Compact representations of market securities using smooth component extraction

Hariton Korizis, Nikolaos Mitianoudis, and Anthony G. Constantinides

Communications and Signal Processing Group,
Imperial College London,
Exhibition Road, London SW7 2AZ, UK

ABSTRACT
Independent Component Analysis (ICA) is a statistical method for expressing an observed set of random vectors as a linear combination of statistically independent components. This paper tackles the task of comparing two ICA algorithms, in terms of their efficiency for compact representation of market securities. A recently developed sequential blind signal extraction algorithm, SmoothICA, is contrasted to a classical implementation of ICA, FastICA. SmoothICA uses an additional 2nd order constraint aiming at identifying temporally smooth components in the data set. This paper demonstrates the superiority of this novel smooth component extraction algorithm in terms of global and local approximation capability, applied to a portfolio of 60 NASDAQ securities, by utilizing common ordering algorithms for financial signals.

KEY WORDS
Independent Component Analysis, SmoothICA, FastICA, market securities, Finance

1 Introduction

The goal of Independent Component Analysis is to find a linear representation of non-Gaussian variables. Finding such a representation provides an insight to the underlying structure of many signal processing problems. The ICA problem is equivalent to establishing the following generating model for the data:

$$x = As \tag{1}$$

where x and s are n-dimensional random vectors, and components s are assumed mutually independent. A is a constant $n \times n$ full rank matrix, denoting the unknown mixing matrix. Relevant to our investigation is the formulation that x consists of a set of observation vectors generated in the financial markets, which are driven by the hidden underlying sources s. The driving mechanisms s are mixed and contaminated among others by elements, such as news and expectations related to results of companies and sectors, domestic and foreign politics that affect exchange and interest rates, consumer confidence, unexpected events and even the weather that affects the commodities' prices. The transformation:

$$s = Wx \tag{2}$$

can be defined, with W the demixing matrix and $A = W^{-1}$. This method allows at most one Gaussian component, concentrating all the signal innovations which cannot

be accounted for by the original problem assumptions. In the case of signals originating from the financial markets, this assumption can be considered valid for a great majority of the cases, as purely gaussian financial signals are rarely generated.

The assumption of statistical independence of the source signals, can be assumed to be valid in the scope of global economy and the hugely diverse micro- and macroeconomic factors that affect financial processes. Unexplained noise, as well as the markets' response to large trades can be also of significance to researchers and traders. However it is logical that the underlying driving sources' independence assumption in the financial markets can be debated, as every source might exert a small influence on all others. A review of various contrast functions can be found in [1], as the sources s can be separated using various interpretations of statistical independence.

In a recent paper, the authors in [2] proposed a sequential blind signal extraction algorithm that incorporates a smoothness constraint based on the original FastICA algorithm [3]. Along with the negentropy cost function, the added temporal constraint seeks to find smooth orthogonal projections in the mixed data vector x. In [4], this algorithm is referred to as SmoothICA and contrasted to the performance of FastICA, in the search for temporal structure in the underlying sources that give rise to stock evolutions. A portfolio of 20 NASDAQ securities was analyzed and possible advantages of this novel approach were highlighted over FastICA, as it produced components with smoother temporal structure. A small degree of correlation was present among the components extracted, introduced by the balancing of the 4th and 2nd order temporal constraint.

In Principal Component Analysis (PCA) the components that contribute most to the mixtures can be easily ranked according to the eigenvalues of the source vectors. In ICA the same task is not straightforward, as the corresponding projection vector consists of normalized rows to unity [3]. This work uses common ordering algorithms for financial signals and contrasts their error performance for approximating a given portfolio using reduced numbers of components.

This text is organized in the following way. The next section contains an overview of ICA ordering techniques for financial time series. Section 3 contains an overview of the SmoothICA algorithm and Section 4 contains a detailed view of the ordering algorithms that will be utilized in the experimental Section 5. An ordering method for selecting a reduced number of components for reconstruction of a whole portfolio of securities is considered (global approximation), as well as two methods for ordering the components' contributions of each security signal and reconstructing accordingly. Finally, section 6 concludes this study.

2 History of Independent Component ordering in Finance

Several investigations of ICA with application to finance have been performed. The most influential was done in [5], examining the portfolio returns of 28 Japanese stocks. PCA and ICA performances were compared for such signals. From the operation of ICA, the components produced have $var(s_i) = 1$. It is therefore assumed that any information about the contribution of each individual component, to a mixture's variation is engulfed in the mixing matrix A. The authors used the maximum norm L_∞ to sort

the rows of the A, and thus to determine which ICs have the maximum contribution to a selected signal's amplitude. Such a measure is applied in this research.

