Factor-Augmented Forecasting in Big Data

Juhee Bae
University of Glasgow

NBER-NSF Time Series Conference
October, 2021

Motivation and Objectives

- Factor-augmented forecasts
  1. Estimate the latent factors from data.
  2. Augment the estimated factors to forecasting equations.

- Estimated factors determine forecasting accuracy.

- Objectives:
  1. Evaluate comprehensively forecasting power of many factor estimation techniques, under the same forecasting framework.
  2. How estimated factors affect forecasting power?
  3. Which factor-augmented forecasting method tends to give the best results?

Model

Factor Models

- Factors are a few latent reference variables that drive comovements of data.

$$\begin{aligned} y_{t+1} &= \underset{1\times k}{\beta'} \underset{k\times 1}{F_t} + u_{t+1} \\ \underset{N\times 1}{x_t} &= \underset{N\times k}{\Phi} \underset{k\times 1}{F_t} + \underset{N\times 1}{e_t} \end{aligned} \quad \Rightarrow \quad \begin{aligned} \underset{T\times 1}{y} &= \underset{T\times k}{F} \underset{k\times 1}{\beta} + \underset{T\times 1}{u} \\ \underset{T\times N}{X} &= \underset{T\times k}{F} \underset{k\times N}{\Phi'} + \underset{T\times N}{E} \end{aligned}$$

y_{t+1}: a target variable being forecasted,
x_t: N predictors
F_t: k latent common factors in predictors.

Factor Models

- Factors are a few latent reference variables that drive comovements of data.

$$y_{t+1} = \boxed{\underbrace{\underset{1\times m}{\delta'} \; \underset{m\times 1}{G_t}}_{\beta' F_t}} + u_{t+1} \qquad \Rightarrow \qquad \underset{T\times 1}{y} = \underbrace{\underset{T\times m}{G} \; \underset{m\times 1}{\delta}}_{F\beta} + \underset{T\times 1}{u}$$

$$\underset{N\times 1}{x_t} = \underset{N\times k}{\Phi} \; \underset{k\times 1}{F_t} + \underset{N\times 1}{e_t} \qquad \qquad \underset{T\times N}{X} = \underset{T\times k}{F} \; \underset{k\times N}{\Phi'} + \underset{T\times N}{E}$$

y_{t+1}: a target variable being forecasted,
x_t: N predictors
F_t: k latent common factors in predictors.

- The best forecast for y_{t+1} at t is $\beta' F_t$. Let G_t be m factors such that $\boxed{\delta' G_t = \beta' F_t}$.

Factor Models

- Factors are a few latent reference variables that drive comovements of data.

$$\underset{}{y_{t+1}} = \boxed{\underbrace{\underset{1\times m}{\delta'}\ \underset{m\times 1}{G_t}}_{\beta' F_t}} + u_{t+1} \qquad \Rightarrow \qquad \underset{T\times 1}{y} = \underbrace{\underset{T\times m}{G}\ \underset{m\times 1}{\delta}}_{F\beta} + \underset{T\times 1}{u}$$

$$\underset{N\times 1}{x_t} = \underset{N\times k}{\Phi}\ \underset{k\times 1}{F_t} + \underset{N\times 1}{e_t} \qquad\qquad \underset{T\times N}{X} = \underset{T\times k}{F}\ \underset{k\times N}{\Phi'} + \underset{T\times N}{E}$$

- The best forecast for y_{t+1} at t is $\beta' F_t$. Let G_t be m factors such that $\boxed{\delta' G_t = \beta' F_t}$.
- Why F_t and G_t should be distinguished?
  - Our goal is **NOT** recovering F_t. What we need is a **specific linear combination of F_t**, which is $\boldsymbol{\beta' F_t}$.
  - **Some factor estimations** recover $\boldsymbol{HF_t}$ with some H. We recover $\beta' H^{-1} H F_t = \beta' F_t$. Then $\boldsymbol{G_t = HF_t}$, $\delta' = \beta' H^{-1}$ and $\boldsymbol{m = k}$.
  - But **some factor estimations** recover $\beta' F_t$ with G_t such that $\boldsymbol{m \leq k}$.

Factor Models

- Factors are a few latent reference variables that drive comovements of data.

$$y_{t+1} = \boxed{\underbrace{\underset{1\times m}{\delta'} \; \underset{m\times 1}{G_t}}_{\beta' F_t}} + u_{t+1} \qquad \Rightarrow \qquad \underset{T\times 1}{y} = \underbrace{\underset{T\times m}{G} \; \underset{m\times 1}{\delta}}_{F\beta} + \underset{T\times 1}{u}$$

$$\underset{N\times 1}{x_t} = \underset{N\times k}{\Phi} \; \underset{k\times 1}{F_t} + \underset{N\times 1}{e_t} \qquad \underset{T\times N}{X} = \underset{T\times k}{F} \; \underset{k\times N}{\Phi'} + \underset{T\times N}{E}$$

- The best forecast for y_{t+1} at t is $\beta' F_t$. Let G_t be m factors such that $\boxed{\delta' G_t = \beta' F_t}$.

