\omega}, \hat{\Sigma})} = \frac{(\text{Actual volatility of } \hat{\omega})}{(\text{Predicted volatility of } \hat{\omega})} \approx \sqrt{\frac{1^\top \Sigma^{-1} \mathbf{1} + 2 \cdot 1^\top \epsilon \mathbf{1}}{1^\top \Sigma^{-1} \mathbf{1} + 1^\top \epsilon \mathbf{1}}} \geq 1,$$

which is monotone increasing as $\|\epsilon\|$ gets larger (when $\epsilon \Sigma \epsilon$ negligible)

- Notice that the actual volatility of $\hat{\omega}$ (i.e., $\sigma(\hat{\omega}, \Sigma)$) may not have a monotone relationship with $\|\epsilon\|$ as

$$\sigma(\hat{\omega}, \Sigma) = \underbrace{\sigma(\hat{\omega}, \hat{\Sigma})}_{\substack{\text{red} \\ \downarrow \text{ as } \|\epsilon\| \uparrow}} \cdot \underbrace{(\text{VR})}_{\substack{\text{blue} \\ \uparrow \text{ as } \|\epsilon\| \uparrow}}$$

- This implies that (VR) is a better proxy than $\sigma(\hat{\omega}, \Sigma)$ for measuring the estimation error in $\hat{\Sigma}^{-1}$ across different approaches

A (SIMPLE) NUMERICAL EXPERIMENT

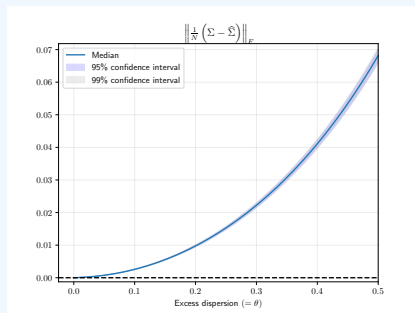
- $N = 2000$ stocks in the market
- $K = 5$ uncorrelated factors
 - ▶ Idiosyncratic returns are uncorrelated with each other as well as with the factor returns
- True factor loadings are drawn from a normal distribution with the true dispersion θ_0
- Construct the estimated covariance matrix $\hat{\Sigma}$ contaminated by adding an excess dispersion θ

$$\text{(Total dispersion)} = \underbrace{\text{(True dispersion)}}_{=\theta_0} + \underbrace{\text{(Excess dispersion)}}_{=\theta}$$

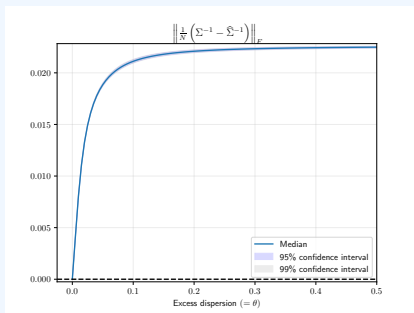
- Scenarios drawn from 1000 different seeds

A SIMPLE NUMERICAL EXPERIMENT (CONT.)

(Average) Frobenious norm of the estimation error



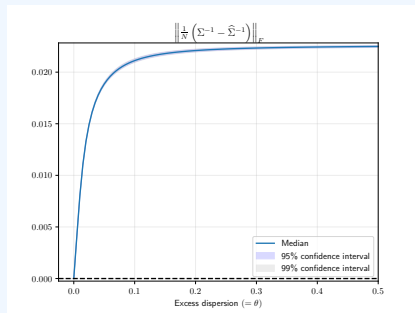
(a) Covariance matrix error



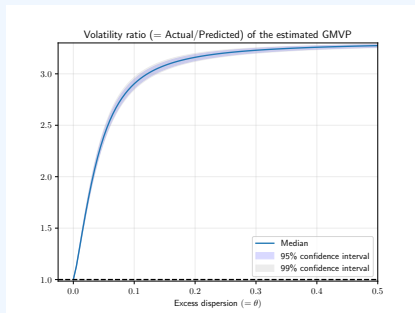
(b) Precision matrix error

A SIMPLE NUMERICAL EXPERIMENT (CONT.)

Volatility ratio is a definitive & observable measure of the error



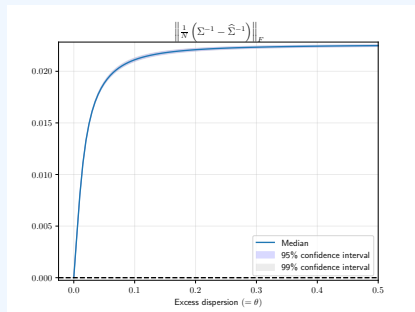
(a) (Ideal) SOR Bias



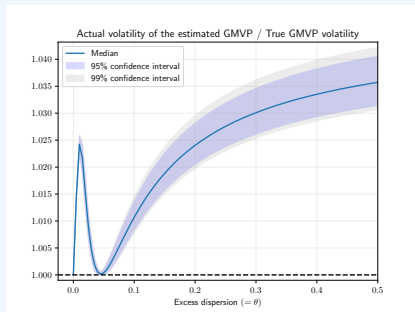
$$(b) \frac{(\text{Actual vol.})_{MV}}{(\text{Predicted vol.})_{MV}} = \frac{\sigma(\hat{\omega}, \Sigma)}{\sigma(\hat{\omega}, \hat{\Sigma})}$$

A SIMPLE NUMERICAL EXPERIMENT (CONT.)

Actual volatility may be misleading to measure the estimation error



(a) (Ideal) SOR Bias



(b) $\frac{(\text{Actual vol.})_{MV}}{(\text{True vol.})_{MV}} = \frac{\sigma(\hat{\omega}, \Sigma)}{\sigma(\omega, \Sigma)}$

AN 'ILLUSION' OF THE ACTUAL VARIANCE

- For $\Sigma \neq \hat{\Sigma}$, suppose that $\Sigma \hat{\Sigma}^{-1}$ **has an eigenvector close to the vector of ones**; i.e., $\Sigma \hat{\Sigma}^{-1} \mathbf{1} \approx \lambda \mathbf{1}$ for some $\lambda > 0$
- In this case, we have

$$\mathbf{1}^\top \hat{\Sigma}^{-1} \Sigma \hat{\Sigma}^{-1} \mathbf{1} \approx \frac{(\mathbf{1}^\top \hat{\Sigma}^{-1} \mathbf{1})^2}{\mathbf{1}^\top \Sigma^{-1} \mathbf{1}}$$

- This implies that

$$\begin{aligned} \text{(Actual variance of } \hat{\omega}) &= \frac{\mathbf{1}^\top \hat{\Sigma}^{-1} \Sigma \hat{\Sigma}^{-1} \mathbf{1}}{(\mathbf{1}^\top \hat{\Sigma}^{-1} \mathbf{1})^2} \\ &\approx \frac{(\mathbf{1}^\top \hat{\Sigma}^{-1} \mathbf{1})^2}{\mathbf{1}^\top \Sigma^{-1} \mathbf{1}} \frac{1}{(\mathbf{1}^\top \hat{\Sigma}^{-1} \mathbf{1})^2} \\ &= \frac{1}{\mathbf{1}^\top \Sigma^{-1} \mathbf{1}} = \text{(The variance of the True GMVP)} \end{aligned}$$

- We define $\|\epsilon\| = \|\hat{\Sigma}^{-1} - \Sigma^{-1}\|$ as the (ideal) SOR bias measure, which is unobservable in practice
 - ▶ When $\|\epsilon\| > 0$ underestimates the predicted volatility of the estimated GMVP, the volatility ratio (VR) is a good proxy of $\|\epsilon\|$
 - ▶ Bias Statistic (BS) is a statistical proxy for VR in the empirical analysis
 - ▶ If the true covariance matrix Σ is time-varying, MRAD and/or Q-statistics may be better than BS for measuring the SOR bias for empirical studies [▶ Definitions of BS / MRAD / Q-statistics](#)
- On the other hand, the actual volatility, which can be statistically proxied by the realized volatility (RV), of the estimated GMVP cannot serve as a definitive measure of $\|\epsilon\|$

LONG-HISTORY PCA (LH-PCA)

(RQ1) Under the ‘large- N and large- T ’ framework, can PCA be a consistent method for estimating the dynamic factor model based on variable factor and idiosyncratic volatility structure with time-varying factor loadings?

