Information Criterion for Selection of Ubiquitous Factors

21/09/2014

Auteurs :

Abstract

Collection

Authors

Metrics

License

Sponsors

MaxEnt 2014

- Abstract
- Collection
- Authors
- Metrics
- License
- Sponsors

Abstract

Collection

Information Criterion for Selection of Ubiquitous Factors

Authors

Julio Stern

Information Criterion for Selection of Ubiquitous Factors

Non-Negative Matrix Factorization and Term Structure of Interest Rates

Cognitive-Constructivism, Quine, Dogmas of Empiricism, and Muenchhausen’s Trilemma

Hellinton
La SEE (Société de l'Electricité, de l' Electronique et des Technologies de l' Information et de la Communication – Association reconnue d'utilité publique, régie par la loi du 1er juillet 1901) met à la disposition de ses adhérents et des abonnés à ses publications, un ensemble de documents numériques accessibles à partir de son portail des publications. Ces documents incluent notamment les articles des revues REE, 3 El et e-STA disponibles sous forme numérique ainsi que des publications additionnelles regroupées dans l'espace eREE. Les présentes conditions précisent les conditions de diffusion et d'utilisation de ces documents et des informations qu'ils contiennent. L'accès à ces documents, qu'il se fasse de façon gratuite ou dans le cadre d'abonnements ou d'achats faits à titre onéreux, implique l'acceptation sans restriction de ces dispositions.

Droits de propriété et de diffusion des contenus téléchargés sur le portail des publications

Les contenus rendus accessibles sur le portail des publications sont, en règle générale, protégés par le droit d'auteur. En tant que producteur, et le cas échéant d'auteur, des informations rassemblées dans les contenus accessibles par ce portail, SEE se réserve l'exclusivité des droits de copie et de diffusion de tout ou partie de ces contenus.

Les contenus sont rendus accessibles à titre individuel, pour les besoins de la personne en détenant des droits d'accès en cours de validité. Aussi, la modification, la reproduction et/ou la diffusion via Internet ou le Web, intranet, extranet ou toute autre forme numérique ou imprimée, de tout ou partie des contenus téléchargés sont interdites. Une tolérance est consentie quant à la reproduction d'extraits limités de ces contenus, dans le cadre de travaux ou d'activités auxquels ils sont utiles, à la condition que l'origine de ces reproductions partielles soit mentionnée de façon lisible et sans ambiguïté. Figureront en particulier : la REE (ou toute autre revue accessible sur le portail) en tant que la source, la référence de la publication et le nom de l'auteur (s'il figure dans la
Ces dispositions s’appliquent également aux figures, illustrations, logos ou images.

**Publication externe des contenus du portail des publications**

Tout extrait des contenus du portail destiné à être utilisé dans des publicités, des communiqués de presse ou du matériel de promotion nécessite un accord préalable écrit de la SEE. Une version préliminaire du document proposé contenant ces extraits doit accompagner chacune de ces demandes. SEE se réserve le droit de refuser un tel usage externe pour quelque raison que ce soit.

**Responsabilités**

La SEE apporte tout le soin possible à la préparation des informations délivrées dans les contenus produits. Cependant elle ne peut être tenue pour responsable d’aucune perte ou frais qui pourrait résulter d'imprécisions, d'inexactitudes, d'erreurs ou de possibles omissions portant sur des informations publiées, ni des résultats obtenus par l'utilisation et la pratique des informations délivrées.

**Utilisation des informations recueillies lors du téléchargement de contenu**

Le portail des publications est susceptible d'utiliser des « cookies » afin notamment de permettre l'utilisation de paniers d'achat et de personnaliser les parcours sur le site. SEE se réserve la possibilité d'utiliser les informations recueillies lors des téléchargements pour ses besoins internes et notamment pour l'amélioration de ses services, sans qu'elles puissent être cédées à des partenaires commerciaux. Conformément à la loi "informatique et libertés" du 6 janvier 1978, chaque utilisateur du portail dispose d'un droit d’accès et de rectification aux informations qui le concernent. Pour exercer ce droit, les utilisateurs doivent s’adresser à SEE – 17 rue de l’amiral Hamelin – 75783 Paris Cedex 16, par simple lettre ou en utilisant le formulaire de contact disponible sur son site.

