Low-Sample Factor Modelling for Asset Pricing

Sumanjay Dutta [1] Shashi Jain [1]

[1] Indian Insitute of Science, Bengaluru

February 13, 2024

Factor Models: Motivation

- Factor analysis explains returns with a small number of fundamental variables called factors or risk factors.
- Factor Models in Finance can be classified as: a) Fundamental b) Statistical and c) Macroeconomic Factor Models.
- We want to explain the returns of different stocks in terms of fewer variables (known as factors)..
- Instead of using the past history of return, we use other variables (factors) as the regressors.
- Statistical factor models estimate *latent* factors based on the cross-section of returns and use them as regressors for explaining individual stock returns.

Uses of Factor Models

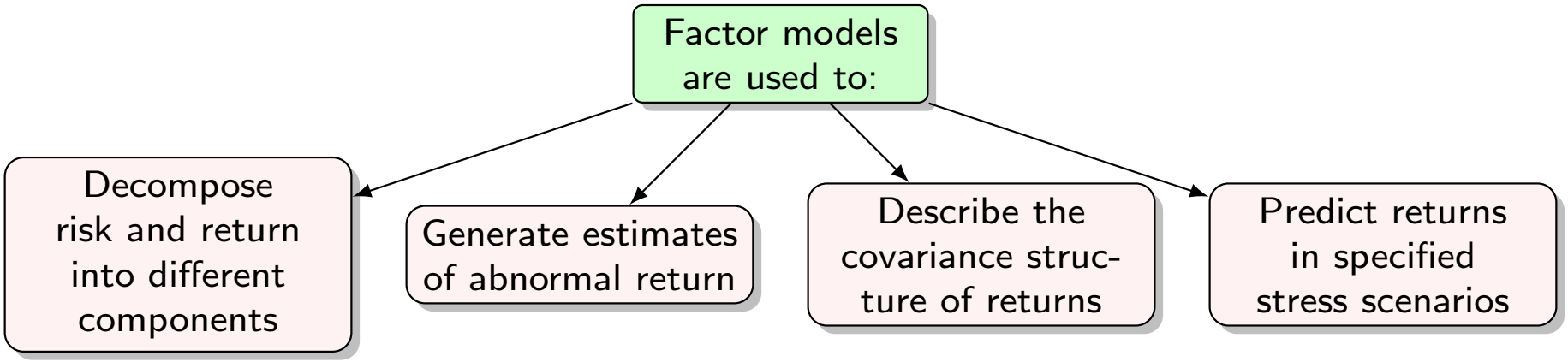


Figure 1: Factors models have generally been used for multiple purposes in financial modelling. Some of the uses have been listed above. Other uses include portfolio allocation, etc.

Factor Models

- Suppose there are p stocks, and n periods. Let r_{it} be the excess return [1] of asset i at time t. Throughout the slides we assume r_{it} is stationary.
- A general form of the factor model is

$$r_{it} = \beta_{0i} + \beta_{1i} f_{1t} + \beta_{2i} f_{2t} + \ldots + \beta_{ki} f_{kt} + \epsilon_{it} \tag{1}$$

  where we assume there are k factors, and f_{jt} is the j-th factor at time t.
- β's represent the factor loadings, ϵ's represent the asset specific factor.
- The $p \times p$ covariance matrix of asset returns has the form

$$\Sigma_{R_t} = \Sigma_{FM} = B \Sigma_f B' + D \tag{2}$$

  where B is $p \times k$ matrix, F is a $k \times k$ matrix and D is $p \times p$ matrix. B is also called factor loadings matrix.

[1] excess return = actual return - risk free-rate

Low-Samples and High Dimensions - Relevance

- Most methods developed to test and estimate asset pricing models are designed for cases when the sample size n is large compared to the number of assets p.
- However, there are many instances when allowing a large n is not beneficial or, sometimes, just not feasible. Eg. structural breaks.
- Moreover, sometimes a long time series of data simply is not available. Eg. emerging markets or new financial instruments.
- Asset pricing theory often requires the number of assets to be arbitrarily large.

