Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v14n1p13

On the number of independent components: an adjusted coefficient of determination based approach

By Afzal, Iqbal, Afzal

Published: 20 May 2021

Electronic Journal of Applied Statistical Analysis
Vol. 14, Issue 01, May 2021, 13-27
DOI: 10.1285/i20705948v14n1p13

On the number of independent components: an adjusted coefficient of determination based approach

Saima Afzal*[a], Muhammad Mutahir Iqbal[a], and Ayesha Afzal[b]

[a]*Department of Statistics, Bahauddin Zakariya University, Multan, Pakistan*
[b]*Department of Computer Science, Air University, Multan, Pakistan*

Published: 20 May 2021

Independent Component Analysis (ICA) is a comparatively new statistical and computational technique to find hidden components from multivariate statistical data. The technique is also employed as a tool for dimension reduction for efficient data analysis. Reduction in dimensions can be done by assigning ranks to the independent components in some appropriate way and then restricting the data analysis to certain high ranking components only. The problem of determining the number of high ranked ICs that should be retained is the main objective of this paper. A method based upon adjusted coefficient of determination is proposed for the purpose. The performance of the proposed method is demonstrated through experimental evaluation on real-world financial time series data.

keywords: Adjusted coefficient of determination, dimension reduction, financial time series, independent component analysis, multidimensional data, big data analytics

1 Introduction

Researchers working in diverse domains including neurophysiology, biomedical image processing, financial time series analysis, analytical chemistry and signal processing, often have to deal with massive data sets with enormous observations and high dimensionality. In the current era of big data, conventional statistical methods are often

*Corresponding author: saimaafzalbzu@bzu.edu.pk

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

considered inefficient and impractical mainly due to the high computational complexity both in terms of computer memory requirement for data storage as well as computation time for performing analytics. The reason these approaches do not work is not just the high data volume in terms of number of observations involved, but it is the large number of variables involved. This is because statistical analysis approaches often incur exponential time complexity in terms of number of variables involved. It is, therefore, the number of variables which should be taken into account for performing analytics on massive data sets, not simply the number of observations. One important observation to address this high dimensionality problem is that not all the information contained in a data set is usually important, at least for a given situation or type of analysis. Some of the pieces may be set aside without doing any harm to the data. Thus, for multivariate data analysis, the challenge lies in selection of the relevant bits to be retained and isolating the unimportant ones (Vigneau, 2016, Reddy et al., 2020).

Independent Component Analysis (ICA) can be utilized to factorize the data into a linear combination of independent components so that the problem of high dimensions can be tackled. ICA is a computationally efficient technique belonging to the group of Blind Source Separation (BSS) methods for separating data into underlying hidden components. Use of the term 'blind' is justified as we either have no or least knowledge about source variables which we are interested to extract.

Apparently ICA is related to Principal Components Analysis (PCA). ICA is a considerably powerful technique as compared to PCA, though, capable of discovering the underlying factors or sources when PCA does not work.

The main objective of PCA is to seek an orthogonal linear transformation that maximizes the variance of the variables and to produce variables that are mutually uncorrelated, the aim of ICA is to seek for linear transformation which makes the variables statistically independent and non-Gaussian. In ICA, the variables are not ranked. The ICA can be seen as a generalization of PCA. ICs are estimated by rotating the PCs to make them as independent as possible. Hence, PCA is used as a preprocessing step in ICA.

The principle working behind the extraction of Principal Components (PCs) is to maximize the variability accounted for by these components. The PCs are not only constructed through this mechanism but are also put in descending order according to their role in explaining the variation. Thus the PCs have their in-built ordering mechanism. The PCs are easier to handle but the situation of ICs is not that simple because the criteria working for ICs i.e., maximization of non-Gaussianity or information theory measures etc. only produces the components that are independent. Independence cannot be measured hence, the produced components are not put into any logical order.

ICA is employed as a dimension reduction technique. By reducing the dimensions several leading or main components are obtained and these main components disclose the major lot of information about the underlying details of the data set. If we are not interested in retaining all ICs and we need to limit their number for one reason or other then the intuitive question arises regarding number of ICs to be retained. It is a tricky question which has to be answered intelligently. In this paper, we propose a new method to determine the optimal number of independent components that should

be retained for analysis of data sets with high dimensionality. Our proposed method is based on adjusted coefficient of determination. We evaluate the effectiveness of the proposed method through experimental evaluation of financial time series data.