Cheung and Xu [6] presented a criterion for ordering source signals, according to their contribution to the trend reservation of each observed signal. This algorithm uses the MSE criterion and is named Testing-and-Acceptance (TnA), and when applied to foreign exchange rates it produces superior results over the L_∞ norm method. This is the second method which will be used in this paper. The same authors have presented a criterion to select the appropriate dimension for the source signal subset to approximate a portfolio of foreign exchange rates in [8]. An algorithm using the Relative Hamming Distance (RHD) instead, was proposed in [7].

A consequence of the increased interest in this type of component extraction and its demonstrated superiority in terms of source separation over PCA, are applications utilizing the ICA capabilities in econometrics and finance problems; from prediction approaches [9] and Factor Model estimation [10] to the computation of the risk of a portfolio of securities [11] and the application of ICA in the context of state space models for interbank foreign exchange rates to obtain a better separation of the observation noise and the "true" price [12]. It is worth focusing on [11] where the contribution of an individual independent component to the variance of the whole portfolio of securities is calculated. The ICs are ordered according to that contribution, and this operates as a preprocessing step for dimensionality reduction before switching back to the prices' space. This is the third method examined in the current paper, testing global approximation performance.

3 The SmoothICA algorithm

After an initial prewhitening step, SmoothICA solves the following inequality constrained optimization problem:

$$\max_{w} \quad J_1(\underline{w}) \tag{3}$$

$$\text{subject to } J_2(\underline{w}) \leq 0 \tag{4}$$

$$J_3(\underline{w}) = 0 \tag{5}$$

where $J_1(\cdot)$ is the approximated negentropy as proposed by Hyvarinen [3], $J_3(\cdot)$ is the unit-norm constraint and $J_2(\underline{w}) = \mathcal{E}\{(\underline{w}^T \underline{\Delta z})^2\} - \rho\mathcal{E}\{(\underline{w}^T \underline{z})^2\}$ is the second-order smoothness criterion and $\rho \in [0, 1]$ defines the degree of smoothness[4]. Modifying the inequality constraint to the equality constraint $\max(J_2(\underline{w}), 0) = 0$, one can find the desired optima using alternating unconstrained maximization of the Lagrangian function $J_1(\underline{w}) + \lambda \max(J_2(\underline{w}), 0) + \kappa J_3(\underline{w})$, where λ, κ are the Lagrange multipliers. The following Newton-step provides an update:

$$\underline{w}^+ \leftarrow \underline{w} - \left[\frac{\partial^2 J}{\partial \underline{w}^2}\right]^{-1} \frac{\partial J}{\partial \underline{w}} \tag{6}$$

where, in this case, the gradient vector and the Hessian matrix are estimated using the following updates :

$$\frac{\partial J}{\partial \underline{w}} = \mu\mathcal{E}\{\underline{z}G'(\underline{w}^T \underline{z})\} + \lambda(\mathcal{E}\{(\underline{w}^T \underline{\Delta z})\underline{\Delta z} - \rho(\underline{w}^T \underline{z})\underline{z}\})(\text{sgn}(J_2) + 1)$$

$$\frac{\partial^2 J}{\partial \underline{w}^2} = \mu \mathcal{E}\{G''(\underline{w}^T \underline{z})\} I + \lambda (C_{\Delta z} - \rho I)(\text{sgn}(J_2) + 1)$$

where $\mu = \text{sgn}(\mathcal{E}\{G(u)\} - \mathcal{E}\{G(v)\})$. After calculating the estimate for $\underline{w}$, we calculate estimates for λ via alternating optimization. The unit-norm constraint is then imposed as a projection of the $\underline{w}$ estimate on the unit hypersphere, to ensure that rotation and not scale deformation is performed:

$$\underline{w}^+ \leftarrow \underline{w}^+ / \|\underline{w}^+\| \tag{7}$$

Subsequent smooth components are extracted using the orthogonal deflation procedure used by Hyvarinen in [1].

