Information criterion for selection of ubiquitous factors

Hellinton H. Takada and Julio M. Stern

Citation: AIP Conference Proceedings **1641**, 378 (2015); doi: 10.1063/1.4906001
View online: http://dx.doi.org/10.1063/1.4906001
View Table of Contents: http://scitation.aip.org/content/aip/proceeding/aipcp/1641?ver=pdfcov
Published by the AIP Publishing

Information Criterion for Selection of Ubiquitous Factors

Hellinton H. Takada[a, b] and Julio M. Stern[b]

[a]*Quantitative Research, Itaú Asset Management, São Paulo, Brazil*
[b]*Institute of Mathematics and Statistics, University of São Paulo, São Paulo, Brazil*

Abstract. Factor analysis is a statistical procedure to describe observed data in terms of unobserved variables called factors. Naturally, it is necessary to determine the number of factors to represent the system. There are several existent criteria to deal with the tradeoff between reduction of approximation error and avoidance of overparameterization. However, given the factors there is a lack of an approach to verify if they are really equally inherent to the entire data. In this paper, the term ubiquitous factors is coined to describe such equally omnipresent factors. An information criterion is proposed to fill the existent blank. Additionally, we show the possibility to use the criterion to compare ubiquity of factors from two different techniques: principal component analysis and non-negative matrix factorization. Finally, the proposed criterion is extended to identify factors more suitable to describe only a partition of the data.

Keywords: Information theory, Entropy, Financial markets.
PACS: 89.70.-a, 89.70.Cf, 89.65.Gh

INTRODUCTION

Originally, factor analysis (FA) was developed in social sciences and psychology [1]. It is a statistical procedure to describe observed data in terms of unobserved variables called factors [2,3]. The objective of FA is to reduce the dimensionality of the original data $X = [x_{ij}] \in \mathbb{R}^{m \times p}$, $m \wedge p \in \mathbb{N}^+$, using an approximation $\tilde{X} = [\tilde{x}_{ij}] \in \mathbb{R}^{m \times p}$ such that:

$$X \approx \tilde{X} = Z\Lambda, \tag{1}$$

where $Z = [z_{ij}] \in \mathbb{R}^{m \times k}$ is the matrix of factors or unobserved (latent) variables; $\Lambda = [\lambda_{ij}] \in \mathbb{R}^{k \times p}$ is the matrix of factor loadings or weights; k represents the number of factors $(k \leq p)$. In the literature, there are some factorization techniques to find Z and Λ. The most popular approach is the principal component analysis (PCA) which was introduced by Pearson [4] and developed by Hotelling [5]. An example of a more recent technique is the non-negative matrix factorization (NNMF) introduced by Paatero and Tapper [6] and popularized by Lee and Seung [7].

In exploratory FA, it is necessary to determine the number of factors k. PCA has a long list of possible approaches to select k: Akaike information criterion [8], minimum description length [9], imbedded error function [10], cumulative percent variance [11], scree test on residual percent variance [12], average eigenvalue [13],

Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MaxEnt 2014)
AIP Conf. Proc. 1641, 378-386 (2015); doi: 10.1063/1.4906001

parallel analysis [14], autocorrelation [15], cross validation based on the PRESS and R ratio [16], variance of the reconstruction error [17], etc. On the other hand, NNMF also has some alternatives to choose: three Bayesian information criterion [18], relative root of sum of square differences [19], volume-based method [20], cophenetic correlation coefficient method [21], bi-cross-validation method [22], etc.

Obviously, the existent criteria deal with the tradeoff between reduction of approximation error and avoidance of overparameterization. However, it is not true that the factors produced using the mentioned criteria are necessarily equally inherent to all data. In the FA literature, the factors are usually referred as common trends. However, that is not true because sometimes obtained factors describe only part of the p columns of X. In this paper, given the factors a criterion is presented to find the most ubiquitous (or omnipresent) factor or factors to all of the p columns of X. Additionally, it is possible to use the proposed criterion to compare the ubiquity degree of factors obtained from different factorization techniques.