⇒ **Yes, under reasonably mild assumptions.**

▶ Dynamic factor model

(RQ2) Does the LH-PCA approach with T_L outperform variants of the traditional PCA method with T_M in mitigating the SOR bias in the presence of both broad (strong) and narrow (weak) factors?

⇒ **Yes, our simulation study and empirical findings support this result.**

▶ Weak factor model

- We show that **LH-PCA** consistently estimates factor loadings under dynamic factor models
 - ▶ Latent factor model
 - ▶ DFM
 - ▶ Weaker loadings
 - ▶ It justifies the use of PCA in the setting of variable factor and idiosyncratic volatilities with **dynamic & weaker** factor loadings
- We demonstrate, both in simulation and empirically, that the use of **long histories** (T_L) substantially mitigates the SOR bias
 - ▶ We estimate factor loadings with a longer history ($T_L \sim 1500$ days)
 - ▶ ... and predict the portfolio volatility on the next day using the **Responsive Covariance Adjustment (RCA)** scheme with a short data history (half-life $T_S = 40$ days)
 - ▶ RCA

MAIN THEOREM IN THE 'LARGE- N & LARGE- T ' FRAMEWORK

- Consider the $N \times N$ *population* covariance matrix within the observation time window T given by

$$\Sigma_N(T) = \frac{(R_{T,N})^\top (R_{T,N})}{T}$$

and define

$$\bar{\Sigma}_N(T) = \bar{X}_N^\top \text{diag} \left(\frac{1}{T} \sum_{t=1}^T (\mu_t)^2 \right) \bar{X}_N + \text{diag} \left(\frac{1}{T} \sum_{t=1}^T (v_t)^2 \right).$$

Theorem (Convergence of population covariance matrices)

By letting $N, T \rightarrow \infty$, we have

$$\left\| \frac{1}{N} \left(\Sigma_N(T) - \bar{\Sigma}_N(T) \right) \right\|_F^2 = O_p \left(\frac{1}{\min\{N, T\}} \right)$$

FOR PRECISION (= INVERSE COVARIANCE) MATRICES

- For all N, T , suppose that both $\Sigma_N(T)$ and $\bar{\Sigma}_N(T)$ are invertible, and there exists some $\varepsilon > 0$ such that

$$\min \left\{ \lambda_{\min}(\Sigma_N(T)), \lambda_{\min}(\bar{\Sigma}_N(T)) \right\} > \varepsilon,$$

where $\lambda_{\min}(A)$ represents the smallest eigenvalue of A

Corollary (Convergence of precision matrices)

By letting $N, T \rightarrow \infty$, we have

$$\left\| \frac{1}{N} \left(\bar{\Sigma}_N^{-1}(T) - \Sigma_N^{-1}(T) \right) \right\|_F^2 = O_p \left(\frac{1}{\min\{N, T\}} \right).$$

(Note) The additional assumption regarding the uniform boundedness (away from zero) of the minimum eigenvalue is consistent with the presence of a fixed number of factors with non-zero factor volatilities.

- Justification of using (LH-)PCA for estimating \bar{X}_N

Corollary (Variable volatility with time-varying factor loadings)

Fix $k \in \{1, \dots, K\}$. By letting $N, T \rightarrow \infty$, we can choose the k^{th} eigenvector of

$$\Sigma_N(T) - \text{diag} \left(\sum_{t=1}^T \frac{(v_{t \cdot})^2}{T} \right)$$

converging in probability to the k^{th} row of $(\bar{X}_N)_k$.

(Note) The average error in estimating \bar{X}_N by (LH-)PCA vanishes at the rate

$$\mathcal{O}_p \left(\frac{1}{\min\{N^{1+\alpha_K}, T\}} \right).$$

Refer to **Proposition 2** of Bai & Ng (2023).

SIMULATION STUDY

■ True (Strong & Weak) Factor Structure

- ▶ $N = 2000$ stocks in the market
- ▶ 4 Broad & 27 (= 11+16) Narrow Factors

▶ Factor structure

■ Variable Volatility Factor Structure

- ▶ Markov Regime Switching (MRS)

▶ MRS

■ Time-varying Factor-loading dynamics

- ▶ Mean-reverting Ornstein–Uhlenbeck process

▶ Mean-reverting Process

■ We institute the two-history scheme to forecast the Σ and Σ^{-1} that will be realized **on the next day**

▶ RCA

- ▶ Estimated factor loadings come from LH-PCA over $T_L = 1500$ days window
- ▶ Estimated GMVP based on RCA from EWMA with $T_S = 40$ -day half-life

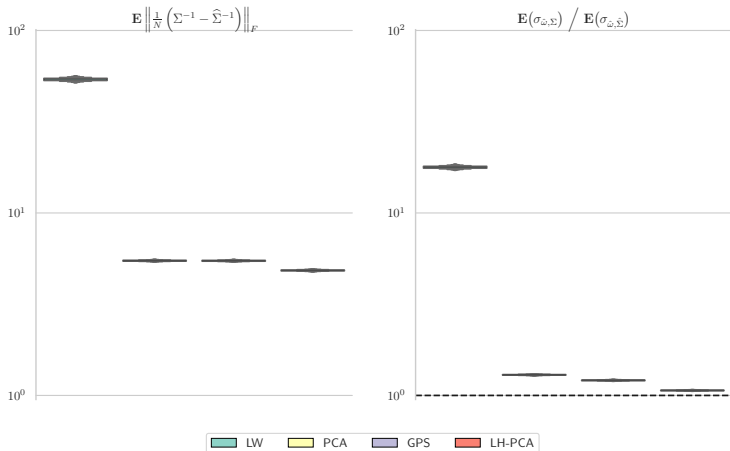
■ Number of (eigen-)factors to extract

▶ Number of factors

- ▶ In our study, we employ the Bai & Ng (2002) estimator and determine the number of PCA eigenfactors within each moving window

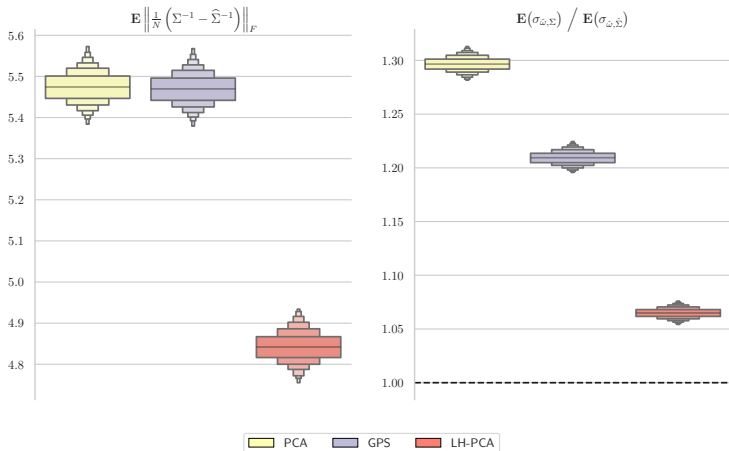
EXTENDED SIMULATION (CONT.)