The rest of the paper is organized as follows. Section 2 provides a brief review of the literature concerning the techniques to determine the number of ICs to be retained in a model. Section 3 provides detailed description of our proposed method. Section 4 discusses experimental evaluation results of the proposed method in the context of financial time series data. Section 5 concludes the article and discusses future work directions.

2 Literature Review

The literature is not silent on the issue of number of ICs to retain for dimensionality reduction. Rather, several tests and information criteria have been proposed to obtain an answer. In contrast with the principal components, extraction of independent components lack any natural ordering – less informative independent components may get extracted before a more informative component or vice versa. Therefore, the techniques used for determining the number of significant principal components can not be employed for ICA except for the use of permutation testing (Vitale et al., 2017).

In the literature about dynamic factor models, several tests and information criteria have been proposed to determine the number of ICs to be retained.

Karhunen et al. (1997) established that two famous information-theoretic criteria, Akaike Information Criteria (AIC) and Minimum Description Length (MDL) produces good estimates of the number of sources for noisy mixtures under certain conditions. Roberts (1998) also proposed a method based on Bayesian criteria.

In a later work, Lee (2003) introduced the Simulated Ordered Negentropy of ICs (SONIC) method to determine the optimal number of ICs. SONIC method is based on the Gap statistic proposed by Tibshirani et al. (2001) to estimate the number of clusters in a multivariate data set. The expected value of negentropy is compared with its value of estimated ICs in the method. The Gaussian variable and measured variable both having the same mean and standard deviation provide a way to find negentropy by using their difference. Thus, this will be essentially a non-negative quantity.

As ICA models can be based upon any number of ICs. Thus their comparison may be required to assess the quality of a particular model. One such method for comparing ICA models is given by Westad and Kermit (2003) who proposed an uncertainty parameter given below.

$$s^2(s_a) = [\sum_{m=1}^{M} (s_a - s_{a(-m)})^2][\frac{M-1}{m}] \tag{1}$$

Here $s^2(s_a)$ is the estimated uncertainty variance of the a^{th} ICA loading, M is the number of cross-validation segments, s_a is the a^{th} loading vector of the ICA model built with all objects, while $s_{a(-m)}$ corresponds to the a^{th} loading vector from the model built after removal of cross-validation segment m.

In another related work Wang et al. (2006) builds ICA models in an incremental manner; first model takes only one IC, the next takes two ICs and the last takes all ICs (A)). These consecutive ICA models built iteratively are then used to reconstruct the X matrix with each of them, yielding X_a when ICA model is built with "a" ICs where $1 \leq a \leq A$, to compute a Residual Sum of Squares (RSS). RSS is essentially the difference between the original and the reconstructed X for each model. The model corresponding to a minimum RSS is considered optimal in this method.

Several other approaches have been proposed in literature for the determination of the optimal number of ICs each having its particular area of suitability. For example, the SONIC method is applicable only when expected value of negentropy is already known. This requirement limits SONIC's application to very few situations and makes it irrelevant in most of the practical situations.

Bouveresse et al. (2012) proposed two generalized methods for finding the optimal number of ICs that only on the data set characteristics and the extracted vectors. The first method is based on an intelligent division of the data matrix into blocks of approximately equal size with intent to make the block as representative of the data as possible. If n is the expected number of ICs to be ultimately retained and m is the block size and if the maximum number of ICs derived from each block is p, i.e. $p \leq m$ then $p \geq n$. Thus each block would have p models with $1, 2, \ldots, p$ ICs involved in. Models corresponding to $1, 2, \ldots, p$ ICs for every block would have different IC loadings. The models with same number of ICs in, coming from different blocks, attempt to represent the true data. The correlation among the specific ICs from each of the blocks with same number of ICs involved may be checked. The number of ICs in models with higher such correlation would suggest the optimum number of overall ICs to be retained. In their second method, Bouveresse, et al. makes use of Durbin Watson (DW) statistic to gauge the noise ratio in signals. Here, DW statistic is defined as follows:

$DW = \frac{\sum_{t=2}^{n}(s_t - s_{t-1})^2}{\sum_{t=2}^{n} s_t^2}$, where s_t and s_{t-1} represent ICs at time t and $t-1$ respectively.