The paper is organized as follows: firstly, the ubiquitous factor criterion (UFC) is introduced. Then, the UFC is applied to PCA and NNMF in the context of financial time series to find the more nearly ubiquitous factors. In the sequence, the UFC is extended to enable the identification of specific factors for partitions of the p columns of X. Finally, the conclusion together with more comments about the results are given at the end.

UBIQUITOUS FACTORS

Ubiquitous Factor Criterion

In this section, the ubiquitous factor criterion (UFC) is introduced. The factor model given by (1) is usually implemented with the following restrictions on factor loadings:

$$\sum_{j=1}^{p} \lambda_{ij}^2 = 1, \forall i \in \{1,2,\dots,k\}. \tag{2}$$

Considering the restriction (2) and noticing that $\lambda_{ij}^2 \geq 0, \forall i \in \{1,2,\dots,k\}, \forall j \in \{1,2,\dots,p\}$, it is possible to define $\mathcal{U}_i$ for each factor i using the discrete Shannon entropy [23] as follows:

$$\mathcal{U}_i \triangleq -\sum_{j=1}^{p} \lambda_{ij}^2 \log \lambda_{ij}^2. \tag{3}$$

The Shannon entropy quantifies the expected value of information contained in the sequence $\lambda_{i1}^2, \lambda_{i2}^2, \dots, \lambda_{ip}^2$. In the previous definition, it is usual to consider $0 \log 0 \triangleq 0$. Using (3), it is possible to state the UFC:

> Given a number k of factors and calculating $\mathcal{U}_i, i = 1,2,\dots,k$, the higher the value of $\mathcal{U}_i$, the more nearly ubiquitous (or omnipresent) the factor i.

It is also important to notice that the lower the value of $\mathcal{U}_i$, the more specific the factor i. In the next section, a sample application using financial time series is presented.

Sample Application

In this section, the UFC is applied to PCA and NNMF to find the most ubiquitous factors in financial time series. PCA has been applied to several problems in finance from yield curves to investment risk factors. On the other hand, NNMF was applied in [24] to identify factors in stock market data. The prices considered here are from some exchange tradable funds (ETFs) from the Brazilian stock exchange (BM&F Bovespa) for the period from 01/02/2012 to 03/19/2014. Specifically, the ETFs chosen are: 1) BOVA11, 2) BRAX11, 3) CSMO11, 4) DIVO11, 5) FIND11, 6) GOVE11, 7) ISUS11, 8) MATB11, 9) MILA11, 10) MOBI11, 11) PIBB11 and 12) SMAL11. Consequently, $m = 556$ and $p = 12$. Additionally, all the prices were normalized to begin at 1; the resulting factors are in variance decreasing order; the restriction (2) is respected; for comparison purposes, it will be adopted $k = 3$ for both PCA and NNMF.

Singular value decomposition (SVD) is a technique from linear algebra used to obtain the principal components [25]. The SVD factorization results:

$$\hat{X} = USV', \tag{4}$$

where $\hat{X}$ is obtained mean centering the data matrix X; $U = [u_{ij}] \in \mathbb{R}^{m \times m}$; $S = [s_{ij}] \in \mathbb{R}^{m \times p}$; $V = [v_{ij}] \in \mathbb{R}^{p \times p}$; $UU' = I_m$; $VV' = I_p$; the columns of U and V are orthonormal eigenvectors of XX'; S is a diagonal matrix containing the square roots of the corresponding eigenvalues from U or V such that $s_{11} \geq s_{22} \geq \cdots \geq s_{pp}$, since usually $m \geq p$. Given $k \leq \min(m, p)$, the PCA k-factor model is:

$$\hat{X} \approx U\tilde{S}\tilde{V}', \tag{5}$$

where $\tilde{S} = [s_{ij}] \in \mathbb{R}^{m \times k}$ and $\tilde{V} = [v_{ij}] \in \mathbb{R}^{p \times k}$. The columns of $U\tilde{S}$ are the factors and the columns of $\tilde{V}$ are the corresponding factor loadings. Consequently, the UFC statistics for PCA are given by:

$$\mathcal{U}_i^{PCA} = -\sum_{j=1}^{p} v_{ji}^2 \log v_{ji}^2, \forall i \in \{1,2, \dots, k\}. \tag{6}$$