Prediction error vs. Volatility Ratio (Actual / Predicted); 1000 seeds



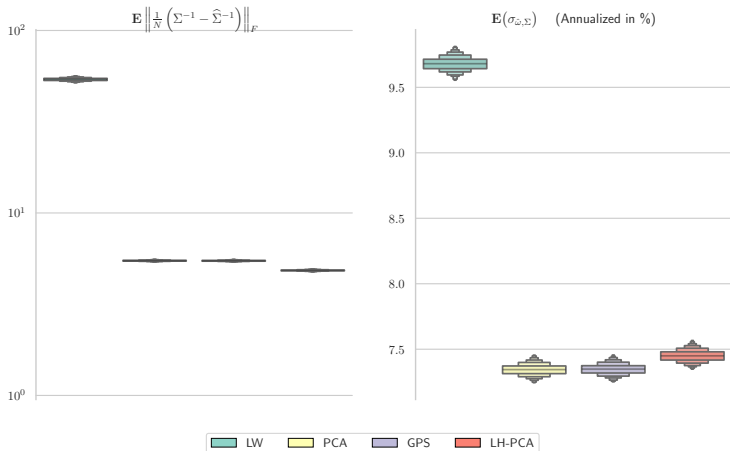
EXTENDED SIMULATION (CONT.)

Prediction error vs. Volatility Ratio (Actual / Predicted); 1000 seeds



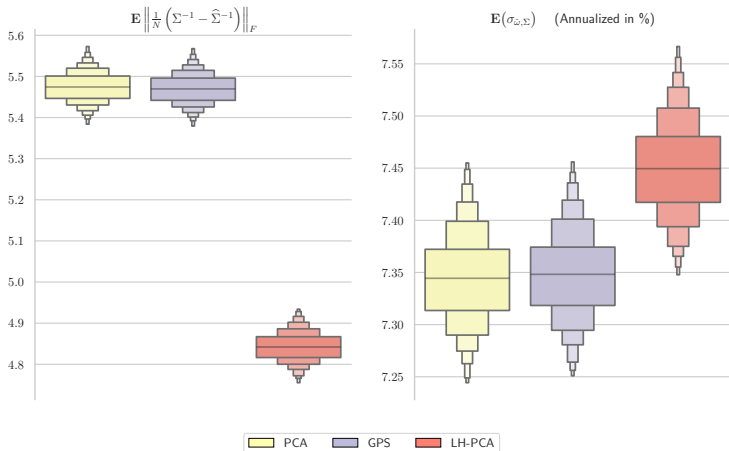
EXTENDED SIMULATION (CONT.)

Prediction error vs. Actual vol. of the estimated GMVP; 1000 seeds



EXTENDED SIMULATION (CONT.)

Prediction error vs. Actual vol. of the estimated GMVP; 1000 seeds

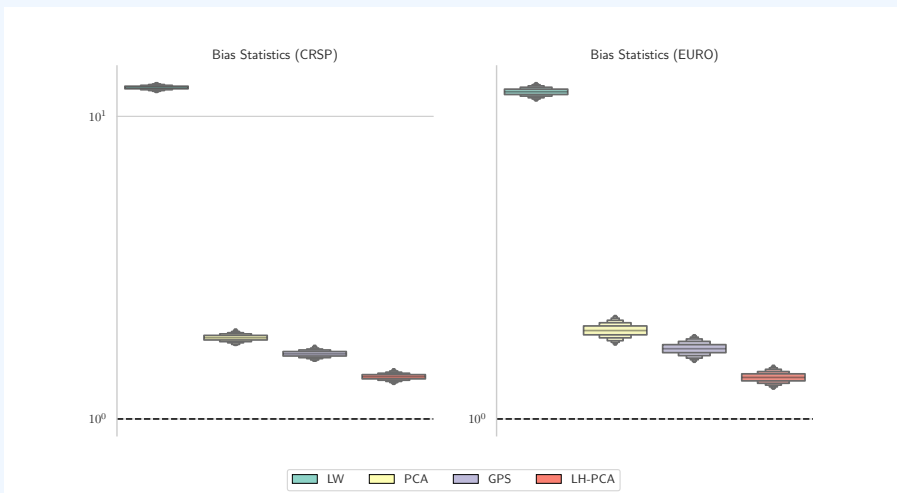


EMPIRICAL ANALYSIS

- Datasets: CRSP and EURO (Compustat Global) data 2001-2021 (21 years)
 - ▶ We need six-year history to make predictions, so predictions are for 15 years 2007-2021
- We consider overlapping six-year windows, including stocks that are present throughout each window (Average: 2,260 CRSP and 2,520 EURO)
 - ▶ Our analysis is largely free of **survivorship bias** by addressing delisting issues
 - ▶ All of the losses leading up to the demise of a stock were included in our analysis for the appropriate day
 - ▶ Details on the data cleaning procedures
- Confidence intervals can be approximated by bootstrapping samples of the realized Z-scores with replacement for cross-validation
 - ▶ Winsorized daily Z-scores at the 0.25th and 99.75th percentiles to mitigate the distortionary impact of outliers

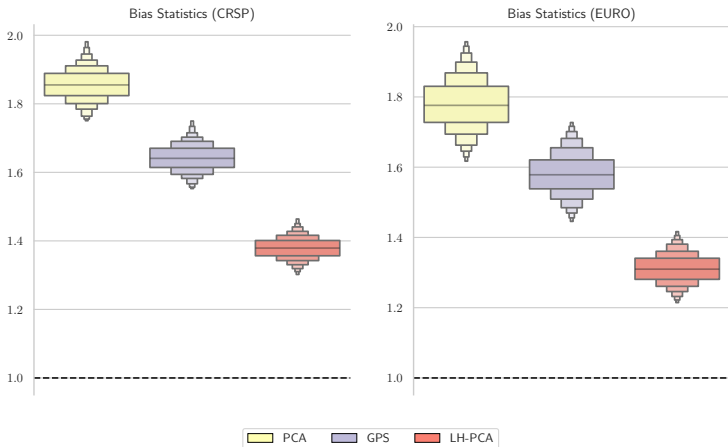
EMPIRICAL ANALYSIS (CONT.)

Bias Statistics (CRSP & EURO; Bootstrapped)



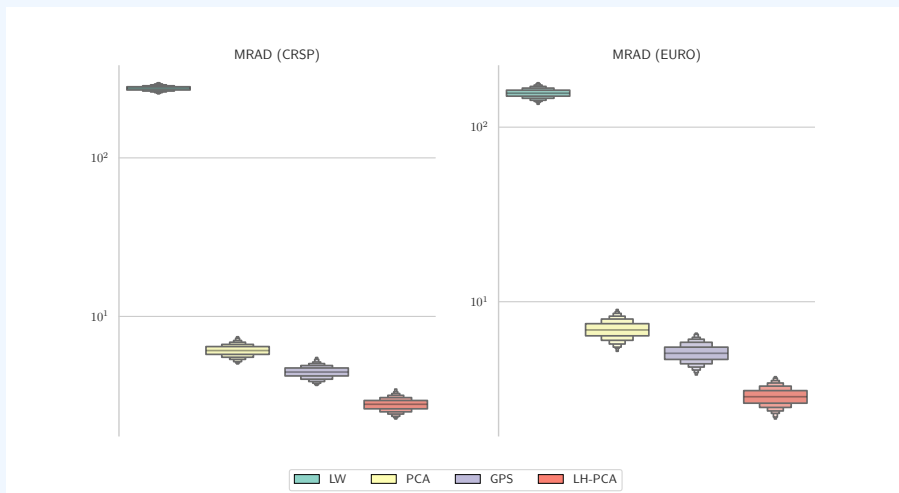
EMPIRICAL ANALYSIS (CONT.)

