Evaluation of latent and observable factors

DUONG Rayhaan

June 2023

Contents

1 Introduction

1.1 Abstract

This paper aims to highlight the theory developed in Bai and Ng's (2005) [1] article, Evaluating Latent and Observed Factors in Macroeconomics and Finance, which focuses on assessing observable factors as proxies within a multi-factor model. Bai and Ng's work contributes to the literature on identifying and estimating latent factors from macroeconomic and financial data, combining principal component analysis with statistical testing methods to evaluate the relevance of observable proxies.

The goal of this study is not to reproduce the full mathematical derivations of Bai and Ng, but rather to review and internalize the key concepts and methodological tools they introduce. This conceptual understanding is crucial for implementing empirical methods in the context of my own research project.

In particular, this work seeks to examine whether certain information indicators can serve as proxies for latent factors in equity and cryptocurrency markets. It relies not only on Bai and Ng's methodology but also on information theory as studied by De Long and Kahneman, to evaluate the ability of observable indicators to capture relevant information on the price in these markets.

1.2 Objectives

The primary objective of this study is to gain a solid understanding and mastery of sophisticated mathematical tools for evaluating factors within a multi-factor model. By familiarizing oneself with these techniques.

2 Theory

2.1 The Factor model

We will not introduce all the assumptions made in the Bai and Ng article, as they are fully detailed there. It is not necessary to know the specific conditions under which their results converge in order to understand the main ideas.

The model is given by :

$$X = F\Lambda^T$$

With X a $T \times N$ matrix representing the observed data, F a $T \times r$ matrix of latent factors, and Λ an $N \times r$ matrix of factor loadings.

The latent factors are often difficult to handle in practice. Some practitioners estimate them using statistical tools such as PCA or factor analysis. The drawback is that these factors usually lack an economic interpretation, making it unclear what drives the prices. This is why we attempt to proxy the latent factors with observable ones. Consequently, we need a way to assess whether an observable factor is indeed a good proxy for the latent factors. What follows presents a method to do so.

There are four main assumptions in the article by Bai and Ng that ensure the convergence results. As a first approximation, we can assume that when dealing with a large amount of market data, these assumptions are reasonable.

Bai and Ng use PCA to estimate the latent factors F_t because these factors capture the common variation across all observed variables. By extracting the principal components of the data matrix, we identify the directions of largest shared variance, which can be used to capture the underlying latent factors in the factor model.

We start by normalizing the X matrix of data by removing the mean and dividing by the standard deviation in order to apply PCA. The normalized correlation matrix, as in Bai and Ng (2002), is given by:

$$\frac{1}{NT} X X^\top$$

Then we calculate the eigenvalues and eigenvectors of this matrix. We rank the eigenvectors in decreasing order of their corresponding eigenvalues and set the estimated factors as:

$$\tilde{F} = \sqrt{T}\, U_{[:,1:r]}$$

where r is the number of latent factors and $U_{[:,1:r]}$ denotes the matrix containing the first r eigenvectors.

where X denotes the $T \times N$ matrix of standardized data. We can note that Bai and Ng provide a rigorous method to determine the number of latent factors r, based on the data.

After estimating the latent factors $\tilde{F}$ using PCA, the factor loadings $\tilde{\Lambda}$ are obtained by projecting the normalized data X onto these factors:

$$\tilde{\Lambda} = \frac{X^\top \tilde{F}}{T}$$

where:

- X is the normalized data matrix of size $T \times N$,
- $\tilde{F}$ is the matrix of estimated latent factors of size $T \times r$,
- T is the number of time periods.

Intuitively, each column of $\tilde{\Lambda}$ corresponds to a latent factor, and each row measures how strongly each asset loads on that factor. This yields the standard PCA decomposition:

$$X \approx \tilde{F} \tilde{\Lambda}^\top$$

2.2 The observed factors

Let's say we observe G, a matrix of size (T, m). We want to know if G is generated by a linear combination of the latent factors. Usually, we want m larger than r, else G can't span the space of the r latent variables.

Bai and Ng present 2 ways of evaluating G, testing G_j individually or a set of vectors. We will present only the individual way, as it is what is needed for my article.

The null hypothesis tested is : G_{jt} is an exact factor : it exists δ_j such that $G_{jt} = F_t \delta_j'$ for all t.