4 Ordering Methods for Independent Component Analysis

In Finance dimensionality reduction is applied for various purposes. It is performed to remove unwanted information and hence get a clearer picture of an underlying process, allowing better modeling and understanding of its statistical nature. It it also applied to represent a large set of assets by an appropriate subset that best defines it and reduce memory requirements and computational burden. Unlike PCA, ICA is not constructed to have an inherent ordering of the ICs. The methods below follow two notions; approximation of a particular security using a few ICs (selected according to their contribution to that particular security) and approximation of a whole portfolio of securities by selecting an appropriate subset of independent components.

4.1 Global Approximation

In ICA the components produced are scaled to unit variance. This means that the additional information about individual contributions of the ICs to the observed signals lies in the mixing matrix A [11]. The variance of the security i is σ_i^2 and the amount of total variance V_j explained by each component s_j can be derived from:

$$\sigma_i^2 = \sum_i a_{ij}^2 \quad \text{and} \quad V_j = \frac{\sum_j a_{ij}^2}{\sum_{i,j} a_{ij}^2} \tag{8}$$

Thus by ordering the ICs according to their individual contributions to the whole portfolio, we can approximate efficiently by selecting a reduced number of components.

4.2 Local Approximation

The L_∞ norm: The weighted ICs, as given by (10) with the largest amplitudes are defined to be the dominant ICs. This of course presents an ordering criterion, as these ICs have the largest effect on the securities. The reconstruction of the ith security from the estimates source signals from:

$$\widehat{x}_i = \sum_{k=1}^{m} a_{ik} s_k \tag{9}$$

where s_k is the kth estimated IC and a_{ik} is the weight in the ith row, kth column of A. The weighted ICs are therefore obtained from:

$$\widehat{s}_{ik} = a_{ik} s_k \quad k = 1..n \tag{10}$$

The L_∞ norm was used in [5] to order the weighted ICs for each particular stock, as this measure reveals the magnitude contribution of each source signal to a particular stock.

The Testing-and-Acceptance algorithm: The TnA algorithm in [6] aims at creating a list L_i, whose elements are the component subscripts decided for decreasing contribution to a specified security signal. Initially, the IC which introduces the minimum MSE error of reconstruction of the selected security if omitted, is selected from the m components. The reconstructed security, while the ith component is omitted, is $\{\widehat{y}_j\}_{j=1,j\neq r}^{m}$. The subscript of this IC is put last in the list L. The next step of the iteration starts with a subset of the ICs that do not include the previously selected component. It finds the next component that, while omitted, causes minimum MSE error of approximation, and puts it second to last in L, and so on. It is a suboptimal heuristic method compared with the exhaustive search, however the TnA algorithm involves just $\frac{m(m+1)}{2} - 1$ compared to $(m+1)!$ steps.

The algorithm operates as follows:

1. Let the set of independent component subscripts $Z = \{j \mid 1 \leq j \leq m\}$, $d = 0$, and the order list $L_i = ()$.
2. For each $j \in Z$ and N being the signal's length, let:

$$v_{ij(t)} = \sum_{m\neq j, m\in Z} \widehat{v}_{im(t)} \ , 1 \leq t \leq N \tag{11}$$

The β which will be stored as the d^{th} element of L_i and removed from the set Z, is selected according to:

$$\beta = \arg\min_{j\in Z} MSE(x_i, v_{ij}) \tag{12}$$

$$d^{new} = d^{old} + 1$$

$$\text{Then let:} \quad L_i^{new} = L_i^{old} + \beta$$
$$Z^{new} = Z^{old} - \{\beta\}$$

3. If $Z \neq \{\}$, goto Step 2; otherwise stop. In order to make the list ordered according to descending contribution, flip it.

5 Experiments

5.1 Description of the data

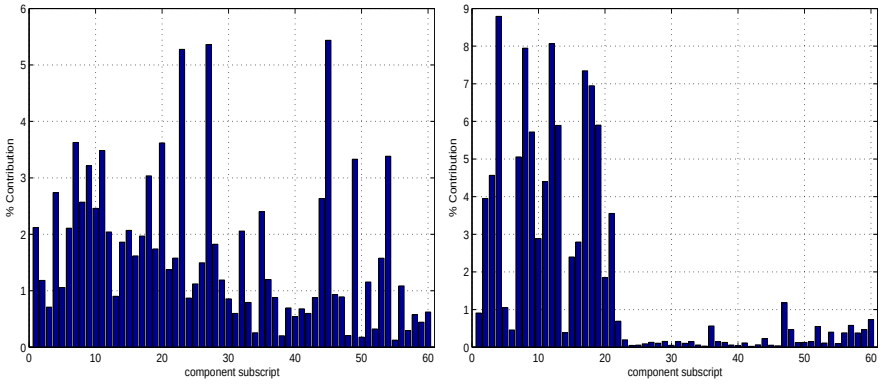
The experiments were performed with daily closing prices of a portfolio of 60 technology stocks[1] from the NASDAQ exchange, for the period ranging from 01/01/2002 to

[1] The portfolio consists of the first 60 stocks (alphabetically) of the NASDAQ US Exchange.