Bias Statistics (CRSP & EURO; Bootstrapped)



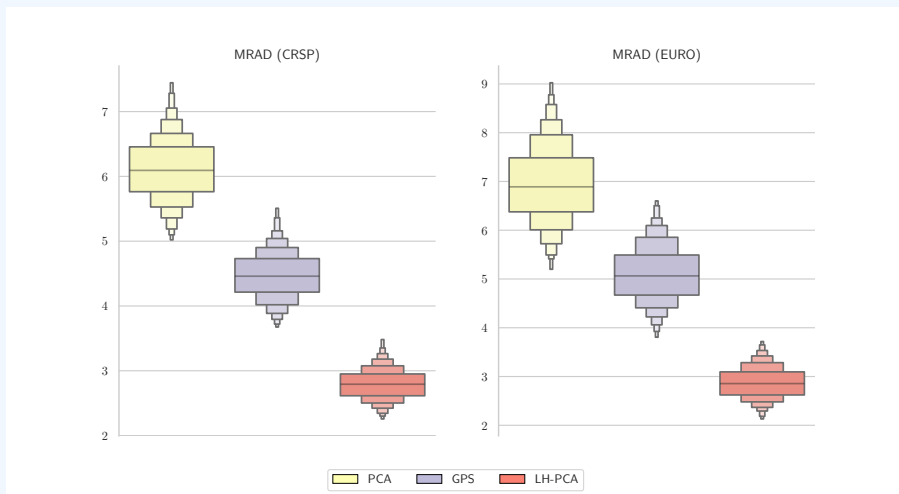
EMPIRICAL ANALYSIS (CONT.)

MRAD (CRSP & EURO; Bootstrapped)



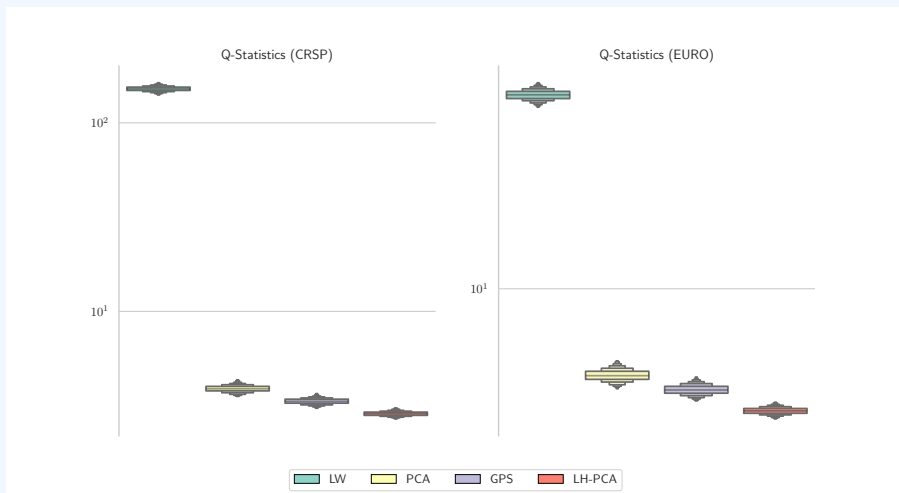
EMPIRICAL ANALYSIS (CONT.)

MRAD (CRSP & EURO; Bootstrapped)



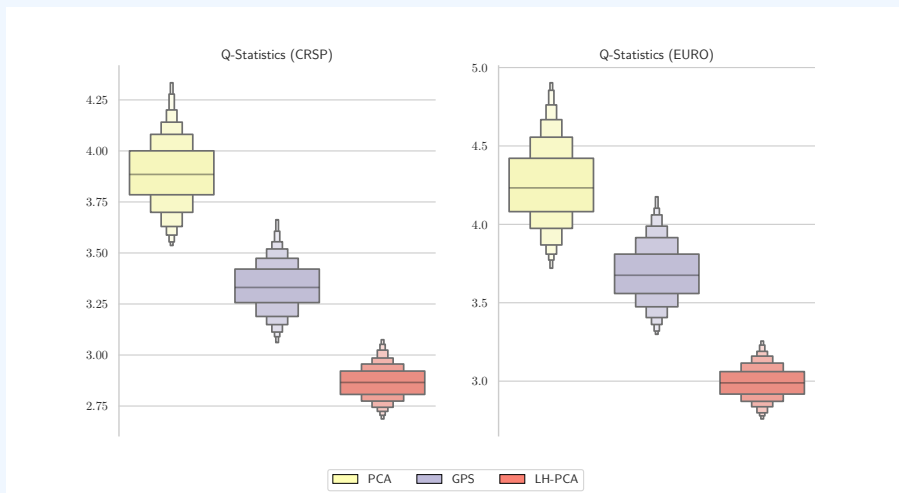
EMPIRICAL ANALYSIS (CONT.)

Q-Statistics (CRSP & EURO; Bootstrapped)



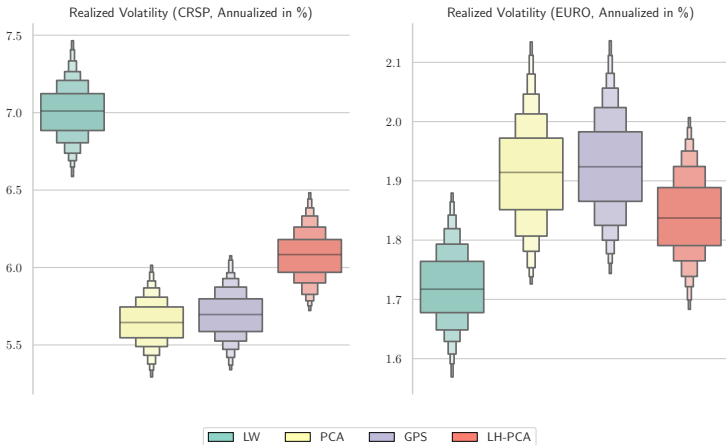
EMPIRICAL ANALYSIS (CONT.)

Q-Statistics (CRSP & EURO; Bootstrapped)



EMPIRICAL ANALYSIS (CONT.)

Realized Volatility (CRSP & EURO; Bootstrapped)
















CONCLUSION & FUTURE RESEARCH

- When $\|\epsilon\| = \|\hat{\Sigma}^{-1} - \Sigma^{-1}\| > 0$ underestimates the predicted volatility of the estimated GMVP, the volatility ratio (VR), which can be statistically approximated by the bias statistics (BS), is a good proxy of $\|\epsilon\|$
 - ▶ The actual volatility, which can be statistically proxied by the realized volatility (RV), of the estimated GMVP cannot serve as a definitive SOR measure
- Variants of PCA with a one-year history of data generally performs poorly in estimating Σ and Σ^{-1} when N gets larger with our empirical datasets
- We show that **LH-PCA** can consistently estimate Σ^{-1} under the dynamic factor structure with arbitrary variable volatility of factors and time-varying & weaker loadings
 - ▶ Theoretical justification + Simulation + Empirical evidence
- Future research topics
 - ▶ LH-PCA leaves room for further improvements
 - ⇒ Optimal choice of T_L guided by data + Eigenvector corrections
 - ⇒ Sparse Dictionary Learning to better estimate the weak (narrow) factors
 - ▶ Asset pricing implications by examining the (weaker) factor prices

Thank you!