Statistical Factor Modelling using PCA

- In a statistical factor model, the factors are the principle components (PC) of the return series.
- Consider a linear combination (portfolio) of p returns at the t-th period

$$\sum_{i=1}^{p} w_i r_{it}$$

  where the weight for the i-th asset is w_i.
- PC is a special linear combination so that
  - each PC is uncorrelated with each other
  - each PC will explain the maximum amount of remaining variance-covariance of r_t after the previous PC.

Sample Principal Components and Estimated Factors

- Sample principal components are computed from the spectral decompositon of the $p \times p$ sample covariance matrix $\hat{\Sigma}$ when $p < n$:

$$\hat{\Sigma} = \hat{P}\hat{\tau}\hat{P}'$$

$$\hat{P} = [\hat{p}_1^*, \hat{p}_2^*, \ldots, \hat{p}_p^*]$$

$$\hat{P}' = \hat{P}^{-1}$$

$$\hat{\tau} = diag(\hat{\tau_1}, \hat{\tau_2}, \ldots, \hat{\tau_N})$$

- The estimated factor realizations are simply the first K sample principal components

$$\hat{f_{kt}} = \hat{p}_k^{*\prime} R_t$$

where $k = 1, \ldots, k$.

$$\hat{f_t} = (\hat{f_{1t}}, \ldots, \hat{f_{Kt}})'$$

Sample Principal Components and Estimated Factors

- The factor loadings for each asset, β_i, and the residual variances, $var(\epsilon_{it}) = \sigma_i^2$ can be estimated via OLS from the time series regression

$$r_{it} = \alpha_i + \beta_i' \hat{f}_t + \epsilon_{it}$$

giving $\hat{\beta}_i$ and σ_i^2 for $i = 1, \ldots, n$.

- The factor model covariance matrix of returns is then

$$\hat{\Sigma}_{FM} = \hat{B}\hat{\Sigma}_f \hat{B}' + \hat{D} \tag{3}$$

Where $\hat{\Sigma}_f = \frac{1}{T-1} \sum_{t=1}^{T} (\hat{f}_t - \bar{f})(\hat{f}_t - \bar{f})'$ and $\bar{f} = \frac{1}{T} \sum_{t=1}^{T} \hat{f}_t$.

Limitations in Low-Samples

- PCA cannot be performed in a low-sample regime since eigenvalue decomposition fails.
- As n goes closer to p, the eigenvalues of the sample covariance matrix becomes noisy.
- The covariance matrix too becomes ill-conditioned, i.e. more sensitive to noise.
- If $p > n$, then the sample covariance matrix of returns becomes singular which complicates traditional factor and principal components analysis.
- Therefore PCA requires a different approach for covariance estimation.

Possible Solution

- The problem with PCA in low-sample settings can be solved by either getting cleaner estimates of: a) covariance matrix and b) inverse covariance matrix.
- The covariance matrix can be estimated by a) Shrinkage Methods which are further classified into linear and nonlinear shrinkage estimators. This approach has been discussed in Ledoit and Wolf (2015).
- Inverse Covariance Matrix (also called Precision Matrix) is estimated using Gaussian Graphical Models (GGMs), which shall be discussed in the slides that follow.

Problems in Existing Literature - Solution through GGM-based PCAs

1. Ledoit et al. (2015) provide a framework for incorporating linear and non-linear shrinkage estimators of the covariance matrix in the PCA framework. Yet, current literature on multivariate statistics does not address the following questions:
2. Is is possible to derive alternatives to the APCA for low-sample settings?
3. If yes, then is it possible to price stocks in high-dimension low-sample settings using statistical factors computed through PCA?
4. While Ledoit et al. (2015) propose some synthetic tests for comparing covariance estimators in terms of eigenvalue properties, they do not extend their work to more practical settings. The question that follows is, is it possible to suggest tests that would be helpful to compare PCA based factor models in empirical settings?
5. To elaborate the previous point, there are two questions that need to be addressed here.
    1. What are the loss metrics that would be considered for comparing PCA-based factor models.
    2. Given that loss-metrics are identified, what methodology would be used to rank different methods based on the suggested metrics ?