As in the case of no noise, the (sum of the) differences in the numerator $s_t - s_{t-1}$ and consequently DW will approach 0 and will depart from 0 otherwise. If p is the maximum number of ICs possible, then p number of IC models are constructed based upon $1, 2, \ldots, p$ ICs and DW statistic is computed for each of the model. The count one less than the number of ICs used in the model producing the smallest value of DW statistic would be considered as the optimum number of ICs to be retained.

Kairov et al. (2017) ranked ICs on the basis of their stability in several runs of computing ICs and then selected an optimal number of components which is consistent with the point of the qualitative change of the stability profile.

In a recent work, Kassouf et al. (2018) proposed three novel techniques to determine the number of ICs to be retained. The first technique namely Random-ICA, is a generalized form of the ICA-by-blocks approach. It splits the original data matrix into two blocks, and then repeat this process numerous times. The second method i.e. KMO-ICA-Residuals was proposed on the basis of Kaiser-Meyer-Olkin (KMO) index of the transpose of residual matrices computed from estimated ICs. ICA-corr-y, the third method selects the optimum number of ICs by the correlations among computed pro-

portions and known information about physio-chemical properties of samples, or among a source signal in the mixture and the signals computed by ICA.

The adjusted coefficient of determination $\bar{R}^2$ is obtained by making some changes to common statistical measure R^2 which summarizes how close the original data points are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression. R^2 is defined as the percentage of the variation of response variable that is explained by a linear model, or

$$R^2 = \frac{ESS}{TSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} \tag{2}$$

where,

$ESS =$ Explained Sum of Squares or sum of squares due to regression (a quantity used to describe how well a model represents the data being fitted).

$TSS =$ Total Sum of Squares (measures how much variation there is in the observed data).

$RSS =$ Residual Sum of Squares (measures the variation in the error between the observed data and fitted values).

Therefore, we can safely say that RSS and R^2 depict the same thing. The main problems with R^2 include the following:

i. Every time when a predictor is added to a model, the R^2 increases, even if it is due to the chance alone. It never decreases. Consequently, a model with more terms may come out to have a better fit simply because it has more terms.

ii. In a model with too many predictors and higher order polynomials, R^2 starts modeling the random noise in the data. This is the condition of overfitting the model and it provides the misleadingly high R^2 values and less ability to make predictions.

The adjusted coefficient of determination i.e., $\bar{R}^2$ compares the explanatory power of regression models that contains different numbers of predictors. The adjusted coefficient of determination is defined as,

$$\bar{R}^2 = 1 - (1 - R^2)\left[\frac{n-1}{n-(p+1)}\right] \tag{3}$$

where,
$n =$ number of observations
$p =$ number of variables

It is a modified version of R^2 that has been adjusted for the number of predictors in the model. $\bar{R}^2$ increases only if the new term improves the model more than it would be expected by chance. It decreases when a predictor improves the model by less than

expected by chance. In a sense $\bar{R}^2$ induces a toxic effect for too many (undue) terms by getting itself decreased as soon as the number of terms crosses a threshold so it provides a sort of protection against over fitting. In a related recent work, Mahmood et al. (2020) also proposed several adjusted coefficients of determination for beta regression models. We have also employed adjusted coefficients of determination for deciding upon the number of ICs to be retained in the model.

3 Proposed Approach to Determine the Number of ICs

After comparing RSS, R^2 and $\bar{R}^2$, we propose a new approach to decide upon the number of ICs to retain in the final model. Our approach relies on adjusted coefficient of determination ($\bar{R}^2$). Below we discuss our proposed approach in detail.

In our proposed approach, the original series is used as original data points, reconstructed series as fitted points, and their difference as error. We need to summarize how close the original and reconstructed series are. $\bar{R}^2$ is used for the purpose where different reconstructions are compared using this summary. The step-by-step procedural details are presented below:

- Compute the ICs for the n series (variables) with some suitable algorithm like JADE (Cardoso and Souloumiac, 1993), SOBI (Belouchrani et al., 1997), FastICA (Hyvärinen and Oja, 1997; Hyvarinen, 1999) etc. The algorithms for computing ICs are briefly discussed in Section 3.1.

