Journal of Intelligent Decision and Computational Modelling

www.jidcm.reapress.com

J. Intell. Decis. Comput. Model. Vol. 1, No. 1 (2025) 50–56.

Paper Type: Original Article

A Machine Learning Approach to GDP Data Analysis Using Independent Component Analysis (ICA)

Fatemeh Asadi*

Department of Statistics, Yazd University, Yazd, Iran; asadi.f.2005@yazd.ac.ir.

Citation:

Received: 13 June 2024 *Revised: 20 September 2024* *Accepted: 25 November 2024*	Asadi, F. (2025). A machine learning approach to GDP data analysis using independent component analysis (ICA). *Journal of Intelligent Decision and Computational Modelling, 1*(1), 50-56.

Abstract

The objective function of numerous well-established Independent Component Analysis (ICA) algorithms, widely recognized as unsupervised machine learning methods, is rooted in specific dependence criteria. This study introduces a distinctive dependence criterion based on the Cumulative Distribution Function (CDF) for characterizing the independence between two random variables. Furthermore, an in-depth exploration of the inherent properties of these variables is conducted. The proposed machine learning algorithms are then applied to real-world time series data, serving as an effective pre-processing clustering method. The algorithm's performance is systematically compared with several previous machine learning-based ICA algorithms.

Keywords: Amari error, Clustering, Cumulative distribution function, Dependence criteria, Independent components analysis.

1 | Introduction

In the field of multivariate data analysis, the complexity arising from high dimensionality and intricate interdependencies among observed values often poses a significant challenge. In order to discern patterns within such data, it is frequently essential to disentangle its various aspects. Independent Component Analysis (ICA) is a widely utilized technique in multivariate statistics for this purpose. Initially proposed by Harlaut in the 1980s, ICA has gained substantial prominence as an unsupervised blind source separation method in the realm of signal processing [1].

For a more comprehensive understanding of ICA and its applications in signal processing, an in-depth exploration of the technique and its role in multivariate data analysis can be found in the see [2], [3], [4] and [5]. These resources offer a detailed overview of ICA, encompassing its theoretical foundations, algorithms, and practical applications in the context of multivariate data analysis and signal processing.

Corresponding Author: asadi.f.2005@yazd.ac.ir

https://doi.org/10.48314/jidcm.v1i1.61

ICA is a matrix factorization approach that optimizes the signals captured by each matrix factor to achieve mutual independence. This technique has found applications in various scientific fields, including medical signals, audio signals, biological assays, space imaging, financial data, climate data, and time series data [6]. These applications demonstrate the versatility and utility of ICA in various scientific fields, making it a valuable tool for data analysis and pattern recognition.

The organization of the paper is as follows: In the introduction, the paper introduces the concept of ICA and its applications in various scientific fields. Section 2 provides a comprehensive review of the ICA technique, its theoretical foundations, and its applications. The paper introduces a dependency criterion based on distribution functions and discusses its estimation in Section 3.

Then, the performance of the proposed algorithm is compared to existing ICA algorithms using the average of Amari errors through a Monte Carlo simulation study. Section 5 will focus on the application of these algorithms in time series data clustering.

Furthermore, this study delves into the application of data clustering methodologies to a set of real-time series data, specifically the Gross Domestic Product (GDP) per capita index, extracted from the statistics of 26 countries spanning from 1975 to 2020. The conclusion of the research is drawn by summarizing the results and analyzing the structures extracted, as detailed in Section 6.

2 | Independent Component Analysis

Each signal is time-varying, and a signal is represented as $s_i = (s_{1i}, s_{2i}, \dots, s_{ni})^T, i = 1,2,\dots, d$, where n is the number of time steps, s_{ij} is time j of the signal s_i, and d is the number of source signals. Given d independent source signals define a matrix $S = (s_1, s_2, \cdots, s_d)$, where $S \in \mathbb{R}^{n\times d}$ is the matrix of the source signals.

The source signals can be mixed. As a result, each source signal has a different effect on the output signals. d mixtures can be represented as $X =: SA$, where $X \in \mathbb{R}^{n\times d}$ is the matrix of the mixed signals, n is the number of mixes and $A \in \mathbb{R}^{d\times d}$ is a mixing coefficients matrix. Therefore, the mixing coefficients are used to transform a linear source signal from S into a mixedsial in $X = (x_1, x_2, \cdots, x_d)$ space as $X = SA$. The target is to extract the source signals by finding the unmixing coefficient matrix (W) used to convert the mixed signals into a set of independent signals, $X \ \rightarrow Y\ Y = XW$, where $W \in \mathbb{R}^{d\times d}$ is the matrix of unmixing coefficients, and it is the inverse of the matrix A [7].