REFERENCES I

-  Sahalia, Yacine, Felix Matthys, Emilio Osambela & Ronnie Sircar (2024), 'When uncertainty and volatility are disconnected: Implications for asset pricing and portfolio performance', *Journal of Econometrics*, p. In Press.
-  Anderson, Evan W., Eric Ghysels & Jennifer L. Juergens (2009), 'The impact of risk and uncertainty on expected returns', *Journal of Financial Economics* **94**(2), 233–263.
-  Bai, Jushan & Serena Ng (2002), 'Determining the Number of Factors in Approximate Factor Models', *Econometrica* **70**(1), 191–221.
-  Bai, Jushan & Serena Ng (2023), 'Approximate factor models with weaker loadings', *Journal of Econometrics* **235**, 1893–1916.
-  Bekaert, Geert, Eric Engstrom & Yuhang Xing (2009), 'Risk, uncertainty, and asset prices', *Journal of Financial Economics* **91**(1), 59–81.
-  Bernardi, Simone, Markus Leippold & Harald Lohre (2019), 'Second-order risk of alternative risk parity strategies', *Risk* **21**(3), 1–25.
-  Brenner, Menachem & Yehuda Izhakian (2018), 'Asset pricing and ambiguity: Empirical evidence', *Journal of Financial Economics* **130**(3), 503–531.
-  Ellsberg, Daniel (1961), 'Risk, ambiguity, and the savage axioms', *The Quarterly Journal of Economics* **75**(4), 643–669.
-  Freyaldenhoven, Simon (2022), 'Factor models with local factors: Determining the number of relevant factors', *Journal of Econometrics* **229**(1), 80–102.

-  Goldberg, Lisa, Alex Papanicolaou & Alex Shkolnik (2022), 'The Dispersion Bias', *SIAM Journal of Financial Mathematics* **13**(2), 521–550.
-  Ledoit, Olivier & Michael Wolf (2004), 'A well-conditioned estimator for large-dimensional covariance matrices', *Journal of Multivariate Analysis* **88**(2), 365–411.
-  Shepard, Peter (2009), 'Second Order Risk', *Working Paper, MSCI Barra* .
-  Uematsu, Yoshimasa & Takashi Yamagata (2023), 'Estimation of sparsity induced weak factor models', *Journal of Business & Economic Statistics* **41**(1), 213–227.

APPENDIX

FIRST ORDER RISK VS. SECOND ORDER RISK

*“... Classical finance assumes the markets to be like a game of chance: Although future events are uncertain, the distribution of these events is known. ... Unfortunately, real financial markets do not behave like a game of chance. ... Our **estimates of financial risk are uncertain**, based on limited historical observation, extrapolated forward.”*

*“... Managing a portfolio to a risk model can tilt the portfolio toward weaknesses of the model. As a result, the optimized portfolio acquires downside exposure to uncertainty in the model itself, what we call **second order risk**.”*

- Shepard (2009)

“Second Order Risk”

▶ Return

*“... Although the notions of uncertainty and volatility are often used interchangeably, the two concepts are inherently different: **volatility** measures the dispersion of short-term shocks around a long-term mean, while **uncertainty** measures the difficulty to forecast the distribution of returns, including its long-term mean.”*

- Aït-Sahalia, Matthys, Osambela and Sircar (2024)

“When Uncertainty and Volatility Are Disconnected: Implications for Asset Pricing and Portfolio Performance”

STANDARD DEVIATION VS. STANDARD ERROR

*“... The **standard deviation** is a measure of the dispersion, or scatter, of the data. ... In contrast, the **standard error** provides an estimate of the precision of a parameter (such as a mean, proportion, odds ratio, survival probability, etc) and is used when one wants to make inferences about data from a sample to some relevant population.”*

- Biau (2011)

“In Brief: Standard Deviation and Standard Error”

[▶ Return](#)

- In practice, the estimation typically focuses on a set of **latent** factors that are computationally convenient (but less interpretable), characterized by orthogonal exposure vectors with the identity covariance matrix
 - ▶ For a given number of securities at each time, without loss of generality, one can apply Gram-Schmidt to transform a correlated factor structure into a new **latent** factor representation,
 - ▶ ... where (latent) factors are uncorrelated in population and exhibit orthogonal (latent) factor loadings
- Latent factor models are extensively used by practitioners in finance
 - ▶ Variants of PCA, a long-established method for dimension reduction, are used in commercially available latent factor models [▶ Return](#)

PCA ESTIMATION FOR LINEAR FACTOR MODELS

- PCA is attractive for predicting short-term risk
- The traditional PCA approaches require temporal stability, assuming a *static* factor structure
 - ▶ In practice, risk factor structure changes its shape rapidly over time indicating a dynamic evolution in the underlying factor structure
- Most practitioners are unwilling to run PCA over data histories longer than one or two years for large portfolios ($T \ll N$)
 - ▶ An exception is Northfield, which uses a hybrid model in which stock returns are first regressed on various fundamental factors
 - ▶ It uses exponentially weighted regressions over 60 months of monthly data; it then extracts five PCA factors from the residuals

▶ Return

ESTIMATION ERROR OF Σ IN FINITE SAMPLES

- Finite-sample estimation error
 - ▶ The **sample covariance matrix** based on the observed data is singular when the dimension (N) is larger than the sample size (T)
 - ▶ (e.g.) Shrinkage estimates of the sample covariance matrix (Ledoit & Wolf 2004, LW)
- PCA estimation of (linear) factor models ($T_M \approx 250 \sim 500$ trading days)
 - ▶ Linear factor model estimated by PCA with T_M (\rightarrow Excess dispersion bias)
 - ▶ Correction of the leading eigenvector (Goldberg, Papanicolaou & Shkolnik 2022, GPS) based on PCA with T_M
- Commercially available factor models typically use a two-history algorithm
 - ▶ A variant of PCA over a medium data history of T_M to estimate the factor loadings (+ bias correction)
 - ▶ Exponentially Weighted Moving Average (EWMA) volatility model with a short half-life, such as $T_S \sim 40$ days, to estimate current factor variances
- We empirically observe that variants of this approach with T_M suffer from a significant amount of SOR bias [▶ Return](#)

■ Motivation

▶ Return

- (i) Estimating Σ is a fundamental problem in statistics and finance (e.g.) statistical inference, efficient asset allocation, portfolio risk management, ...
 - Estimated Σ^{-1} plays the pivotal role in determining the optimized minimum-variance portfolios

- (ii) Sometimes volatility changes become too rapid and extreme to be reliably captured from low frequency observations
 - Daily risk predictions are appropriate for short-term investors, such as hedge funds and other leveraged institutional investors
 - In addition, risk predictions at longer horizons typically take short-term volatility as an input

■ Bates et al. (2013) \Rightarrow (RQ1)

[Return](#)



Contents lists available at [ScienceDirect](#)

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom



Consistent factor estimation in dynamic factor models with structural instability[☆]



Brandon J. Bates^a, Mikkel Plagborg-Møller^b, James H. Stock^b, Mark W. Watson^{c,*}

^a BlackRock, Inc., United States

^b Harvard University, United States

^c Princeton University, United States

ARTICLE INFO

Article history:

Available online 17 April 2013

ABSTRACT

This paper considers the estimation of approximate dynamic factor models when there is temporal instability in the factor loadings. We characterize the type and magnitude of instabilities under which the principal components estimator of the factors is consistent and find that these instabilities can be larger than earlier theoretical calculations suggest. We also discuss implications of our results for the robustness of regressions based on the estimated factors and of estimates of the number of factors in the presence of parameter instability. Simulations calibrated to an empirical application indicate that instability in the factor loadings has a limited impact on estimation of the factor space and diffusion index forecasting, whereas estimation of the number of factors is more substantially affected.