- Arrange the ICs in an appropriate order using any of available ordering method e.g., L_∞ norm used by Back and Weigend (1997), and regression-based ordering method proposed by Afzal and Iqbal (2016) etc.

- Reconstruct the series using the procedure proposed by Back and Weigend (1997). In the absence of any final method to determine the number of ICs, do the reconstruction of each of the series at different arbitrary threshold levels (say p).

- Use the original series as original data points and reconstructed series as fitted points and their difference as error (We need to summarize how close the original and reconstructed series are.) Compare each of the reconstructed series with original series and compute adjusted coefficient of determination ($\bar{R}^2$).

- For each of the given series, p values of $\bar{R}^2$s are available. Plot these values against number of ICs used for the reconstruction.

- Gauge the cost effectiveness of additional ICs to be included in reconstruction process by checking corresponding marginal improvement in the respective value of $\bar{R}^2$. If inclusion of further ICs is not contributing enough in the improvement of the value of $\bar{R}^2$ then stop further inclusion.

- After obtaining a suitable value for the number of ICs to be retained for each series, obtain an average as a cut-off point as number of ICs to retained.

3.1 Computing ICs

As discussed above, ICs can be computed by using any of the standard ICA algorithms as discussed below.

Joint Approximation Diagonalization of Eigen-matrices (JADE) algorithm
JADE algorithm (Cardoso and Souloumiac, 1993) is a higher order statistic based algorithm which employs the blind source separation technique. The principle working behind this algorithm relates to the diagonalization fourth-order tensor i.e., the cumulant tensor. The main advantage of JADE is its computational efficiency for low-dimensional data sets. It does not consider the temporal structure of data.

FastICA Algorithm
FastICA is also a higher order statistic based algorithm (Hyvärinen and Oja, 1997; Hyvarinen, 1999). It looks for an orthogonal rotation of whitened data using a fixed-point iteration scheme, which maximizes a non-Gaussianity measure of the rotated components. Such algorithms are made robust to additive Gaussian noise using this approach. It also does not consider the temporal structure of data.

Second Order Blind Identification (SOBI) Algorithm
SOBI is a second-order statistic based algorithm (Belouchrani et al., 1997). The decomposition in such algorithms achieves the decomposition through information of time-frequency. In this algorithm, the Blind Source Separation (BSS) is performed through the joint diagonalization of time delayed covariance matrices.

4 Evaluation using Financial Data

4.1 Dataset Description

The data of daily closing rates of 161 companies listed in Karachi Stock Exchange of Pakistan over the period of June 11, 2004 to February 15, 2012 has been used for the analysis performed in this research. Appendix A provides the list of selected companies. A matrix of 161 columns denoting the companies and 2004 rows denoting their respective closing rates has been constructed for the analysis. First, the non-stationary stock prices x_{it} ($i = 1, 2, \ldots, 161$ and $t = 0, 1, \ldots, 2003$) are transformed to stock returns defined ahead.

The rates of change of a price series are called *returns*. The price series do not typically fluctuate around a constant level but the returns series often looks stationary. Thus, in many applications the returns series are used instead of price series.

Let $X = x_{it}$ denote the matrix of closing rates of 161 companies at 2004 time points. $\mathbf{x}_i$, a row vector of order 2004 denote the closing rates for i^{th} company at 2004 time points. Let y_{it} be the i^{th} transformed series where $i = 1, 2, \ldots, 161$ and $t = 1, \ldots 200$. Then,

$$y_{it} = ln[x_{it}] - ln[x_{i(t-1)}] \tag{4}$$

This is the first difference of the log price series, and is sometimes also called the *log return*. The matrix of transformed series (returns) is given by Y.

4.2 Construction of ICs and Weighted ICs

Before the computation of ICs, the kurtoses of the 161 series are computed to test the normality of each individual series. The results show that the data is far away from normality. Since the dataset is non-Gaussian distributed, it is reasonable to apply ICA to extract the interesting features from the data.

There are numerous approaches to estimate ICs and unmixing matrix such as maximization of non-Gaussianity, information theoretic measures, maximum likelihood estimation method and tensor based methods. Various algorithms have been introduced in the literature to construct ICs based upon any one or a blend of several of the above-mentioned approaches. We selected three of the existing algorithms namely JADE (Cardoso and Souloumiac, 1993), SOBI (Belouchrani et al., 1997), and FastICA (Hyvärinen and Oja, 1997 and Hyvarinen, 1999) to compute ICs in the current research. The choice is in line with the existing applications available in the literature as employed by Back and Weigend, 1997, Prieto, 2011, Jianwei et al., 2019, and Miao et al., 2020.