Contributions

- This paper addresses three crucial challenges within statistical factor modelling.
- Comparison of the eigenvalue properties of the covariance matrix and precision matrix, which is essential for PCA.
- Conducting principal component analysis (PCA) based on different covariance and precision matrix estimators which serve as an alternative to APCA. (henceforth known as *low-sample PCA methods*.
- Performing statistical factor modelling based on low-sample PCA methods.
- Testing them based on loss-functions which have a financial interpretation.
- Proposing synthetic and empirical experiments for comparing different factor models.

Gaussian Graphical Models

- A GGM in N dimensions is a probability distribution with density

$$p(X = x) = \frac{1}{\sqrt{(2\pi)^N |\Sigma|}} \exp\left(-(x-\mu)^T \Sigma^{-1} (x-\mu)/2\right)$$

  where μ is the mean and Σ is the covariance matrix. The important point is that their conditional independence structure is encoded by $\Theta = \Sigma^{-1}$.
- A zero off-diagonal entry $\Theta_{j,k} = 0$ implies X_j and X_k are conditionally independent given all other variables.

$$Q = \begin{pmatrix} * & * & * & * & * & 0 \\ * & * & * & * & * & 0 \\ * & * & * & 0 & 0 & 0 \\ * & * & 0 & * & 0 & 0 \\ * & * & 0 & 0 & * & 0 \\ 0 & 0 & 0 & 0 & 0 & * \end{pmatrix}$$

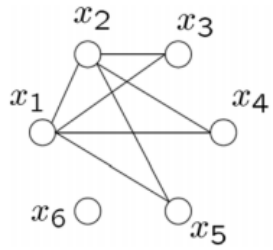


Figure 2: Connection between the graph and the precision matrix

Inverse Covariance Matrix Estimation - Graphical Lasso

- For the entire data X, the likelihood function is $L(\Theta) = f(x)^p$. Taking logarithm and after some algebra

$$\arg\max_{\Theta} l(\Theta) = \log det(\Theta) - tr(S\Theta) - \lambda||\Theta||_1 \tag{4}$$

- It is reasonable to impose structure on $\Theta(\Sigma)$ or assume that they are sparse. That is some of $\Theta_{i,j} = 0$.
- The entries of Θ_{ij} have regression interpretation.
- In particular, Θ_{ij} is proportional to the regression coefficient of variable X_j in the multiple regression of variable X_i on the rest.
- The zeros in coefficient can be forced by a column-by column approach through penalized least square (lasso).

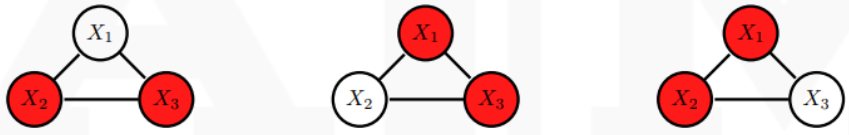


Figure 3: Nodewise Regression

PCA through Inverse Covariance Estimation

- The eigenvalues of Θ are the reciprocal of the eigenvalues of Σ. The eigenvectors remain the same.
- We first estimate $\hat{\Theta}$ for three GGM-based methods (Graphical Lasso, GreedyPrune and HybridMB. Then we perform eigenvalue decomposition of $\hat{\Theta}$.
- Rest of the process is similar to standard PCA.
- Benchmarking of models was done using both synthetic and empirical experiments.
- We compare our approach with Shrinkage and Asymptotic PCA proposed by Conner and Korajczyk (1986).
- We use the linear and non-linear shrinkage methods proposed by Ledoit Wolf (2004,2020).
- Synthetic experiments were performed using a multivariate gaussian data generation process.
- Empirical experiments were performed on daily data for 500 stocks listed at the NSE exchange.