© 2013 Elsevier B.V. All rights reserved.

■ Bai and Ng (2023) \Rightarrow (RQ2)

▶ Return



ELSEVIER

Contents lists available at [ScienceDirect](#)

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom



Approximate factor models with weaker loadings[☆]

Jushan Bai^a, Serena Ng^{b,*}

^a Columbia University, 420 W. 118 St. MC 3308, New York, NY 10027, United States of America

^b Columbia University and NBER, 420 W. 118 St. MC 3308, New York, NY 10027, United States of America



ARTICLE INFO

Article history:

Accepted 10 January 2023

Available online 13 March 2023

JEL classification:

C30

C31

Keywords:

Principal components

Low rank decomposition

Weak factors

Factor augmented regressions

ABSTRACT

Pervasive cross-section dependence is increasingly recognized as a characteristic of economic data and the approximate factor model provides a useful framework for analysis. Assuming a strong factor structure where $A^0 A^0/N^\alpha$ is positive definite in the limit when $\alpha = 1$, early work established convergence of the principal component estimates of the factors and loadings up to a rotation matrix. This paper shows that the estimates are still consistent and asymptotically normal when $\alpha \in (0, 1]$ albeit at slower rates and under additional assumptions on the sample size. The results hold whether α is constant or varies across factor loadings. The framework developed for heterogeneous loadings and the simplified proofs that can be also used in strong factor analysis are of independent interest.

© 2023 Elsevier B.V. All rights reserved.

DYNAMIC FACTOR MODEL OF STOCK RETURNS

- For some $T, N \in \mathbb{N}$, the observable return matrix is given by $R_{T,N} \in \mathbb{R}^{T \times N}$,
 - ▶ At time $t \in \{1, \dots, T\}$, the return of the security $n \in \{1, \dots, N\}$ is generated by the linear (latent) factor model as

$$R_{tn} = \sum_{k=1}^K \phi_{tk} (X_t)_{kn} + \varepsilon_{tn},$$

where $X_t \in \mathbb{R}^{K \times N}$ is the time- t factor exposures corresponding to the N tradable securities

- Factor returns (ϕ_{tk}) + idiosyncratic returns (ε_{tn}):
 - ▶ We assume that $\mathbf{E}(\phi_{tk}) = 0$ and $\mathbf{Var}(\phi_{tk}^2) = \mu_{tk}^2$ for $k = 1, \dots, K$
 - ▶ $\mu_{tk} \in [0, M]$ is the factor volatility of ϕ_{tk}
 - ▶ We further assume that $\mathbf{Var}(\varepsilon_{tn}) = v_{tn}^2$, where $v_{tn} \in [0, M]$ is the idiosyncratic volatility of ε_{tn}

DYNAMIC FACTOR MODEL OF STOCK RETURNS (CONT.)

- Time-varying factor loadings (X_t): (Bates et al., 2013)
 - ▶ For each t , we follow Bates et al. (2013) by specifying the dynamics of X_t as

$$X_t = \bar{X}_N + h_{NT} \xi_t^N,$$

where $h_{NT} \geq 0$ may depend on the pair of (N, T) and $\xi_t^N \in \mathbb{R}^{K \times N}$ is a (possibly degenerate) stochastic process

- Bates et al. (2013) show that some mild regularity conditions satisfy (i) white noise, (ii) random walk, (iii) (single) large break of X_t with practically reasonable form of h_{NT} :
 - ▶ Conditions in **Assumption 4 (Factor Loading Innovations)**
 - ▶ Assumptions of **Corollary 1**

▶ Return

STRONG & WEAK (BASELINE) FACTOR LOADINGS

- Strong (only) factor loadings (Bai & Ng, 2002, and many others)
 - ▶ For each $N \in \mathbb{N}$, \bar{X}_N has rank K and the rows of \bar{X}_N are orthogonal, and there exists some diagonal matrix $D \in \mathbb{R}^{K \times K}$ such that

$$\frac{1}{N} \bar{X}_N \bar{X}_N^\top \xrightarrow{p} D \quad \text{as } N \rightarrow \infty$$

- Strong + Weaker factor loadings (Bai & Ng, 2023)

- ▶ Intuition: For some $\alpha \in (0, 1]$, $\left(\frac{1}{N} \rightarrow \frac{1}{N^\alpha}\right)$

$$\frac{1}{N^\alpha} \bar{X}_N \bar{X}_N^\top \xrightarrow{p} D \quad \text{as } N \rightarrow \infty$$

- ▶ Allowing the strengths of baseline loadings to vary across factors, let $1 \geq \alpha_1 \geq \dots \geq \alpha_K > 0$ so that the weakest baseline loading has strength $\alpha_K \in (0, 1]$
- ▶ Define the $K \times K$ normalization matrix

$$B_N = \text{diag}\left(N^{\frac{\alpha_1}{2}}, \dots, N^{\frac{\alpha_K}{2}}\right)$$

- ▶ There exists some diagonal matrix $D \in \mathbb{R}^{K \times K}$ such that

▶ Return

$$B_N^{-1} \bar{X}_N \bar{X}_N^\top B_N^{-1} \xrightarrow{p} D \quad \text{as } N \rightarrow \infty$$

ASSUMPTIONS OF X_t

- There exist envelope functions $Q_1(N, T)$, $Q_2(N, T)$ and $Q_3(N, T)$ such that the following conditions hold for all N, T and factor indices $p, q, r, \ell = 1, \dots, K$

$$\sup_{s,t \leq T} \sum_{i,j=1}^N \left| \mathbf{E} \left((\xi_s^N)_{ip} (\xi_t^N)_{jq} \phi_{sp} \phi_{tq} \right) \right| \leq Q_1(N, T),$$

$$\sum_{s,t=1}^T \sum_{i,j=1}^N \left| \mathbf{E} \left((\xi_s^N)_{ip} (\xi_s^N)_{jq} \phi_{sp} \phi_{sq} \phi_{tr} \phi_{t\ell} \right) \right| \leq Q_2(N, T),$$

$$\sum_{s,t=1}^T \sum_{i,j=1}^N \left| \mathbf{E} \left((\xi_s^N)_{ip} (\xi_s^N)_{jq} (\xi_t^N)_{ir} (\xi_t^N)_{j\ell} \phi_{sp} \phi_{sq} \phi_{tr} \phi_{t\ell} \right) \right| \leq Q_3(N, T),$$

- The following conditions hold:

$$h_{NT}^2 Q_1(N, T) = \mathcal{O}(N)$$

$$h_{NT}^2 Q_2(N, T) = \mathcal{O}(NT^2)$$

$$\min\{N, T\} h_{NT}^4 Q_3(N, T) = \mathcal{O}(N^2 T^2)$$

- Kan and Smith (2008)
 - ▶ ... prove under the i.i.d. normality assumption of returns that the **finite-sample** minimum-variance frontier is a significantly biased estimator of the true population frontier
 - ▶ As such, optimizing a stock portfolio using an estimated (sample) covariance matrix, or an estimated covariance matrix that has been dimensionally reduced by PCA, is problematic in terms of the out-of-sample performance
- Ledoit and P ech e (2011) and Wang and Fan (2017) among others
 - ▶ ... propose corrections of the **eigenvalues** (= variances) to mitigate this problem
- Goldberg, Papanicalaou and Shkolnik (2022; hereafter GPS)
 - ▶ ... identified excess dispersion in the estimated dominant eigenvector as a key source of the problems with optimized portfolios
 - ▶ They propose a correction to the estimated **eigenvector** (= exposure) corresponding to the largest eigenvalue
 - ▶ ... in a theoretical one-factor model and in a simulation involving only broad factors