The regression

$$G_{jt} = \gamma_j' \tilde{F}_t + \epsilon$$

Let $\hat{g}_j$ be the least squares estimate of g_j and define

$$\hat{G}_{jt} = \hat{\gamma_j'} \tilde{F}_t$$

Consider the t-statistic

$$t_t(j) = \frac{\hat{G}_{jt} - G_{jt}}{\sqrt{\text{var}(\hat{G}_{jt})}}$$

and its estimator :

$$\hat{t_t(j)} = \frac{\hat{G}_{jt} - G_{jt}}{\sqrt{\hat{\text{var}}(\hat{G}_{jt})}}$$

Working on this t-statistic will provide interesting metrics.

2.3 Proxy evaluation statistics

2.3.1 Frequency of exceeding the critical t-value

$$A(j) = \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}_{\left(|\hat{t}_t(j)| > F^{-1}_{1-\alpha/2}\right)} \tag{1}$$

The $A(j)$ statistic provides a measure of how well the observed factor G_j is captured by its projection $\hat{G}_j$. Specifically, it quantifies the frequency with which G_{jt} deviates from $\hat{G}_{jt}$ beyond the threshold set by α. This allows us to assess the reliability of G_j as a proxy for the latent factors in the model.

If the observed factor G_{jt} is close to its projection $\hat{G}_{jt}$ then most of the variation in G_{jt} can be explained by the latent factors F. This implies that G_j is a reliable proxy for a latent factor, as it captures the essential information embedded in the underlying factor structure.

2.3.2 Maximum t-statistic over time

$$M(j) = \max_{1 \leq t \leq T} | \hat{t}_t(j) |$$

It tests if $\hat{G}_{jt}$ is far from G_{jt}. A small $M(j)$ reinforces the idea that G_j is a good proxy for a latent factor, because the component of G_j not explained by F is small at all times.

Table 1: Critical values for the $M(j)$ test

T	50	100	200	400
0.01	3.775	3.935	4.109	4.219
0.05	3.283	3.467	3.656	3.830
0.10	3.076	3.278	3.475	3.632

This table is obtained thanks to the fact that under the right assumptions (the series exhibits no serial correlation), $M(j)$ is following the law of $|N(0,1)|$.

2.3.3 Noise to signal ratio and explained variance

$$\text{NS}(j) = \frac{\text{var}(\hat{d}(j))}{\text{var}(\hat{G}(j))} \quad \text{et} \quad R^2(j) = \frac{\text{var}(\hat{G}(j))}{\text{var}(G(j))}.$$

The $\text{NS}(j)$ statistic measures the noise-to-signal ratio: it is zero if G_j is an exact factor, and large values indicate substantial deviations from the latent factors. Similarly, $R^2(j)$ equals one for an exact factor and zero if G_j is irrelevant, allowing the practitioner to assess the quality of G_j as a proxy by comparing $\text{NS}(j)$ and $R^2(j)$ to chosen thresholds.

3 Illustrative example

We perform simulations to assess the finite-sample properties of the tests. The latent factors $F_{kt} \sim \mathcal{N}(0,1)$, $k = 1, \ldots, r$, and the idiosyncratic errors $e_{it} \sim \mathcal{N}(0, \sigma^2_{e(i)})$ are serially uncorrelated and independent across i, j. When $\sigma^2_{e(i)} = \sigma^2_e$ for all i, the data are homogeneous. Factor loadings $\lambda_{ij} \sim \mathcal{N}(0,1)$, and the data are generated as

$$x_{it} = \lambda_i' F_t + e_{it}.$$

We standardize x_{it} to have zero mean and unit variance before estimating factors via principal components. Observed factors are generated as

$$G_{jt} = d_j' F_t + \eta_{jt}, \quad \eta_{jt} \sim s_{\eta(j)} \mathcal{N}\Big(0, \text{var}(d_j' F_t)\Big),$$

where d_j is a $r \times 1$ vector of weights. In our experiments, we test $m = 7$ observed variables.

j	1	2	3	4	5	6	7
d_{j1}	1	1	1	1	1	1	0
d_{j2}	1	0	1	0	1	0	0
s_η	0	0	0.2	0.2	2	2	1

Table 2: Parameters for the observed factors G_{jt}.

I In this toy example, G_1 and G_2 are exact linear combinations of the latent factors. G_3 and G_4 are linear combinations but with a bit of noise. G_5 and G_6 are also linear combinations but with much more noise. G_7 is uncorrelated to the latent factors.