The matrix of estimated ICs is,

$$S = s_{it} \tag{5}$$

and the estimated mixing matrix is given as,

$$A = a_{ik} \tag{6}$$

The matrix of the estimated ICs with rows ordered according to "Regression Based Ordering of ICs" proposed by Afzal and Iqbal (2016) is given below:

$$S^{'} = s^{'}_{it} \tag{7}$$

The mixing matrix with ordered rows is given as,

$$A^{'} = a^{'}_{ik} \tag{8}$$

The calculation of weighted ICs is in line with the procedure proposed by Back and Weigend (1997). The elements of the i^{th} row of ordered mixing matrix are used as weights to compute weighted ICs for i^{th} company. The matrix of weighted ICs for i^{th} company is thus,

$$W_i = a^{'}_{ik} s^{'}_{kt} \tag{9}$$

4.3 Reconstruction of the Series

The original series can be reconstructed without loss if all ICs are used in the process. However, this practice is useless as it does not add any value to the analysis. Number of ICs used in the reconstruction should necessarily be lower than the total (161 in this case). In the absence of any final method of deciding number of ICs to be retained and used in the reconstruction, it is done at nine arbitrary cut-off points as 10(10)90 percent that is 16(16)144. For possible generalization, if the number of retained ICs is denoted

by 'l' then possible values of l at nine retention levels will be $16, 32, \ldots, 144$. At 10% retention level i.e., at $l = 16$, the first 16 elements of the t^{th} column of weighted ICs are cumulated to form Cumulative IC at time t for company i. Thus for the i^{th} company, at retention level l, and at time t, the i^{th} cumulative IC is computed as:

$$\hat{Y}_{it} = \sum_{k=1} l a_{ik}^{'} s_{kt}^{'} \tag{10}$$

Let $\hat{X}$ be the matrix of reconstructed series of closing rates of 161 companies of KSE each having 2004 observations then for a given retention level and at time t, the reconstructed series of closing rates is given by:

$$\hat{X}_{it} = \hat{x}_{i(t-1)} \text{antilog}(\hat{y}_i t) \tag{11}$$

$\hat{x}_{i0}$ is required to proceed any further and is borrowed as the starting point from the original series. For example, when we talk about the company, ABOT ($i = 1$) it is 140.35.

4.4 Deciding the Number of ICs to be Retained

The reconstruction of original series is done in nine different ways i.e., using 10, 20, 30, 40, 50, 60, 70, 80 and 90 percent of ICs for the purpose. Each of the reconstructed series is then compared with the original series and $\bar{R}^2$ is calculated. Thus, nine values of $\bar{R}^2$ for every company are obtained. For a given company, say ABOT, nine values of $\bar{R}^2$s are available. These values are plotted against number of ICs used for the reconstruction. The procedure is replicated for each of construction algorithm namely JADE, FastICA, and SOBI. We analyzed $161 \times 3 \times 9 = 4347$ $\bar{R}^2$s in $161 \times 3 = 483$ layouts. One such layout is presented in Figure 1 where reconstruction through JADE algorithm is discussed. Two more layouts are shown in Figures 2 and 3 presenting similar results for FastICA and SOBI algorithms.

As discussed earlier, the cost effectiveness of additional ICs to be included in reconstruction process is gauged by checking corresponding marginal improvement in the respective value of $\bar{R}^2$. If inclusion of further ICs is not contributing enough in the improvement in the value of $\bar{R}^2$ then further inclusion may be stopped. Figures 1 to 3 can be observed to visualize the impact at a glance.

Figure 1 shows that for company ABOT the cut-off point falls near 50%. The cut-off points obtained in this manner are summarized as Table 1.