The obtained factors and factor loadings for PCA are in **FIGURE 1** and **FIGURE 2**, respectively. The UFC statistics $\mathcal{U}_i^{PCA}$ are in **TABLE 1**. It is possible to notice that the first factor is the most nearly ubiquitous one. On the other hand, the third factor is the second most nearly ubiquitous one while the second factor is the third in terms of nearly ubiquity.

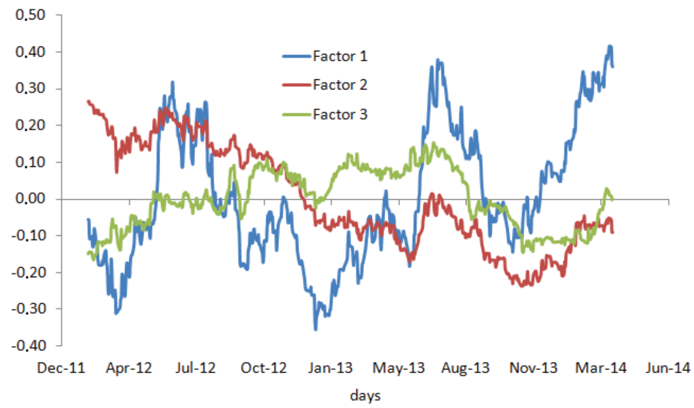

FIGURE 1. Factors obtained using PCA.

FIGURE 2. Factor loadings obtained using PCA.

Since the matrix of historical prices is nonnegative $X = [x_{ij}] \in \mathbb{R}_{\geq 0}^{m \times p}$ and given the integer $k \leq \min(m, p)$, the NNMF problem is to find the following approximation:

$$X \approx \tilde{X} = FB, \tag{7}$$

where $\tilde{X} = [\tilde{x}_{ij}] \in \mathbb{R}_{\geq 0}^{m \times p}$; $F = [f_{ij}] \in \mathbb{R}_{\geq 0}^{m \times k}$; $B = [b_{ij}] \in \mathbb{R}_{\geq 0}^{k \times p}$. It is possible to notice that the columns of F represent the factors and the rows of B the factor loadings.

The NNMF optimization procedures minimizes the approximation error between X and $\tilde{X}$. In a generalized way, the Bregman divergence $D_\varphi(X \| \tilde{X})$ is used as the objective function to be minimized [26,27]. Considering only separable Bregman divergences,

$$D_\varphi(X \| \tilde{X}) = \sum_{ij} D_\varphi(x_{ij} \| \tilde{x}_{ij}) = \sum_{ij} \{\varphi(x_{ij}) - \varphi(\tilde{x}_{ij}) - \nabla\varphi(\tilde{x}_{ij})[\varphi(x_{ij}) - \varphi(\tilde{x}_{ij})]\}, \tag{8}$$

where $\varphi(\cdot)$ is a strictly convex function with a continuous first derivative. Formally, the resulting optimization problems are:

$$\min_{F,B\geq 0}\{D_\varphi(X\|FB) + J_F(F) + J_B(B)\} \tag{9}$$

or

$$\min_{F,B\geq 0}\{D_\varphi(FB\|X) + J_F(F) + J_B(B)\} \tag{10}$$

where $J_F(\cdot)$ and $J_B(\cdot)$ are penalty functions to enforce certain application-dependent characteristics of the solution, such as sparsity and/or smoothness. It is also important to remember that the Bregman divergences are not symmetric in general. Here, we consider $D_\varphi(X\|FB)$.