05/04/2005. The portfolio of prices is centered and whitened so that uncorrelated, unit variance signals are obtained. The SmoothICA algorithm is performed on the whitened data as outlined in [2] and [4]. Flexibility is added with ρ starting at 0.05 and increasing in case of non converging components. Using a large portfolio, the issue of low correlation among the components extracted, as mentioned in [4], does not appear. Thus a high number of mixtures is needed for this algorithm to properly converge, due to the added smoothness constraint. The correlation matrix among the sources is now a proper identity matrix. The FastICA algorithm is also applied on the whitened data, which gives the reference results for comparison in terms of approximation fitness for the all the ranges of subset order possible.

5.2 Global Approximation Comparison

After obtaining successful convergence for both algorithms, the percentages of variance contribution of each of their components are calculated, using the expression for V_j in (8). The result is presented on Figure 1. While in the FastICA case there is an almost equal parsing of the variance contributions among the ICs, a significant amount of variance is concentrated in approximately the first 20 ICs that SmoothICA extracts. Indicatively, 35 FastICA ICs contain 70% of the portfolio's variance, while by using the additional smoothness temporal constraint only 12 components are required. This signifies the great advantage in terms of dimensionality reduction and global approximation using a smaller subset of signals. SmoothICA, as observed, produces components that have an inherent ordering of the source signals and can provide a more efficient representation of a portfolio of securities. It can be used among other tasks, as an alternative to dimensionality reduction for simpler modeling or extraction of seasonal and structural variations (currently done during the pre-whitening step by PCA).

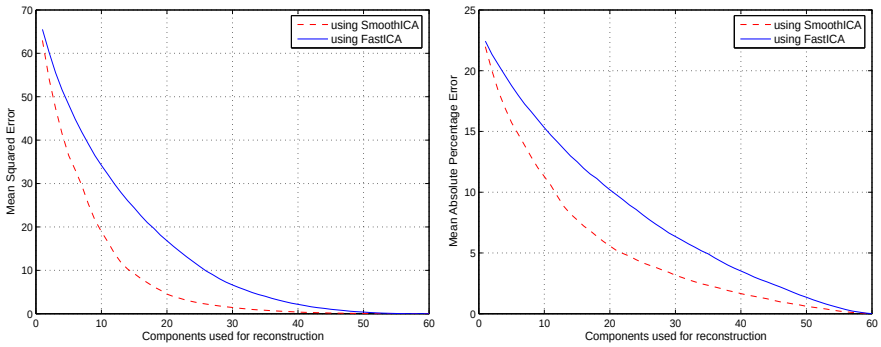


(a) For the FastICA case. (b) For the SmoothICA case.

Fig. 1. Global approximation performance. Percentage variance contributions of each source signal.

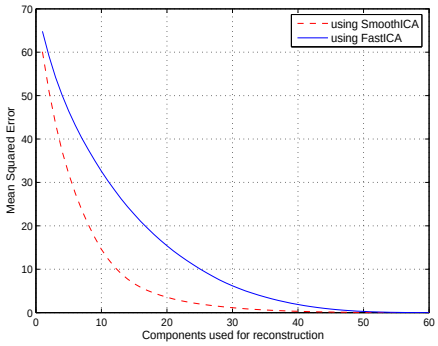
5.3 Local Approximation Comparison

The local approximation performances of both algorithms are compared using both appropriate ordering methods found in the literature. The performance has been evaluated using four error criteria calculating a mean approximation error across all securities on the portfolio. On the x-axis lie the numbers of ICs used for approximation of each security signal; from only 1 to all the ICs (60). The error criteria evaluated are Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Percentage Error (MPE) and Mean Absolute Percentage Error (MAPE). For economy of space only MSE and MAPE are presented. The former is a commonly used fitness measure penalizing large deviations from observed security prices in a greater extent, while the latter being an easily understood intuitive measure. The lists containing the ordered contributions are calculated for both algorithms, using both L_∞ norm measure and TnA heuristic algorithm. The results on Figure 2 demonstrate a clearly superior local approximation performance. Equally consistent results are obtained for the RMSE and MPE approximation measures not presented here, using both ordering methods.