Methods used in the Study

Method	Class	Reference
GPPCA	GGM	Kelner et. al (2020)
GLPCA	GGM	Friedman et. al (2007)
HMBPCA	GGM	Kelner et. al (2020)
LWNLPCA	Nonlinear Shrinkage	Ledoit and Wolf (2020)
LWLPCA	Linear Shrinkage	Ledoit and Wolf (2004)

Table 1: Methods used in the paper. We have used three GGM-based methods which estimate the inverse covariance matrix directly and then use it in PCA. Two shrinkage methods have also been considered.

Choice of Loss Functions - Empirical Experiment

- We use three loss functions:
- **Asset Mispricing** - RMS-Alpha measures to what extent the average returns are not attributable to the extracted factors.

$$\text{RMS Alpha} = \sqrt{\frac{1}{p}\sum_{i=1}^{p}\hat{\alpha}_i^2}$$

where p represents the number of stocks.

- **Explanatory Power** - The Total Adjusted R^2 measures fraction of return variance explained by the estimated models:

$$Total\ adj - R^2 = 1 - \frac{\sum_{it}(R_{it} - \hat{\alpha}_i - \hat{\beta_{i1}}\hat{f_{1t}} + \ldots + \hat{\beta_{i1}}\hat{f_{1t}})^2}{\sum_{it} R_{it}^2}$$

- **Correlation Structure** - Sphericity measures the correlation among the idiosyncratic errors of the stocks under consideration.[2]

$$Sphericity = \det(\Omega)$$

where Ω represents the correlation matrix.

[2] Based on Bartlett's Sphericity. The closer the correlation matrix of errors is to the identity matrix, the lesser the chance of mispricing. See Ruppert (2011) for proof.

Synthetic Experiments

- We use the daily closing prices of the constituent stocks of the Nifty500 index [3] The duration of study is 3rd January 2017 to November 16th 2023 (1871 days).
- We then estimate the sample covariance matrix ($\hat{\Sigma}$) and sample mean returns ($\hat{\mu}$) from the daily returns. These estimates are then used as the *true* parameters (μ and Σ) for the multivariate Gaussian distribution to generate the samples.
- The dimension size is 500 and we keep varying the sample size from 250 till 1000.
- The estimation accuracy of different covariance matrix estimation methods is then measured by three methods, viz, the Eigenvalue Distance (ED), Factor Coefficient Distance (FCD) Factor Matrix Distance (FD).
- We perform a similar experiment with lower dimension size (Nifty top 100 companies by Market Capitalization).

[3] These are constituents of Nifty500 as on 1st April 2021.

Synthetic Experiments

Design

Relative to the true covariance structure, the SE and FD are given by:

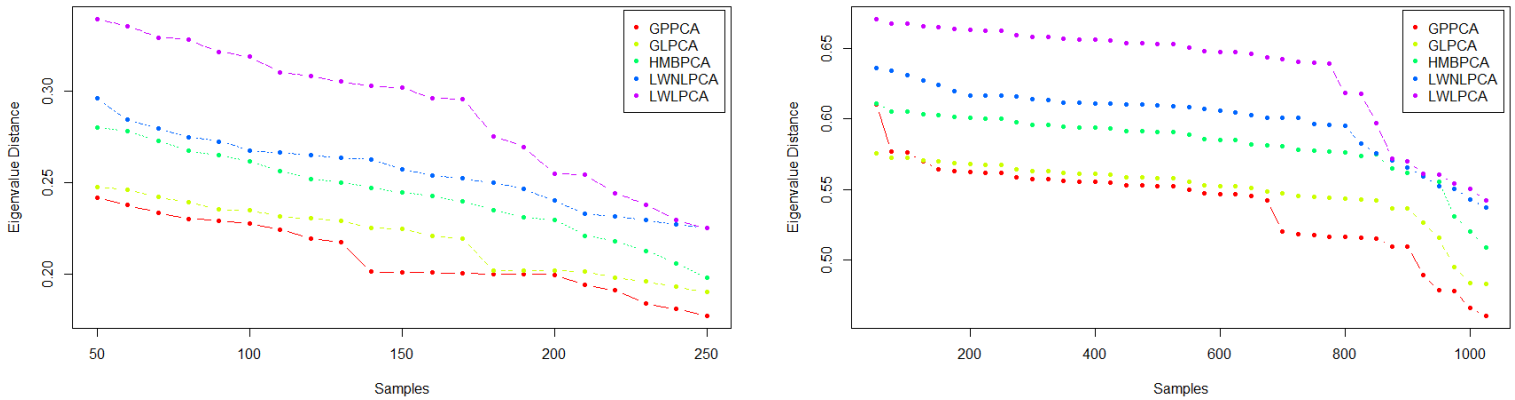
$$\text{ED} = \frac{1}{p}\sum_{i=1}^{p}[\hat{\tau}_i - \tau_i]^2$$