Paris, le 28 avril 2013

**Sponsors**

**Scientific sponsors**

Logistic sponsors
Information Criterion for Selection of Ubiquitous Factors

Hellinton H. Takadaa, b and Julio M. Sternb

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 and 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 and it is proposed an information criterion to fill the existent blank. Additionally, it is also shown 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 and it is a statistical procedure to describe observed data in terms of unobserved variables called factors. The objective of FA is to reduce the dimensionality of the original data, using an approximation such that: (1) where is the matrix of factors or unobserved (latent) variables, is the matrix of factor loadings or weights, represents the number of factors and . In the literature, there are some factorization techniques to find and . The most popular approach is the principal component analysis (PCA) and it was introduced by Pearson and developed by Hotelling. An example of a more recent technique is the non-negative matrix factorization (NMF) introduced by Paatero and Tapper and popularized by Lee and Seung. In exploratory FA, it is necessary to determine the number of
factors. PCA has a long list of possible approaches to select: Akaike information criterion, minimum description length, imbedded error function, cumulative percent variance, scree test on residual percent variance, average eigenvalue, parallel analysis, autocorrelation, cross validation based on the PRESS and R ratio, variance of the reconstruction error, etc. On the other hand, NNMF has also some alternatives to choose: three Bayesian information criterion, relative root of sum of square differences, volume-based method, cophenetic correlation coefficient method, bi-cross-validation method, 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 columns of . In this paper, given the factors a criterion is presented to find the most ubiquitous (or omnipresent) factor or factors to all of the columns of . 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 most ubiquitous factors. In the sequence, the UFC is extended to enable the identification of specific factors for partitions of the columns of . 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 is usually implemented with the following restrictions on factor loadings: (2) Considering the restriction and noticing that , it is possible to define for each factor using the discrete Shannon entropy as follows: (3) The Shannon entropy quantifies the expected value of information contained in the sequence . In the previous definition, it is usual to consider . Using , it is possible to state the UFC: Given a number of factors and calculating , the higher the value of , the more ubiquitous (or omnipresent) the factor . It is also important to notice that the lower the value of , the more specific the factor . 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 to identify factors in stock market data. The prices considered here are from some exchange tradable funds (ETFs) from 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) FND11, 6) GOVE11, 7) ISUS11, 8) MATB11, 9) MILA11, 10) MOBI11, 11) PIBB11 and 12) SMAL11. Consequently, and . Additionally, all the prices were normalized to begin at , the resulting factors are in variance decreasing order, the restriction is respected and, for comparison purposes, it will be adopted for both PCA and NNMF. Singular value decomposition (SVD) is a technique from linear algebra used to obtain the principal components . The SVD factorization results: (4) where is obtained mean centering the data matrix , , , , , , the columns of and are orthonormal eigenvectors of and is a diagonal matrix containing the square roots of the corresponding eigenvalues from or such that , since usually . Given , the PCA -factor model is: (5) where and . The columns of are the factors and the columns of are the corresponding factor loadings. Consequently, the UFC statistics for PCA are given by: (6) The obtained factors and factor loadings for PCA are in FIGURE 1 and FIGURE 2, respectively. The UFC statistics are in TABLE 1. It is possible to notice that the first factor is the most ubiquitous one. On the other hand, the third factor is the second most ubiquitous one while the second factor is the third in terms of ubiquity. FIGURE 1. Factors obtained using PCA. FIGURE 2. Factor loadings obtained using PCA. Since the matrix of historical prices is nonnegative and given the integer , the NNMF problem is to find the following approximation: (7) where and . It is possible to notice that the columns of represent the factors and the rows of the factor loadings. The NNMF optimization procedures minimizes the approximation error between and . In a generalized way, the Bregman divergence is used as the objective function to be minimized [26,27]. Considering only separable Bregman divergences, (8) where is a strictly convex function with a continuous first derivative. Formally, the resulting optimization problems are: (9) or (10) where and 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. Consequently, it will be considered here . Adopting and , 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, it will be adopted the ALS (the use of other algorithms does not provide great differences to the sample example presented here) and the UFC statistics for NNMF are (11) The obtained factors and factor loadings for NNMF are in FIGURE 3 and FIGURE 4, respectively. The UFC statistics are in TABLE 1. It is possible to notice that factors are already in the decreasing 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. first factor ( ) 2.1599 2.3991 0.3250 0.0858 second factor ( ) 1.6152 2.3628 0.8697 0.1221 third factor ( ) 1.6904 2.1394 0.7945 0.3455 Finally, it is also possible to notice that the ubiquity degree for NNMF factors are higher when compared with the statistics for PCA. Consequently, for the considered data the NNMF factors represent better ubiquitous factors than PCA. In other words, the NNMF factors are better 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 some popular examples. Additionally, the use of information theory in cluster analysis is not new and, particularly, the Kullback-Leibler divergence has already been applied [29]. However, the problem here is quite different: given the factors it is proposed a criterion to select the best factor that describes partitions of the columns of . For each factor , it is possible to define a statistic based on the discrete Kullback-Leibler [30] divergence: (12) The discrete Kullback-Leibler divergence is a non-symmetric measure of the difference between two mass distributions. Using , it is possible to state the specific factor criterion (SFC): Given a number of factors and calculating , the lower the value of , the more specific the factor to a partition of the columns of described by . The vector is chosen to create partitions of the columns of . In the following, some particular