SOURCES OF DISPERSION BIAS: FACTOR LOADING ESTIMATION

- PCA estimation: $R \sim VW$ by minimizing $\|R - VW\|_2^2$ ▶ Return
 - ▶ V is a $T \times K$ matrix of estimated factor returns with $|V_{.k}| = 1$
 - ▶ W is a $K \times N$ matrix of estimated stock factor loadings
 - ▶ U is an $N \times K$ matrix, where $V = RU$

$$W = U^T R^T R = U^T \left(X^T \phi^T \phi X + X^T \phi^T \varepsilon + (X^T \phi^T \varepsilon)^T + \varepsilon^T \varepsilon \right)$$

- By assumption, factor returns and idiosyncratic returns are uncorrelated:

$$\phi^T \varepsilon \equiv 0$$

- In a finite sample, the factor returns and idiosyncratic returns appear correlated, and

$$\phi^T \varepsilon \neq 0$$

- ▶ The estimated factor loadings are contaminated and exhibit excess dispersion in finite sample
- ▶ Estimated factor loadings of some stocks appear to be smaller than they actually are, and the optimizer will choose to increase their weights in the portfolio (and vice versa) \Rightarrow Underestimation of MV portfolio risk

SOURCES OF DISPERSION BIAS: NARROW FACTORS

- Narrow (weak) factors cannot be captured cleanly by PCA
 - ▶ To properly evaluate the risk of a portfolio, and especially to produce optimized portfolios, we need our estimated covariance matrix to accurately reflect the **narrow factors**
- When $K \ll T \ll N$, the idiosyncratic returns of some stocks will appear to be very significantly correlated with the returns of some weak factors
 - ▶ The finite-sample error contaminates the narrow and broad factor loadings together, but the **relative contamination of the narrow factor is much greater** than that of the broad (e.g., market) factor
 - ▶ This contamination assigns nonzero estimated factor loadings on the narrow factors, when the stocks are NOT exposed to them
 - ▶ One can think of this as excess dispersion of the estimated factor loadings that are, in fact, zero (= infinite relative error)

▶ Return

ACTUAL VARIANCE OF THE ESTIMATED GMVP

- Define $\tilde{\Sigma} \triangleq \hat{\Sigma}\Sigma^{-1}\hat{\Sigma}$ as the *sandwich* covariance matrix
- The **actual** variance of the **estimated** GMVP

$$\begin{aligned}(\text{Actual variance of } \hat{\omega}) &= \sigma^2(\hat{\omega}, \Sigma) \\ &= \hat{\omega}^\top \Sigma \hat{\omega} \\ &= \frac{1^\top \hat{\Sigma}^{-1} \Sigma \hat{\Sigma}^{-1} 1}{(1^\top \hat{\Sigma}^{-1} 1)^2} \\ &= \frac{1^\top \tilde{\Sigma}^{-1} 1}{(1^\top \hat{\Sigma}^{-1} 1)^2} \\ &= \frac{\sigma^4(\hat{\omega}, \hat{\Sigma})}{\sigma^2(\tilde{\omega}, \tilde{\Sigma})} = \frac{(\text{Predicted variance of } \hat{\omega})^2}{(\text{Predicted variance of } \tilde{\omega})},\end{aligned}$$

where $\tilde{\omega}$ is the weight vector of the estimated GMVP based on $\tilde{\Sigma}$

$$\Rightarrow \underbrace{\sigma^2(\hat{\omega}, \hat{\Sigma})}_{\text{Medium Var.}} \text{ is the geometric mean of } \underbrace{\sigma^2(\hat{\omega}, \Sigma)}_{\text{Large Var.}} \text{ and } \underbrace{\sigma^2(\tilde{\omega}, \tilde{\Sigma})}_{\text{Small Var.}}$$

ERROR QUANTIFICATION OF $\sigma^2(\tilde{\omega}, \tilde{\Sigma})$

- By the definition of $\epsilon \triangleq \hat{\Sigma}^{-1} - \Sigma^{-1}$, we have

$$\begin{aligned}\tilde{\Sigma}^{-1} &= \hat{\Sigma}^{-1} \Sigma \hat{\Sigma}^{-1} \\ &= (\Sigma^{-1} + \epsilon) \Sigma (\Sigma^{-1} + \epsilon) \\ &= \Sigma^{-1} + 2\epsilon + \epsilon \Sigma \epsilon \approx \Sigma^{-1} + 2\epsilon ,\end{aligned}$$

where the approximation is justified by small $\|\epsilon\|$

- It follows that

$$\underbrace{\sigma^2(\tilde{\omega}, \tilde{\Sigma})}_{\text{Small Var.}} = \frac{1}{\mathbf{1}^\top \tilde{\Sigma}^{-1} \mathbf{1}} \approx \frac{1}{\mathbf{1}^\top \Sigma^{-1} \mathbf{1} + 2 \cdot \underbrace{\mathbf{1}^\top \epsilon \mathbf{1}}_{\substack{\uparrow \text{ as } \|\epsilon\| \uparrow}}} } \leq \underbrace{\sigma^2(\hat{\omega}, \hat{\Sigma})}_{\text{Medium Var.}} ,$$

and the following ratio is (typically) monotone increasing in $\|\epsilon\|$:

$$\frac{\sigma^2(\hat{\omega}, \hat{\Sigma})}{\sigma^2(\tilde{\omega}, \tilde{\Sigma})} = \frac{(\text{Medium Var.})}{(\text{Small Var.})} \geq 1$$

- Define the volatility ratio (VR) as

$$\begin{aligned}(\text{VR}) &\triangleq \frac{\sigma(\hat{\omega}, \Sigma)}{\sigma(\hat{\omega}, \hat{\Sigma})} = \frac{\text{(Actual volatility of } \hat{\omega}\text{)}}{\text{(Predicted volatility of } \hat{\omega}\text{)}} \\ &= \sqrt{\frac{\text{(Large Var.)}}{\text{(Medium Var.)}}} = \sqrt{\frac{\text{(Medium Var.)}}{\text{(Small Var.)}}} = \frac{\sigma(\hat{\omega}, \hat{\Sigma})}{\sigma(\tilde{\omega}, \tilde{\Sigma})} \geq 1,\end{aligned}$$

which is monotone increasing as $\|\epsilon\|$ gets larger

TRUE FACTOR STRUCTURE IN (BASELINE) SIMULATION

- Our setup of the return generating process is empirically calibrated

▶ Return

Broad Factors ($K_0 = 4$)			Country Factors ($K_1 = 16$)		
	Volatility	No. of Stocks		Volatility	No. of Stocks
Market	16.00%	2000	Country 1	13.00%	154
Style 1	8.00%	2000	Country 2	12.38%	69
Style 2	4.00%	2000	Country 3	18.53%	159
Style 3	4.00%	2000	Country 4	15.91%	76
Industry Factors ($K_2 = 11$)			Country 5	24.93%	178
Industry 1	11.51%	143	Country 6	18.14%	99
Industry 2	12.74%	291	Country 7	11.86%	177
Industry 3	13.00%	131	Country 8	17.57%	19
Industry 4	15.08%	144	Country 9	16.43%	203
Industry 5	13.99%	115	Country 10	24.41%	73
Industry 6	17.80%	251	Country 11	14.31%	56
Industry 7	14.07%	258	Country 12	12.03%	247
Industry 8	12.22%	53	Country 13	13.33%	171
Industry 9	9.80%	109	Country 14	13.79%	75
Industry 10	13.11%	171	Country 15	18.29%	135
Industry 11	21.23%	334	Country 16	22.59%	109