$$\text{FCD} = ||\tilde{FC} - FC||^2$$

$$\text{FMD} = ||\tilde{F} - F||^2$$

$\hat{\tau}_i$ and τ_i represent the eigenvalues computed from the proposed PCAs and the the *true* eigenvalues respectively. Here, $\tilde{\Sigma}$ represents the estimated factor matrix obtained from shrinkage or GGM-based method. $\tilde{FC}$ represents the estimated factor coefficient matrix obtained from shrinkage or GGM-based method. F represents the true matrix for the factors. FMD measures the distance of the estimated factors from the factors estimated using the true covariance matrix.

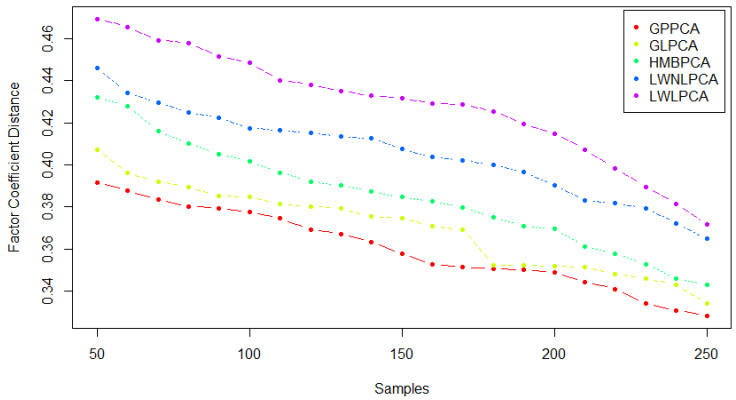
Results - Synthetic Experiments - Eigenvalue Distance



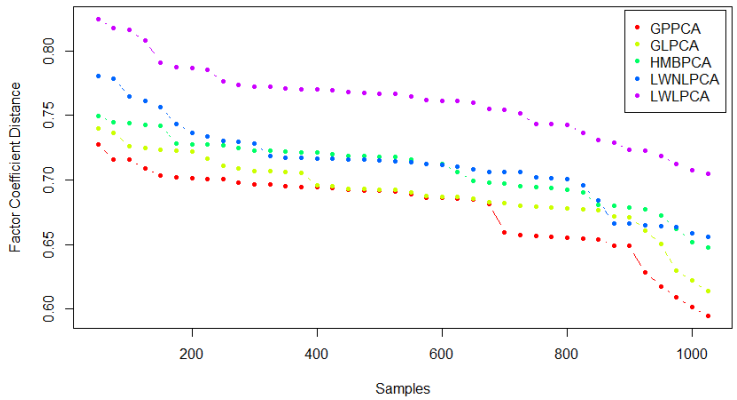
(a) Eigenvalue Distance for Dimension Size 100 (b) Eigenvalue Distance for Dimension Size 500