- Why F_t and G_t should be distinguished?
    - Factor estimation methods estimate G_t, which may not be identical to F_t.
    - ***k*:** The number of the original factors. (dimension of F_t)
    - ***m*:** The optimal number of estimated factors by the given factor estimation method. (dimension of G_t)

Estimation of Factors

1. Principal Component Analysis (PCA)

- $\hat{\lambda}_1, \ldots, \hat{\lambda}_K$ are the k largest eigenvalues of $S_{NT} = X'X/(NT)$.
- $\hat{\alpha}_j = \sqrt{N} \times$ eigenvector corresponding to $\hat{\lambda}_j$.
- Let $\hat{A} = [\hat{\alpha}_1, \ldots, \hat{\alpha}_k]$, where $\hat{A}$ is $N \times k$ matrix.
- Then PCA factors $\hat{G}_{PCA}$ are estimated as $\hat{G}_{PCA} \equiv X\hat{A}/N$
- k is the asymptotically optimal number of PCA factors. It can be estimated by many different methods (called Decision Rules).
- PCA factors are estimated without using the information on the relationship between X and y. (unsupervised method)

2. Partial Least Squares (PLS) Factors: Ahn and Bae (2020)

- Ahn and Bae (2020) provide asymptotic properties of the PLS factors:
  - Note that $y = \boxed{F\beta} + u$ and we should recover $F\beta$.
  - Let $\boxed{F\beta = \Sigma_{j=1}^{J} F_{(j)}\beta_{(j)}}$, where the factors in $F_{(j)}$ have the same variances σ_j^2 ($\sigma_j^2 \neq \sigma_{j'}^2$), J: number of distinct factor variances
  - $\tilde{\alpha}_j = S_{NT}^{j-1} \times X'y/(N^{1/2}T)$, where $S_{NT} = X'X/(NT)$ and $\tilde{A} = [\tilde{\alpha}_1, ..., \tilde{\alpha}_J]$.
  - $\underset{T\times J}{\tilde{G}_{PLS}} \equiv X\tilde{A} = \boxed{\underset{T\times J}{(F_{(1)}\beta_{(1)}, ..., F_{(J)}\beta_{(J)})}} \underset{J\times J}{\tilde{B}} + \underset{negligible}{E\tilde{A}}$
  - J (number of distinct factor variances) is the asymptotically optimal number of PLS factors to use.
  - Simulation results: Forecasting with $X\tilde{\alpha}_1$ **(PLS1)** very often outperforms forecasting with more PLS factors. This is so even if $J > 1$.

Experiment Design

Forecasting Models

- Denote 12-month-ahead variable as y_{t+12}. Using t dated predictors, we will forecast $\hat{y}_{t+12|t}$, using all the available information up to t.

- The forecast for y_{t+12} is

$$\hat{y}_{t+12} = \hat{\theta} + \underbrace{\hat{\delta}'\hat{G}_t}_{\hat{\beta}'\hat{F}_t} + \sum_{j=1}^{p} \hat{\gamma}_j' y_{t-j+1}$$

  where $\hat{G}_t$ is the estimated factors.

- Estimate these unobservable G_t out of X using 7 factor estimation methods.

- But we should decide
  1. m: Decision rules for number of estimated factors
  2. p: Bayesian Information Criteria

Results

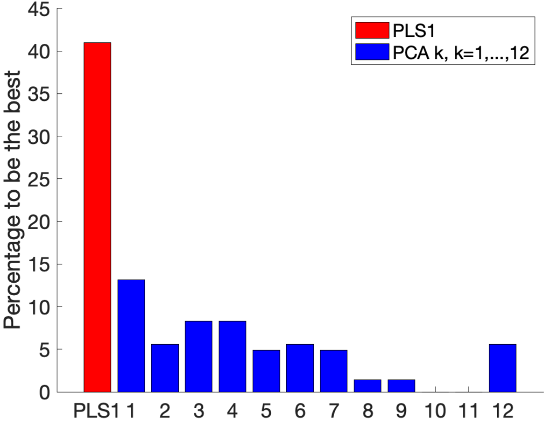
Which method tends to give the best results?

Strong Forecasting Performance of PLS1

- There are 101 combinations of all factor estimation methods and decision rules, for a target variable.
- There are 148 target variables.
- PLS1 is the best out of 101 combinations for 23 variables (around 16%), most frequently.
- **The main takeaway: PLS1 gives forecasting performance very close to the best result out of 101 combinations.**

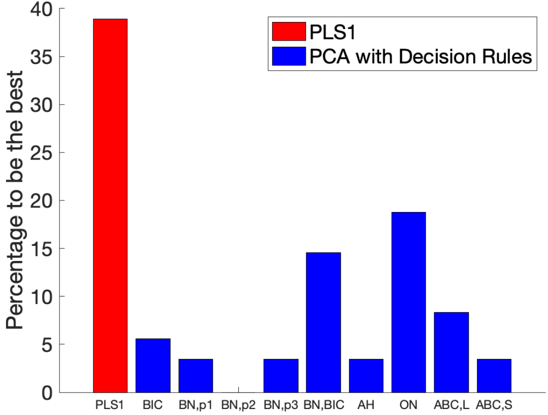
PLS1 vs PCA *k*, All Variables

- I compared forecasting results of **PLS1** and **PCA *k*, with *k* = 1, 2, ..., 12**, for all target variables.
- For each variable, I found the best method out of 13 methods.
- The following is the percentage that each method was the best.



PLS1 vs PCA with Decision Rules, All Variables

- I compared forecasting results of **PLS1** and **PCA with all possible decision rules**, for all the target variables.



BN: Bai and Ng (2002), AH: Ahn and Horenstein (2013),
ON: Onatski (2010), ABC: Alessi et al (2010)

References

[1] Ahn, S. C., & Bae, J. (2020). Forecasting with Partial Least Squares When a Large Number of Predictors Are Available. Working paper. https://www.juheebae.com

[2] Ahn, S. C., & Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica*, 81(3), 1203-1227.

[3] Alessi, L., Barigozzi, M., & Capasso, M. (2010). Improved penalization for determining the number of factors in approximate factor models. *Statistics & Probability Letters*, 80(23-24), 1806-1813.

[4] Bai, J., & Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1), 191-221.

[5] Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics*, 92(4), 1004-1016.