The computation is available in the Jupyter Notebook accessible in the GitHub.

j	1	2	3	4	5	6	7
$A(j)$	0.000	0.000	0.000	0.000	0.030	0.145	0.665
$M(j)$	0.471	0.402	0.895	1.651	3.201	3.864	15.481
$NS(j)$	0.029	0.023	0.135	0.241	1.100	2.114	33.384
$R^2(j)$	0.972	0.977	0.881	0.806	0.476	0.321	0.029

Table 3: Simulation results for the observed factors G_{jt}.

The results are not surprising, G_1, G_2, G_3 and G_4 seem to be good proxies for latent variables. The A and M metrics are relatively low for G_5 and G_6, which indicates that

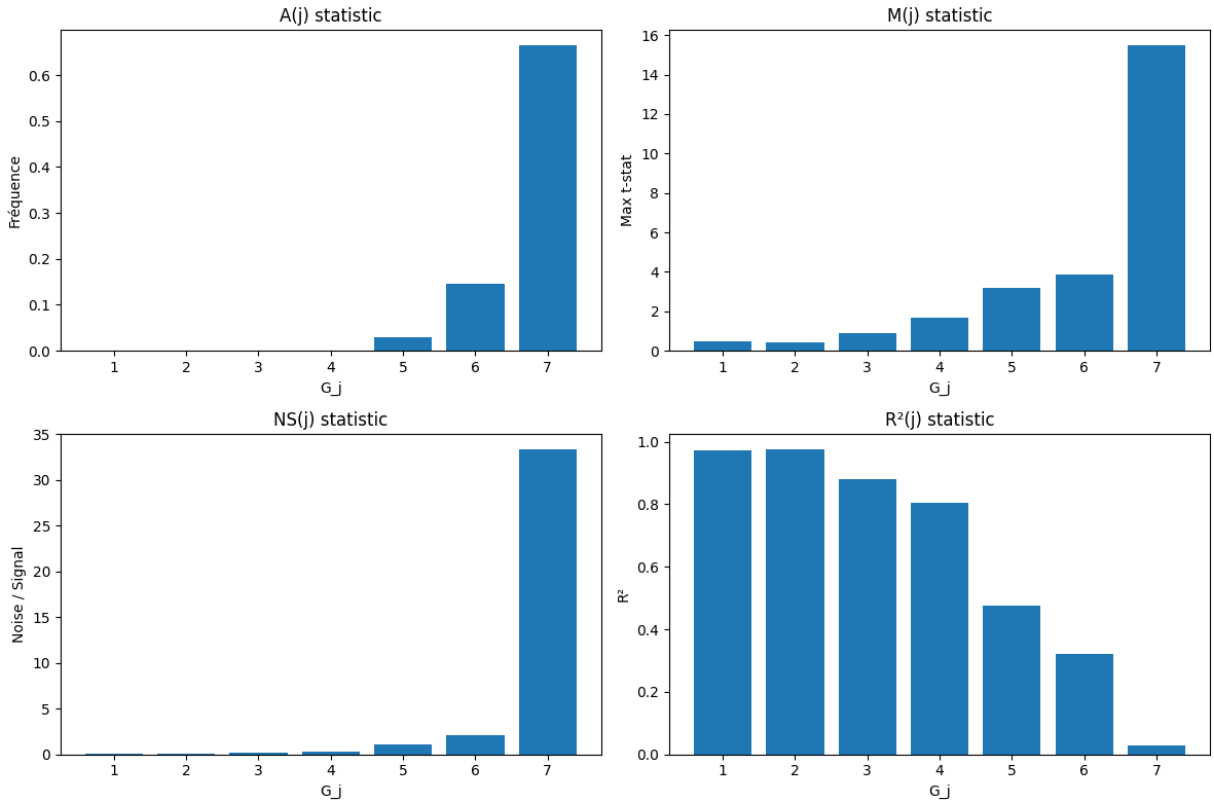

Figure 1: Simulation results for the observed factors G_{jt}.

the deviations of these observed factors from the latent factors are not extreme. However, the high noise-to-signal ratio (NS) makes their quality as proxies less clear, despite the low A and M values. The metrics indicate clearly that G_7 is not a proxy for the latent variables.

4 Conclusion

This example illustrates the effectiveness of the method developed by Bai and Ng. It allows us to assess, using four different metrics, how well an observable variable can serve as a proxy for latent factors. Having gained familiarity with this methodology, we now apply it to the equity and cryptocurrency markets to evaluate information indicators studied by De Long, Kahneman, and other researchers.

References

[1] J. Bai, S. Ng, *Evaluating latent and observed factors in macroeconomics and finance*, Journal of Econometrics, 2005, available online 25 February 2005.