Figure 4: Synthetic Experiment: Eigenvalue Distance for different dimensions

Results - Synthetic Experiments - Factor Coefficient Distance



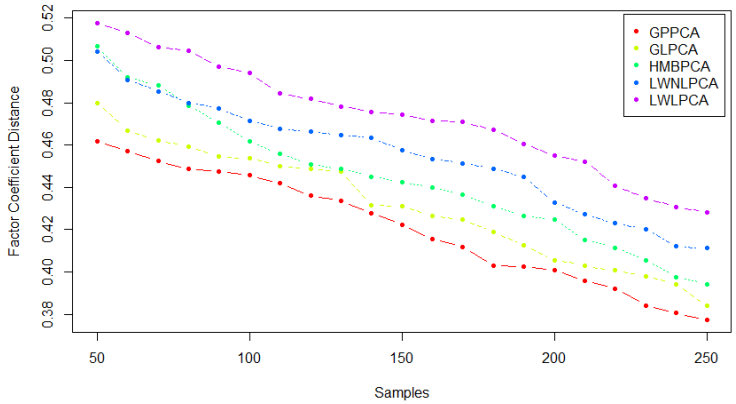
(a) FCD for Dimension Size 100



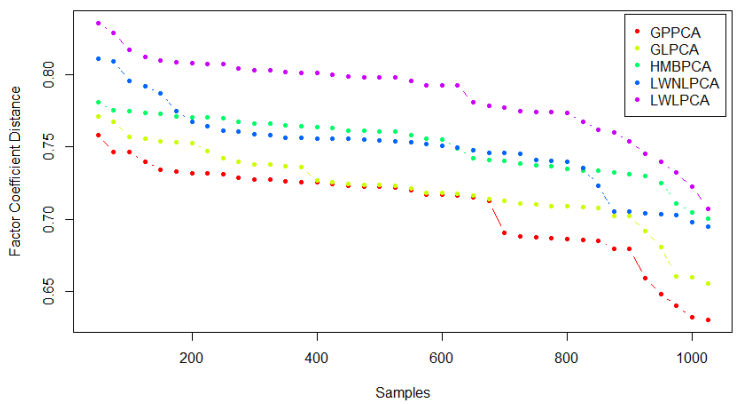
(b) FCD for Dimension Size 500

Figure 5: Synthetic Experiment: FCD for different dimensions. Three-factor case.

Results - Synthetic Experiments - Factor Coefficient Distance



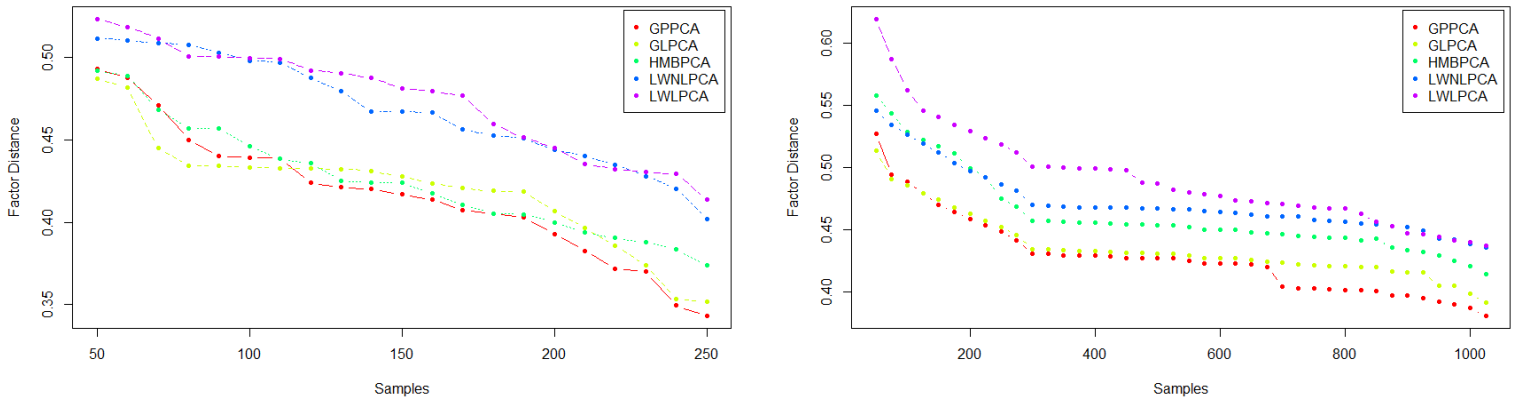
(a) FCD for Dimension Size 100



(b) FCD for Dimension Size 500

Figure 6: Synthetic Experiment: FCD for different dimensions. Four-factor case.