All ICA algorithms produce an unmixing matrix W that can be applied to matrix X to recover estimates of the independent components. Many types of research exist about different types of ICA procedures and their interpretations. Most ICA algorithms minimize a contrast function that measures the degree of dependency between components. The efficiency of the ICA algorithms depends on the choice of the contrast function and the algorithm used for the implementation of the optimization problem. Matrix W can be estimated using several main independence approaches, which leads to the creation of unmixed matrices with slight differences. In all these approaches, an unmixed matrix is obtained. The data is projected on that matrix, and the independent signals are extracted; see [1], [7], and [8].

In ICA, the Maximum likelihood (ML) method is used to estimate the matrix W, which provides the best fit for the extracted Y signals; see [9], [10], and [11]. Because the Mutual Information (MI) criterion measures the independence between two random variables, many algorithms for ICA are constructed based on minimizing MI to estimate the initial signals [12]. Therefore, independent components can be obtained by minimizing MI between different components; see [1] and [7]. In this setting, the matrix W is obtained such that the MI criterion is close to zero.

3 | Proposed Dependency Criterion

It is inevitable to encounter independent components that are unknown and do not have the usual statistical distributions in ICA. The objective function of many ideas in ICA is based on the density functions. These led to a hard way of estimation because the calculation of the estimators can be included with errors. A solution to this is to use the Cumulative Distribution Functions (CDFs) in ICA algorithms. Unlike the density function estimators, the estimators based on the empirical distribution functions are very simple and fast in statistical inference.

Moreover, the empirical distribution functions converge to the theoretical CDFs almost surely. On the other hand, we know that in ICA, a dependency criterion will be helpful if it characterizes the independence between two random variables. Motivated by these facts, we will introduce a criterion based on the CDFs that characterizes the independence between two random variables if it would be equal to zero; see [5], [13], [14].

Let X_1 and X_2 be two random variables with the joint distribution function F, the marginal distribution functions F_1 and F_2, respectively. We define a dependency criterion based on CDFs, denoted by GDC as the following form:

$$\mathrm{GDC}(X_1, X_2) = \iint \left(\cosh\left(\frac{F(x_1, x_2)}{F_1(x_1)F_2(x_2)} - 1\right) - 1\right) dF(x_1, x_2). \tag{1}$$

These divergences vanish if and only if the random variables X_1 and X_2 are independent.

Theorem 1.

$\mathrm{GDC}(X_1, X_2) = 0$, if and only if X_1 and X_2 are independent.

$\mathrm{GDC}(X_1, X_2) = \mathrm{GDC}(X_2, X_1)$.

Proof: The results hold.

3 | Estimate Dependency Criterion

Let $X_{n\times 2}$ be the observation matrix from the random vector (X_1, X_2) with the joint distribution function F and the marginal CDFs F_1 and F_2, respectively. Also, suppose that $x_{ij} = [X]_{ij}, i = 1,2,\dots, n, j = 1,2$. Functions $\hat{F}_1(x_1)$ and $\hat{F}_2(x_2)$ are the marginal empirical distribution functions and $\hat{F}(x_1, x_2)$ will be the joint empirical distribution function. Now, we can estimate GDC as

$$\widehat{\mathrm{GDC}}(X_1, X_2) = \iint \left(\cosh\left(\frac{\hat{F}(x_1, x_2)}{\hat{F}_1(x_1)\hat{F}_2(x_2)} - 1\right) - 1\right) d\hat{F}(x_1, x_2). \tag{2}$$

4 | The Proposed Independent Component Analysis Algorithm Based on $\widehat{\mathrm{GDC}}$

In this section, the general process of the new algorithm for ICA is explained. The algorithm uses the estimator $\widehat{\mathrm{GDC}}$ for GDC, and it is called GDCICA. The general process of the algorithm for a 2-dimensional case is summarized as follows:

Algorithm 1

Input: A matrix $X \in R^{n\times 2}$ where the rows are mixed-centered components.

Make the matrix X white, so that $Y = X \times Z'$, where Z is a whitening matrix.

Define $\widehat{\mathrm{GDC}}(S(\theta))$ as the function of θ, where $S(\theta) = Y \times R(\theta)$, such that $R(\theta)$ is an orthogonal rotation matrix as

$$R(\theta) = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}.$$

Minimize the function $\mathrm{GDC}(X_1, X_2)$ over $\theta \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ and set $\theta_0 = \mathrm{argmin}_\theta\, \widehat{\mathrm{GDC}}(S(\theta))$.

Output: Unmixing $\widehat{W} = R'(\theta_0) \times Z$, and matrix of source signal estimates $\hat{S} = Y \times R(\theta_0)$ [16].