Adopting $\varphi(x) = x^2/2$ and $J_F(\cdot) = J_B(\cdot) = 0$, there are some known algorithms to solve the NNMF problem divided in general classes [28]: gradient descent algorithms, multiplicative update algorithms and alternating least squares algorithms (ALS). Here, the ALS will be adopted (the use of other algorithms does not provide great differences to the sample example presented here) and the UFC statistics for NNMF are

$$\mathcal{U}_i^{NNMF} \triangleq -\sum_{j=1}^{p} b_{ij}^2 \log b_{ij}^2, \forall i \in \{1,2,\dots,k\}. \tag{11}$$

The obtained factors and factor loadings for NNMF are in **FIGURE 3** and **FIGURE 4**, respectively. The UFC statistics $\mathcal{U}_i^{NNMF}$ are in **TABLE 1**. It is possible to notice that factors are already in the decreasing nearly ubiquity degree order.

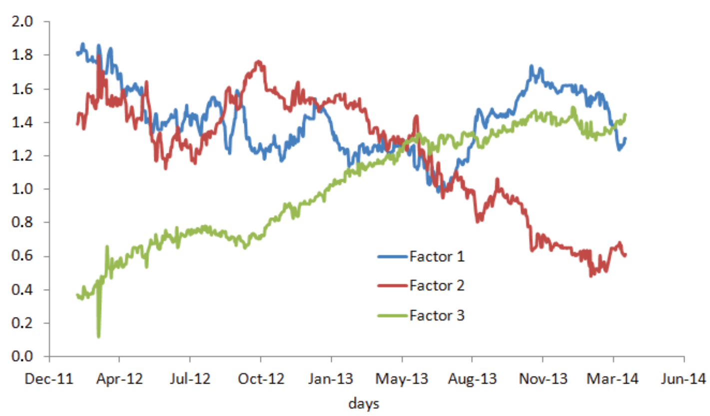

FIGURE 3. Factors obtained using NNMF.

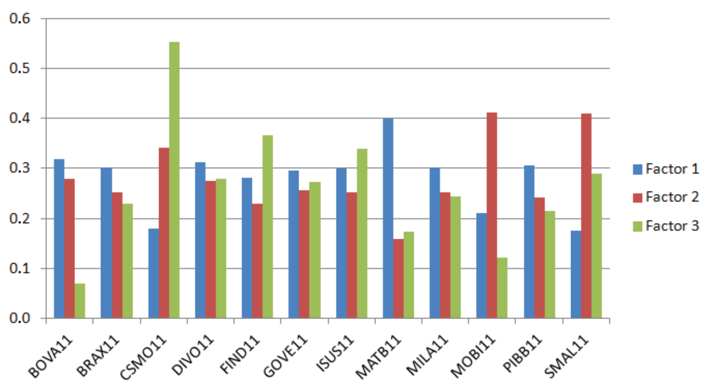

FIGURE 4. Factor loadings obtained using NNMF.

TABLE 1. UFC and SFC statistics for PCA and NNMF factors.

	$\mathcal{U}_i^{PCA}$	$\mathcal{U}_i^{NNMF}$	$\mathcal{S}_i^{PCA}(\boldsymbol{q}^{(e)})$	$\mathcal{S}_i^{NNMF}(\boldsymbol{q}^{(e)})$
first factor ($i = 1$)	2.1599	2.3991	0.3250	0.0858
second factor ($i = 2$)	1.6152	2.3628	0.8697	0.1221
third factor ($i = 3$)	1.6904	2.1394	0.7945	0.3455

Finally, it is also possible to notice that the nearly ubiquity degree for NNMF factors are higher when compared with the statistics for PCA. Consequently, for the considered data the NNMF factors represent better nearly ubiquitous factors than PCA. In other words, in our example, the NNMF factors are better to find common trends than PCA factors.