EXTENDED SIMULATION: VARIABLE VOLATILITY FACTORS

■ Variable Volatility Factor Structure

▶ Return

- ▶ Factor volatility switches between normal and crisis states according to a Markov regime-switching (MRS) mechanism as a latent Markov chain $s_t \in \{0, 1\}$
- ▶ The transition matrix \mathbf{P} containing the probabilities

$$p_{ij} = \mathbb{P}(s_t = i | s_{t-1} = j)$$

of switching from regime j at time $t - 1$ to regime i at time t

$$\mathbf{P} = \begin{bmatrix} p_{0|0} & p_{1|0} \\ p_{0|1} & p_{1|1} \end{bmatrix} = \begin{bmatrix} 0.998 & 0.002 \\ 0.008 & 0.992 \end{bmatrix},$$

where the expected durations given by

$$D_0 = 1/(1 - p_{0|0}) = 500 \text{ (days)} = 2.0 \text{ (years)}$$

$$D_1 = 1/(1 - p_{1|1}) = 125 \text{ (days)} = 0.5 \text{ (years)}$$

- ▶ During the normal period (i.e., $s_t = 0$), the factor volatilities are drawn from the same distribution as the baseline simulation setup
- ▶ For the crisis regime (i.e., $s_t = 1$), we assume that the volatilities of the (market, style, country, industry, idiosyncratic) factors are multiplied by (2.0, 1.5, 1.5, 1.5, 1.25), respectively

EXTENDED SIMULATION: TIME-VARYING FACTOR LOADINGS

- We extend the simulation setup by allowing variable factor loadings over time
 - ▶ We assume that the *broad* factor exposures are time-varying based on the mean-reverting Ornstein–Uhlenbeck process

$$d\beta_t^i = \kappa \left(\bar{\beta}^i - \beta_t^i \right) dt + \sigma^i d\mathbf{Z}_t ,$$

where $\beta_t^i = [\beta_{0,t}^i, \beta_{1,t}^i, \beta_{2,t}^i, \beta_{3,t}^i]^\top$ is the time- t *broad* factor loadings specific to firm i , and \mathbf{Z}_t is a K_0 -dimensional standard Brownian motion

- ▶ For simplicity, we assume that the *narrow* factor loadings are binary constants
- ▶ We set the initial broad factor loadings and the long-run mean level of the factor loadings as the true broad factor exposures in the baseline simulation setup

- Responsive Covariance Adjustments (RCA) [▶ Return](#)
 - ▶ As $N, T \rightarrow \infty$, the LH-PCA scheme consistently estimates the asymptotic covariance matrix, which is given by $\lim_{N, T \rightarrow \infty} \Sigma_N(T)$
 - ▶ This estimate is distinct from the covariance matrix that will be encountered on the subsequent day of estimation, when dealing with dynamic factor models
 - ▶ This disparity has the potential to lead to empirically inaccurate out-of-sample predictions of the true covariance matrix on a daily basis
- We institute a **Responsive Covariance Adjustment** (RCA) estimating current covariance matrix
 - ▶ Exponentially weighted moving average with 40-day half-life
 - ▶ Applied separately to the factor returns, on the one hand, and the idiosyncratic factor returns, on the other hand, to correct for changing factor volatilities

RESPONSIVE COVARIANCE ADJUSTMENT (RCA)

- In real-world financial markets, volatility is constantly changing with clustering property ▶ Return
 - ▶ Standard practice is to address variable factor volatility based on the **Exponentially Weighted Moving Averages** (EWMA) model using a Responsive Covariance Adjustment (RCA)
- Practitioners often assign more weight to recently realized returns than distant observations to capture the contemporaneous volatility structure
 - ▶ Within each trailing window of T days, a decay factor $\eta \in (0, 1)$ yields the adjusted factor-based return covariance
 - ▶ Simply put, the daily forecast of factor-based return covariance matrix estimated on day t is given by

$$\widehat{\Sigma}_{N \times N}(t+1) = \sum_{m=0}^{T-1} \omega_m \xi_{t-m}^T \xi_{t-m},$$

where the weights satisfy $\sum_{m=1}^T w_m = 1$ and decrease by fixed proportion as $\omega_{m+1} = \eta \omega_m$

- ▶ A half-life of T_S trading days corresponds to setting $\eta = \exp\left(\frac{\log 0.5}{T_S}\right)$ as the decay factor.
- ▶ In our study, we set $T_S = 40$ days so that the decay factor is $\eta \approx 0.9828$

■ Determining the Number of PCA Eigenfactors

[▶ Return](#)

- ▶ Existing literature highlights that the consistent principal component estimation of weak factors and their associated loadings depends on the degree of their weakness; refer to, for example, Freyaldenhoven (2022), Uematsu & Yamagata (2023), and Bai & Ng (2023)
- ▶ Corollary 2 in Uematsu & Yamagata (2023) points out that the Bai & Ng (2002) estimators remain valid for determining the number of relevant (eigen)factors under more general (yet reasonably mild) conditions
- ▶ ..., where principal component estimates of the factors are consistent with potentially variable weak factor loadings

DEFINITIONS OF BS / MRAD / Q-STATISTICS

- Z-score:
$$z_t = \frac{(\text{Realized GMVP return})_{t+1}}{(\text{Predicted GMVP volatility})_t} = \frac{R_{t+1} \hat{\omega}_t}{\sigma(\hat{\omega}_t, \hat{\Sigma}_t)}$$
 - ▶ z_t is a realized GMVP return after standardization with its *predicted* volatility

- Bias Statistics:
$$\text{BS}_t(\tau) = \sqrt{\frac{1}{\tau} \sum_{k=t-\tau+1}^t z_k^2}$$
 - ▶ τ is the number of (moving) windows in the testing period
 - ▶ If the forecasts are accurate, the realized BS should be close to one

- Mean Rolling Absolute Deviation:

$$\text{MRAD}_t(u; \tau) = \frac{1}{\tau - u + 1} \sum_{k=t-\tau+u}^t |\text{BS}_k(u) - 1|$$

- ▶ The rolling window moves forward one day at a time to the end of the entire sample period
 - ▶ We set the block size $u = 12$ trading days following conventional practice
- Q-statistics: $z_t^2 - \log z_t^2$ (Patton, 2011) [▶ Return](#)

EMPIRICAL ANALYSIS: DATA CLEANING PROCEDURES

- CRSP: following the standard finance literature, we use stocks that are
 - ▶ identified as common stocks (share code 10, 11)
 - ▶ listed on AMEX, NASDAQ, and NYSE (exchange code 1, 2, and 3)
 - ▶ not investment funds, trusts, REITS (exclude SIC code 6722, 6726, 6798, 6799)
 - ▶ Micro stocks (maximum closing price < \$5) are dropped from our sample
- EURO: in the similar context, we incorporate stocks that are
 - ▶ identified as common, ordinary stocks (tpci code 0)
 - ▶ not investment funds, trusts, REITS (exclude SIC code 6722, 6726, 6798, 6799)
 - ▶ holidays are removed
 - ▶ Illiquid stocks (the percentage of the imputed closing prices > 50% of the total sample for each security) are dropped from our sample
- We include all of the tradable stocks in the market
 - ▶ When a stock does not trade on a given day, we use the average of the closing Bid and Ask as the closing price for CRSP dataset
 - ▶ Since closing Bid and Asks are not available for EURO dataset, we interpolate the closing price linearly between the last previous and next future closing price