Results - Synthetic Experiments - Factor Matrix Distance

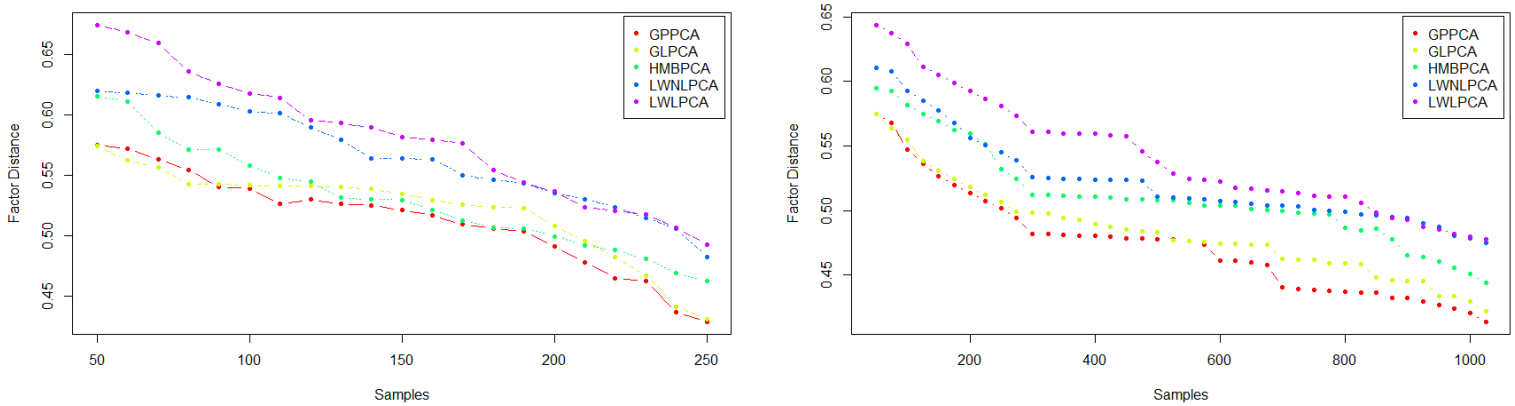


(a) FMD for Dimension Size 100

(b) FMD for Dimension Size 500

Figure 7: Synthetic Experiment: FMD for different dimensions. Three-factor case.

Results - Synthetic Experiments - Factor Matrix Distance



(a) FMD for Dimension Size 100

(b) FMD for Dimension Size 500

Figure 8: Synthetic Experiment: FMD for different dimensions. Four-factor case.

Findings based on Synthetic Experiments

- Eigenvalues calculated from GGM-based PCAs are closer to the *true* eigenvalues than linear and non-linear shrinkage based PCAs for both low and high sample settings.
- Asymptotic PCA performs worse than all the proposed alternatives but starts improving as the number of samples increases.
- Distance between true and estimated factors is explained better by GGM-based PCAs, followed by Shrinkage Methods. Here too, APCA starts improving as sample size increases.
- Our observations remain consistent for both dimension sizes.
- We use a three-factor model in this case.[4]

[4] We use the Cattell-Nelson-Gorsuch (CNG) test to find the optimal number of factors. Based on our dataset, we find that the optimal number of factors comes out as three.

Empirical Experiments

- We use daily returns data of the constituents of the Nifty500 index and perform our proposed PCAs on it. on it.
- Using in-sample factor estimates, we compute Out-of-Sample RMS-Alpha, Total Adjusted R^2 and Sphericity based on next day's intra-day data.
- We get a time series of Out-of-Sample performance metrics.
- With a large number of competing models, it becomes difficult to conclusively rank order their performance.
- The test for Superior Predictive Ability (SPA) identifies benchmark model.
- But the SPA test cannot rank models.
- We use Hansen's Model Confidence Set (MCS) Approach for ranking competing models.