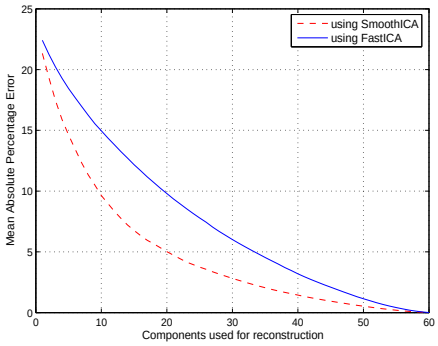


(a) Mean MSE using L_∞ norm ordering.

(b) Mean MAPE using L_∞ norm ordering.



(c) Mean MSE using TnA algorithm.



(d) Mean MAPE using TnA algorithm.

Fig. 2. Local approximation performance. SmoothICA is shown to be superior in terms of more efficient representation of each source signal.

6 Conclusions

Through the addition of the 2nd order temporal constraint, which seeks to identify temporally smooth underlying sources, the SmoothICA algorithm is more efficient that the FastICA in approximating a portfolio of securities from an appropriate subset of the estimated sources (section 5.2). This novel algorithm estimates smoother underlying sources that have an inherent ordering, as a high percentage of the portfolio's variance is contained in the first few components, compared to FastICA which has a significantly higher variance spreading among its ICs. To contain 70% of the portfolio's variance in just 12 components, while the classical FastICA requires 35, and 90% of the variance in just 21 contrasted to 49 components, is a significant improvement in terms of global approximation. In the local approximation part of this paper (section 5.3), each security in the portfolio is reconstructed by appropriate subsets of the source signals of dimensions 1 to 60. For each dimension selected the mean MSE and MAPE approximation error across the portfolio is plotted against the subset dimension. For both component ordering methods examined, the errors calculated show consistent superiority of the SmoothICA algorithm for efficient compact representation of a portfolio of securities. Furthermore, the gradients in the plots of Figure 2 support the global case conclusions.

References

1. A. Hyvarinen and E. Oja, Independent component analysis: algorithms and applications, *Neural Networks*, 13(4-5), 2000, 411–430.
2. N. Mitianoudis and T. Stathaki and A. G. Constantinides, Smooth signal extraction from instantaneous mixtures, *IEEE Signal Processing Letters*, 14(4), 2007.
3. A. Hyvarinen, Fast and robust fixed-point algorithms for independent component analysis, *IEEE Transactions on Neural Networks*, 10(3), 1999, 626–634.
4. H. Korizis and A. G. Constantinides and N. Christofides, Smooth Component Extraction from a Set of Financial Data Mixtures, *Proc. Signal Processing, Pattern recognition and Applications*, Innsbruck, Austria, 2006, 554–136.
5. A. D. Back and A. S. Weigend, A First Application of Independent Component Analysis to Extracting Structure from Stock Returns, *Int. J. Neural Systems*, 8(4), 1997, 473–484.
6. YM. Cheung and L. Xu, The MSE Reconstruction Criterion for Independent Component Ordering in ICA Time Series Analysis, *NSIP*, 8(4), 1999, 793–797.
7. ZB. Lai and YM. Cheung and L. Xu, Independent Component Ordering in ICA Analysis of Financial Data, *Computational Finance*, Chapter 14, pp. 201-212, The MIT Press, 1999.
8. YM. Cheung and L. Xu, An empirical method to select dominant independent components in ICA for time series analysis, *Proc. Int. Joint Conf. on Neural Networks '99*, Washington DC, USA, 1999, 3883–3887.
9. S. Malaroiu and K. Kiviluoto and E. Oja, Time series prediction with Independent Component Analysis, *Proc. '99 Conf. on Advanced Investment Technologies*, Gold Coast, Australia, 2000.
10. SM. Cha and LW. Chan, Applying Independent Component Analysis to Factor Model in Finance, *Proc. IDEAL '00*, Hong Kong, China, 2000, 538–544.
11. E. Chin and A. S. Weigend and H. Zimmermann, Computing portfolio risk using gaussian mixtures and independent component analysis, *Proc. IEEE/IAFE/INFORMS CIFEr '99*, New York, USA, 1999, 74–117.
12. J. Moody and L. Wu, What is the true price? State space models for high frequency FX data, *Proc. IEEE/IAFE CIFEr '97*, New York, USA, 1997.