5 | Simulation Study

In this section, we utilize the Monte-Carlo simulation to compare the performance of GDCICA with FastICA [15], Infomax [16], JADE [17], RADICAL [18], HICA [19], and RLICA [20] using R software. We use 18 different one-dimensional densities provided by [23]. For comparison, we applied the Amari error introduced by [21]. We obtained the average of the Amari errors of the methods FastICA, Infomax, JADE, RADICAL, HICA, RLICA, and GDCICA by 120 replications of their corresponding algorithms.

The averages of errors for the n=1000 sample are reported in *Table* 1. In this table, it is observed that in terms of Amari error, GDCICA outperforms its competitors in 8 out of 18 cases and is better than FastICA, Infomax, JADE, RADICAL, HICA, and RLICA methods. The GDCICA applied to data cases with nonsymmetric distributions labeled as e, j, k, p, and q has yielded the best performance. The algorithm with negative kurtosis distributions labeled as g, k, p, and q yields the best performance. Also, multimodal distributions labeled as f, g, j, and p exhibited the lowest Amari errors.

Also, we calculated the average of Amari errors on all distributions (a) to (r) and provided them in the last row of *Table 1*. We see that GDCICA is the best technique regarding the average of the Amari errors in all distributions. In addition, we randomly selected two distributions among all distributions (a) to (r) and calculated the averages of Amari errors for them, which are reported in the last line of *Table 1*. The results showed that in the choice of random distributions, in terms of Amari error, GDCICA was the best.

Table 1- Averages of Amari errors for d=2 and n=1000 samples. The smallest entry of our functions in each row is boldfaced.

Distributions	Fast ICA	Infomax	JADE	RADICAL	HICA	RLICA	GDCICA
a	0.5401	0.5402	0.5339	0.5440	0.5239	0.5340	0.5514
b	0.5237	0.5237	0.5350	0.5442	0.5234	0.5268	0.5213
c	0.5196	0.5196	0.5203	0.5245	0.5213	0.5225	0.5295
d	0.5443	0.5445	0.5594	0.5537	0.5383	0.5427	0.5737
e	0.5435	0.5397	0.5060	0.5092	0.5312	0.5397	0.5438
f	0.5340	0.5340	0.5378	0.5400	0.5454	0.5448	0.5172
g	0.4781	0.4781	0.4782	0.4767	0.4784	0.4807	0.4716
h	0.5458	0.5458	0.5500	0.5352	0.5184	0.5417	0.5304
I	0.5128	0.5126	0.5076	0.5154	0.5183	0.4989	0.5265
j	0.5160	0.5272	0.5064	0.5080	0.5149	0.5090	0.4953
k	0.5316	0.5315	0.5303	0.5413	0.5436	0.5429	0.5171
L	0.5083	0.5075	0.5137	0.5073	0.5096	0.5018	0.5243
M	0.5076	0.5076	0.5127	0.5014	0.5325	0.5160	0.5089
N	0.5360	0.5365	0.5365	0.5215	0.5477	0.5262	0.5457
O	0.4784	0.4784	0.4878	0.5011	0.4981	0.4969	0.4783
P	0.5475	0.5472	0.5498	0.5554	0.5461	0.5468	0.5337
q	0.5342	0.5331	0.5337	0.5479	0.5452	0.5492	0.5108
R	0.5192	0.5192	0.5156	0.5308	0.5183	0.5100	0.5151
Average	0.5234	0.5237	0.5230	0.5254	0.5253	0.5239	0.5219
Rand	0.5151	0.5171	0.5153	0.5184	0.5185	0.5167	0.5080

6 | Real Data

Time series clustering has been practical and often attractive in producing beneficial information in various domains. Since ICA naturally takes the temporal dependency into account through its underlying model when decomposing variables, ICA gives the opportunity to generate independent components quickly and then group them based on the temporal dependence among them [22]. ICA is used as a data pre-processing step to improve the clustering of temporal data.

Using ICA not only provides the possibility of clustering of time series but also supplies information on common characteristics. This technique is a multi-stage procedure that uses the GDCICA algorithm on real data as pre-processing in time series clustering. The mixing matrix coefficients obtained by GDCICA can be exerted as an input variable in a clustering method. Here, we provide the time series clustering using the GDCICA algorithm as a pre-processing, and then we apply the PAM algorithm (K-medoids) introduced by [26] to the final clustering. After estimating the mixing matrix using our algorithm, we apply the PAM clustering algorithm on the unmixing matrix to determine the suitable number of clusters and to select the best clustering in terms of the Silhouette criterion using the NbClustR package. Therefore, the best clustering is reported as the final result.

We apply the GDCICA algorithm on a batch of time series to prepare them for clustering using R software. We consider the GDP per capita time series for 27 countries from 1975 to 2020. (Australia, Austria, Belgium, Burkina Faso, Cameroon, Canada, Chile, China, Denmark, Finland, France, Germany, Ghana, India, Indonesia, Italy, Japan, South Korea, Malaysia, Pakistan, Qatar, Saudi Arabia, Singapore, Sweden, Turkey, the United Arab Emirates, and the United States).