SPECIFIC FACTOR CRITERION

Cluster analysis has the objective of grouping objects in partitions. In the literature, there are several related algorithms: hierarchical clustering and k-means are popular examples. Additionally, the use of information theory in cluster analysis is not new. Particularly, the Kullback-Leibler divergence has already been applied to cluster analysis [29]. However, the problem here is quite different: given the factors, a criterion is proposed to select the best factor that describes partitions of the p columns of X. For each factor i, it is possible to define a statistic based on the discrete Kullback-Leibler [30] divergence:

$$\mathcal{S}_i(\boldsymbol{q}) \triangleq \sum_{j=1}^{p} \lambda_{ij}^2 \log\left(\lambda_{ij}^2 / q_j\right). \qquad (12)$$

The discrete Kullback-Leibler divergence is a non-symmetric measure of the difference between two mass distributions. Using (12), it is possible to state the specific factor criterion (SFC):

> Given a number k of factors and calculating $\mathcal{S}_i(\boldsymbol{q}), i = 1,2,\dots,k$, the lower the value of $\mathcal{S}_i$, the more specific is the factor i to a partition of the p columns of X described by $\boldsymbol{q}$.

The vector $\boldsymbol{q}$ is chosen to create partitions of the p columns of X. In the following, some particular cases of $\boldsymbol{q}$ are empirically studied using the same data from the previous section. Considering a vector $\boldsymbol{q}^{(e)}$ given by:

$$q_j^{(e)} = 1/p\,, \forall j = 1, \dots, p, \tag{13}$$

the SFC acts as the UFC. The SFC statistics obtained are presented in **TABLE 1** and they bring the same conclusions obtained using the UFC statistics.

Arbitrarily, choosing a vector $\boldsymbol{q}^{(i)}$ such that

$$q_j^{(i)} = \begin{cases} (1 - 9 * \varepsilon)/3\,, & j \in \{3,5,7\} \\ \varepsilon, & \text{otherwise} \end{cases}, \tag{14}$$

and a second vector $\boldsymbol{q}^{(d)}$

$$q_j^{(d)} = \begin{cases} (1 - 9 * \varepsilon)/3\,, & j \in \{1,10,12\} \\ \varepsilon, & \text{otherwise} \end{cases}, \tag{15}$$

where ε is a very small positive number (considered here 10^{-6}), the SFC statistics were calculated and the results are in **TABLE 2**. Clearly, the factor that best describes the partition given by $\boldsymbol{q}^{(i)}$ is the factor 3 and the partition $\boldsymbol{q}^{(d)}$ is the factor 2. Observing **FIGURE 3**, it is possible to notice an increasing trend (given by factor 3) and a decreasing trend (given by factor 2). Obviously, the ETFs CSMO11, FIND11 and ISUS11 have predominantly increased, while BOVA11, MOBI11 and SMAL11 have predominantly decreased in the considered historical data. Consequently, the SFC identified the factors that best describe the common trend of each set of ETFs chosen.

TABLE 2. SFC statistics for PCA and NNMF factors.

	$\mathcal{S}_i^{NNMF}(\boldsymbol{q}^{(i)})$	$\mathcal{S}_i^{NNMF}(\boldsymbol{q}^{(d)})$
first factor ($i = 1$)	8.8522	9.1667
second factor ($i = 2$)	8.4927	**6.1651**
third factor ($i = 3$)	**4.6169**	10.3602

CONCLUSIONS

In the literature, there are several existent criteria to find the number of factors considering the tradeoff between reduction of approximation error and avoidance of overfitting. However, given the factors there is a lack of an approach to verify if they are really ubiquitous to the entire data. In this paper, the ubiquitous factor criterion is introduced to fill the blank. Additionally, a criterion is also proposed to identify more suitable factors to describe only a partition of the data. Applications of the criteria using financial time series show their usefulness to select the best overall and partition specific trends and to compare different factorization techniques such as PCA and NNMF.