Table 1 gives an average for JADE algorithm close to 55%, for FastICA approximately 54%, and for SOBI it is about 60%. Overall average of three algorithms for six companies portrayed above is 56%. The overall average for all 161 companies for three algorithms gives a percentage close to 50. So a cut-off point is chosen at this percentage namely use 80 ICs for reconstruction purposes gives satisfactory reconstruction. Figure 4 illustrates the performance of the reconstructed series (red) in comparison with the original series (black), for ABOT Compancy with fifty percent reduction in the number of ICs using the proposed method. It can be seen that the reconstructed series in case of JADE and

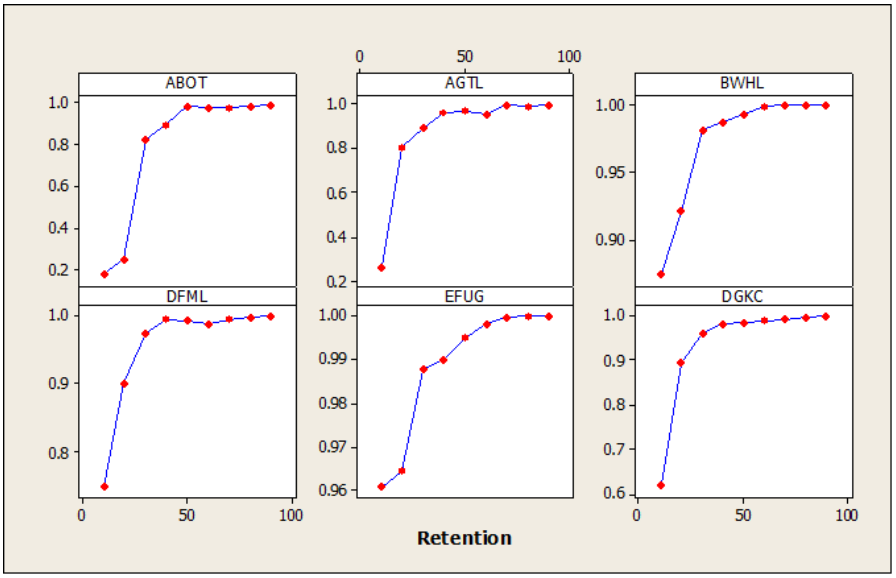

Figure 1: $\bar{R}^2$ as a function of Retention Level (Case of JADE)

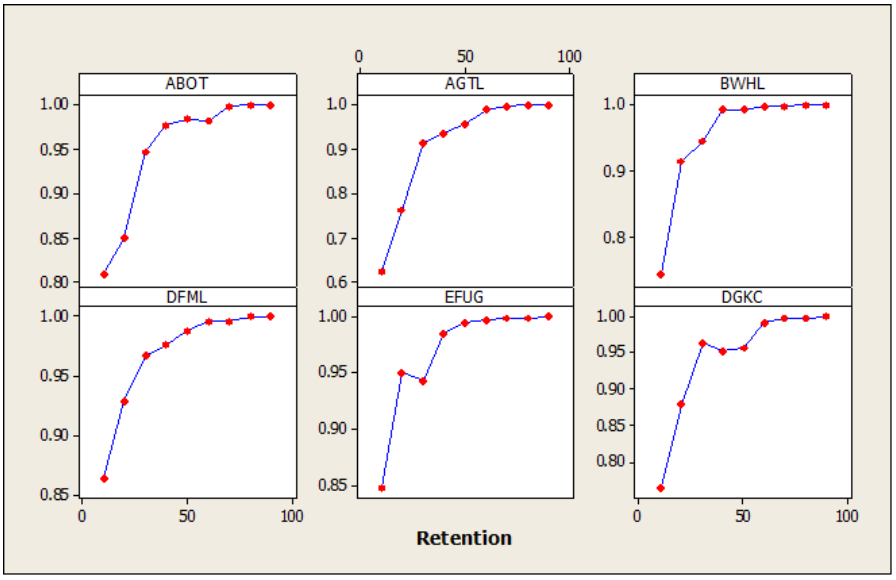

Figure 2: $\bar{R}^2$ as a function of Retention Level (Case of FastICA)

FastICA algorithms are very close to the original series. However, the performance of SOBI algorithm is not so well.