Results for Out-of-Sample RMS-Alpha

Nifty500				Nifty100			
Method	Rank_R	v_R	MCS_R	Method	Rank_R	v_R	MCS_R
GPPCA	1	-0.52	1	GPPCA	1	-1.35	1
GLPCA	2	-0.29	1	GLPCA	2	-1.09	0.99
HMBPCA	3	0.09	0.99	HMBPCA	3	-0.81	0.97
LWNLPCA	4	0.12	0.98	LWNLPCA	4	0.05	0.95
LWLPCA	5	0.25	0.98	LWLPCA	5	0.21	0.95
APCA	6	0.37	0.96	APCA	6	1.05	0.90

Table 2: Out-of-Sample MCS test for RMS-Alpha for three-factor model. The above table represents the superior set of models (SSM) determined using the MCS method. The subscript R refers to the elimination rules associated with a test statistic defined in Hansen (2011). This test statistic is a contrast of model i's sample loss to that of the average across all models. MCS_R represents the level of belongingness in the SSM. v_R represents the loss relative to other models.The lower v_R, the better the model.

Results for Out-of-Sample Adjusted R^2

Nifty500				Nifty100			
Method	Rank_R	v_R	MCS_R	Method	Rank_R	v_R	MCS_R
GPPCA	1	-1.91	1	GPPCA	1	-1.42	1
GLPCA	2	-1.73	0.99	GLPCA	2	-1.03	0.99
LWNLPCA	3	-1.25	0.99	HMBPCA	3	-0.55	0.99
HMBPCA	4	-0.51	0.98	LWLPCA	4	-0.11	0.98
LWLPCA	5	-0.29	0.95	LWNLPCA	5	0.097	0.95
APCA	6	0.05	0.92	APCA	6	0.35	0.97

Table 3: Out-of-Sample MCS test for Adjusted R^2 for three-factor model.

Results for Out-of-Sample Sphericity

Nifty500				Nifty100			
Method	Rank_R	v_R	MCS_R	Method	Rank_R	v_R	MCS_R
GLPCA	1	-1.99	1	GLPCA	1	-1.38	1
GPPCA	2	-1.51	0.99	HMBPCA	2	-1.05	0.99
LWNLPCA	3	-0.87	0.99	GPPCA	3	-0.43	0.99
LWLPCA	4	-0.25	0.98	LWLPCA	4	0.01	0.98
LWLPCA	5	-0.02	0.96	LWNLPCA	5	0.25	0.97
APCA	6	0.19	0.94	APCA	6	0.45	0.95

Table 4: Out-of-Sample MCS test for Sphericity for three-factor model.

Findings from Empirical Experiments

- In terms of RMS-Alpha, Total Adjusted R^2 and Sphericity, GLPCA and GPPCA perform better than LWNLPCA and LWLPCA.
- GPPCA performs better GLPCA in terms of RMS-Alpha and Total Adjusted R^2. But GLPCA performs better than GPPCA.
- Compared to APCA, which was proposed for low-sample settings, GGM and shrinkage based methods perform better for all metrics.
- Effect on Asset Pricing: GGMs and shrinkage based methods lead to better pricing of individual stocks than APCA.
- The performance of different methods remains similar across different dimensions.
- When extended to four and five-factor cases, the ranking of algorithms remains similar.

References

- Olivier Ledoit and Michael Wolf. *A well-conditioned estimator for large-dimensional covariance matrices*. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.
- Ledoit and Michael Wolf. Spectrum estimation: A unified framework for covariance matrix estimation and pca in large dimensions. *Journal of Multivariate Analysis*, 139:360–384, 2015.
- Olivier Ledoit and Michael Wolf. *Analytical nonlinear shrinkage of large-dimensional covariance matrices*. 2020.
- David Ruppert and David S Matteson. Statistics and data analysis for financial engineering, volume 13. Springer, 2011.
- Jonathan Kelner, Frederic Koehler, Raghu Meka, and Ankur Moitra. Learning some popular gaussian graphical models without condition number bounds. *Advances in Neural Information Processing Systems*, 33:10986–10998, 2020.
- Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008