REFERENCES

1. C. Spearman, ""General intelligence," objectively determined and measured," *Am. J. Psychol.* **15** (2), 201–292 (1904).
2. Z. Ghaharamani and G. E. Hilton, "The EM algorithm for mixtures of factor analyzers," *Tech. Rep.* CRG-TR-96-1, Dept. of Computer Science, Univ. of Toronto, 1997.
3. B. S. Everitt, *Latent Variable Models*, London: Chapman and Hall, 1984.
4. K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philos. Mag.* **2** (11), 559–572 (1901).
5. H. Hotelling, "Analysis of a complex of statistical variables into principal components," *J. Educational Psychol.* **24** (6), 417–441 (1933).
6. P. Paatero and U. Tapper, "Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values," *Environmetrics* **5** (2), 111–126 (1994).
7. D. Lee and H. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature* **401**, 788–791 (1999).
8. H. Akaike, "Information theory and an extension of the maximum likelihood principle," *Proc. 2nd International Symposium on Information Theory*, 267–281 (1974).
9. J. Rissanen, "Modeling by shortest data description," *Automatica* **14**, 465–471 (1978).
10. E. R. Malinowski, "Determination of the number of factors and the experimental error in a data matrix," *Analytical Chemistry* **49** (4), 612–617 (1977).
11. E. R. Malinowski, *Factor Analysis in Chemistry*, New York: Wiley-Interscience, 1991.
12. R. B. Cattell, "The scree test for the number of factors," *Multivariate Behavioral Research* **1**, 245–276 (1966).
13. H. F. Kaiser, "The application of electronic computers to factor analysis," *Educational and Psychological Measurement* **20** (1), 141–151 (1960).
14. W. R. Zwick and W. F. Velicer, "Comparison of five rules for determining the number of components to retain," *Psychological Bulletin* **99** (3), 432–442 (1986).
15. R. I. Shrager and R. W. Hendler, "Titration of individual components in a mixture with resolution of difference spectra, pKs, and redox transitions," *Analytical Chemistry* **54** (7), 1147–1152 (1982).
16. S. Wold, "Cross validatory estimation of the number of components in factor and principal components analysis," *Technometrics* **20**, 397–406 (1978).
17. S. J. Qin and R. Dunia, "Determining the number of principal components for best reconstruction," *J. Process Control* **10**, 245–250 (2000).
18. J. Bai and S. Ng, "Determining the number of factors in approximate factor models," *Econometrica* **70** (1), 191–221 (2002).
19. X. Shao, G. Wang, S. Wang and Q. Su, "Extraction of mass spectra and chromato-graphic profiles from overlapping GC/MS signal with background," *Analytical Chemistry* **76** (17), 5143–5148 (2004).
20. P. Fogel, S. S. Young, D. M. Hawkins and N. Ledirac, "Inferential, robust non-negative matrix factorization analysis of microarray data," *Bioinformatics* **23** (1), 44–49 (2007).
21. J. Brunet, P. Tamayo, T. R. Golub and J. P. Mesirov, "Metagenes and molecular pattern discovery using matrix factorization," *Proc. National Academy of Sciences of the United States of America* **101** (12), 4164–4169 (2004).
22. A. B. Owen and P. O. Perry, "Bi-cross-validation of the SVD and the non-negative matrix factorization," *Tech. Rep.*, Stanford Univ., 2008.
23. C. E. Shannon, "A mathematical theory of communication," *Bell System Tech. J.* **27**, 379-423 (1948).
24. K. Drakakis, S. Rickard, R. de Fréin and A. Cichocki, "Analysis of financial data using non-negative matrix factorization," *International Mathematical Forum* **3** (38), 1853–1870 (2008).
25. G. H. Golub and C. F. V. Loan, *Matrix Computations*, The Johns Hopkins Univ. Press, 1996.

26. I. S. Dhillon and S. Sra, "Generalized nonnegative matrix approximations with Bregman divergences," *Advances in Neural Information Processing Systems* **18**, 283–290 (2005).
27. L. Li, G. Lebanon and H. Park, "Fast Bregman divergence NMF using Taylor expansion and coordinate descent," *Proc. 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, August, 2012.
28. M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, R. J. Plemmons, "Algorithms and applications for approximate nonnegative matrix factorization," Computational Statistics and Data Analysis **52**, 155–173 (2007).
29. A. Sheehy, "Maximal Kullback-Leibler divergence cluster analysis," *Tech. Rep.* 113, Dept. of Statistics, Univ. of Washington, 1987.
30. S. Kullback and R. A. Leibler, "On information and sufficiency," *Annals Mathematical Statistics* **22**, 79-86 (1951).