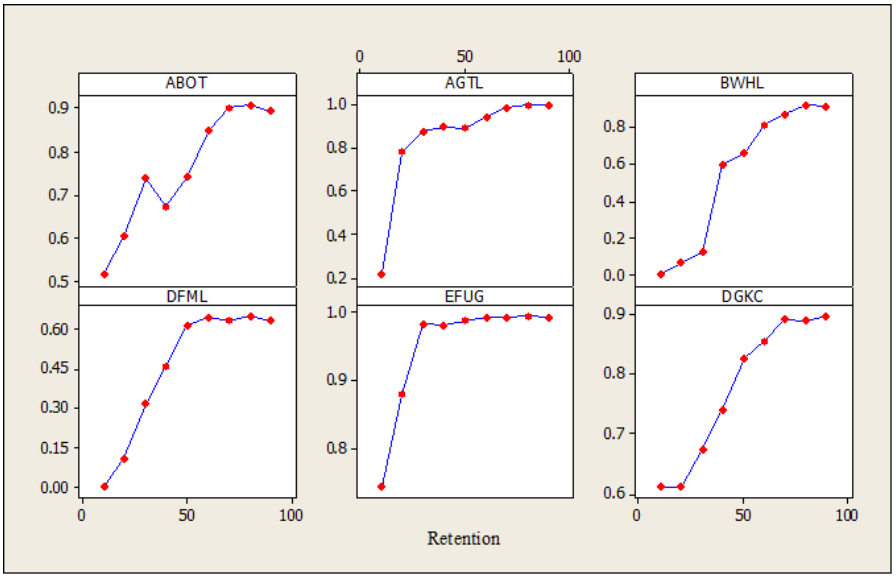

Figure 3: $\bar{R}^2$ as a function of Retention Level (Case of SOBI)

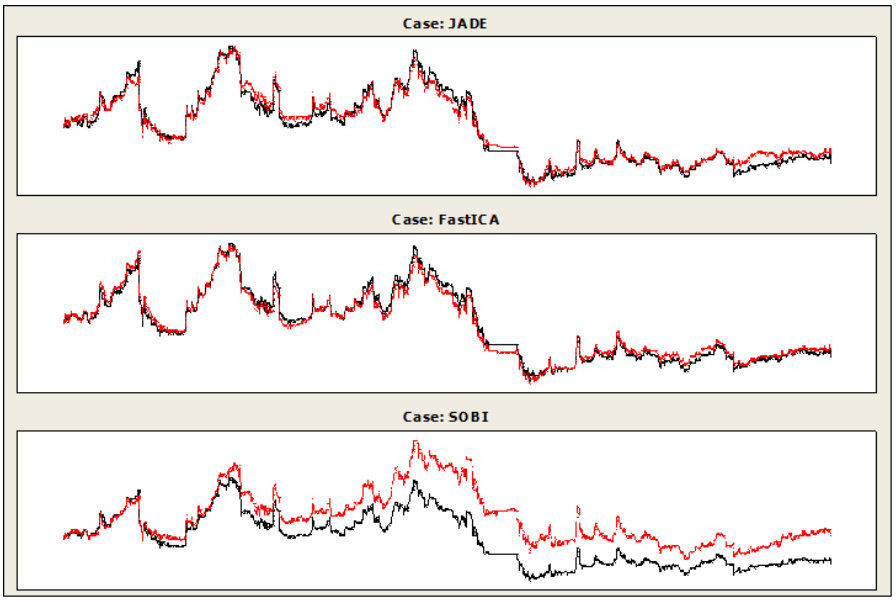

Figure 4: Original (Black) and Reconstructed (Red) Series for Abbot Laboratories with Fifty Percent Reduction in Number of ICs.

Table 1: Retention Levels for Six Randomly Selected Companies Using JADE, FastICA and SOBI Algorithm.

Company	Retention Levels		
	JADE	FastICA	SOBI
ABOT	50%	70%	70%
AGTL	70%	60%	70%
BWHL	30%	40%	70%
DFML	60%	60%	50%
EFUG	60%	40%	30%
DGKC	40%	60%	70%

5 Conclusion

In this work, a new method based upon $\bar{R}^2$ is proposed to determine the optimal number of ICs to be retained in a final model for performing data analytics. To validate the effectiveness of the proposed method, we evaluated its performance through experimentation on financial time series data set. We performed a decomposition of 161 stock returns into statistically independent components. Our proposed approach significantly reduced the computational complexity of the data set by allowing fifty percent reduction in the number of ICs to retain for analysis. Though the results are very encouraging, a room for further investigation still remains there.

In our future work, we intend to investigate further into this problem by introducing a statistical test procedure to check the significance, suitability, and adequacy of the retained ICs.

References

Afzal, S. and Iqbal, M. M. (2016). A new way to order independent components. *Journal of Applied Statistics*, 43(9):1753–1764.

Back, A. D. and Weigend, A. S. (1997). A first application of independent component analysis to extracting structure from stock returns. *International journal of neural systems*, 8(04):473–484.

Belouchrani, A., Abed-Meraim, K., Cardoso, J.-F., and Moulines, E. (1997). A blind source separation technique using second-order statistics. *IEEE Transactions on signal processing*, 45(2):434–444.

Bouveresse, D. J.-R., Moya-González, A., Ammari, F., and Rutledge, D. N. (2012). Two novel methods for the determination of the number of components in indepen-

dent components analysis models. *Chemometrics and Intelligent Laboratory Systems*, 112:24–32.

Cardoso, J.-F. and Souloumiac, A. (1993). Blind beamforming for non-gaussian signals. In *IEE proceedings F (radar and signal processing)*, volume 140, pages 362–370. IET.

Hyvarinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE transactions on Neural Networks*, 10(3):626–634.

Hyvärinen, A. and Oja, E. (1997). A fast fixed-point algorithm for independent component analysis. *Neural computation*, 9(7):1483–1492.

Jianwei, E., Ye, J., and Jin, H. (2019). A novel hybrid model on the prediction of time series and its application for the gold price analysis and forecasting. *Physica A: Statistical Mechanics and Its Applications*, 527:121454.

Kairov, U., Cantini, L., Greco, A., Molkenov, A., Czerwinska, U., Barillot, E., and Zinovyev, A. (2017). Determining the optimal number of independent components for reproducible transcriptomic data analysis. *BMC genomics*, 18(1):712.

Karhunen, J., Cichocki, A., Kasprzak, W., and Pajunen, P. (1997). On neural blind separation with noise suppression and redundancy reduction. *International Journal of Neural Systems*, 8(02):219–237.

Kassouf, A., Jouan-Rimbaud Bouveresse, D., and Rutledge, D. N. (2018). Determination of the optimal number of components in independent components analysis. *Talanta*, 179:538 – 545.

Lee, S.-M. (2003). *Estimating the number of independent components via the SONIC statistic*. PhD thesis, Master's thesis, University of Oxford, United Kingdom.

Mahmood, S. W., Seyala, N. N., and Algamal, Z. Y. (2020). Adjusted R2-type measures for beta regression model. *Electronic Journal of Applied Statistical Analysis*, 13(2):350–357.

Miao, F., Zhao, R., Jia, L., and Wang, X. (2020). Fault diagnosis of rotating machinery based on multi-sensor signals and median filter second-order blind identification (mf-sobi). *Applied Sciences*, 10(11):3735.

Prieto, E. G. (2011). *Independent component analysis for time series*. PhD thesis, Universidad Carlos III de Madrid.

Reddy, G. T., Reddy, M. P. K., Lakshmanna, K., Kaluri, R., Rajput, D. S., Srivastava, G., and Baker, T. (2020). Analysis of dimensionality reduction techniques on big data. *IEEE Access*, 8:54776–54788.

Roberts, S. J. (1998). Independent component analysis: source assessment and separation, a bayesian approach. *IEE Proceedings-Vision, Image and Signal Processing*, 145(3):149–154.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423.

Vigneau, E. (2016). Dimensionality reduction by clustering of variables while setting aside atypical variables. *Electronic Journal of Applied Statistical Analysis*, 9(1):134–

153.

Vitale, R., Westerhuis, J. A., Naes, T., Smilde, A. K., de Noord, O. E., and Ferrer, A. (2017). Selecting the number of factors in principal component analysis by permutation testing - Numerical and practical aspects. *Journal Of Chemometrics*, 31(12).

Wang, G., Cai, W., and Shao, X. (2006). A primary study on resolution of overlapping gc-ms signal using mean-field approach independent component analysis. *Chemometrics and intelligent laboratory systems*, 82(1-2):137–144.

Westad, F. and Kermit, M. (2003). Cross validation and uncertainty estimates in independent component analysis. *Analytica chimica acta*, 490(1-2):341–354.