The data comes from the World Bank organization. We aim to cluster countries by their GDP per capita time series over the past 46 years. To do this, the primary GDP per capita data were standardized. Then, the GDCICA algorithm was used on the principal components, and the coefficients of the mixing matrix in GDP per capita standardized data were obtained; then, they were applied as input to the PAM clustering algorithm. Though these time series data follow various distributions, the proposed GDCICA algorithms can extract the sources well for clustering them.

To check the efficiency of this method, time series plots in 5 clusters are drawn in *Fig. 1*. Based on trend plots obtained from all clusters in standardized GDP per capita, the pre-processing technique done by the proposed algorithm for clustering countries worked very well because is intuitively clear that countries in the identical cluster have similar trends in standardized GDP per capita.

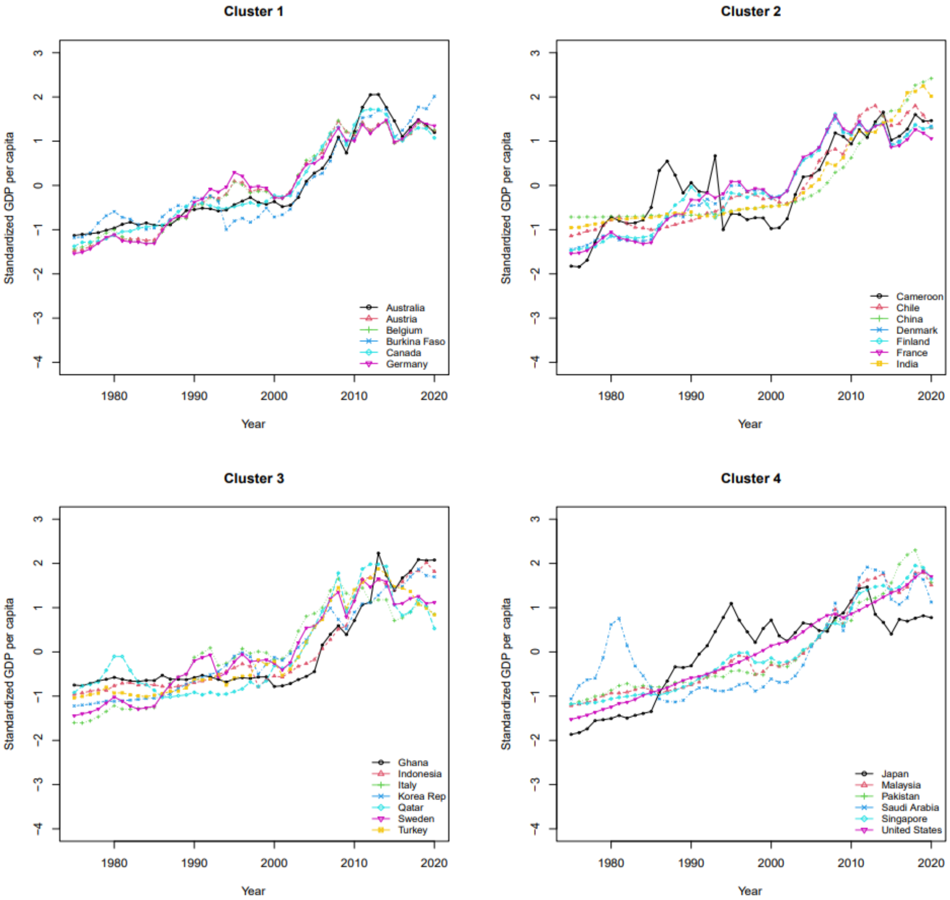

Fig. 1. Trend plots of standardized GDP per capita time series in 4 clusters.

7 | Conclusion

The objective function of many well-known ICA algorithms is based on density functions. Their estimation makes the algorithms proceed longer, and the implementation is time-consuming. To solve this problem, it was proposed to use dependence criteria for two continuous random variables X_1 and X_2 based on the CDFs. Then, we presented an algorithm for the ICA problem based on this dependence criterion, called GDCICA.

To evaluate the performance, the proposed algorithm was compared to some of the ICA algorithms by the Monte Carlo simulation. Results showed that in most cases, the proposed algorithm attains better performance than the usual ICA algorithms in different classes of distributions in terms of the average of the Amari errors. Also, the GDCICA algorithm, in nonsymmetric, multimodal distributions, and the negative kurtosis class exhibited the lowest Amari errors.

We used a batch of time series involving GDP per capita for 27 countries over the past 46 years for clustering. The GDCICA algorithm, as a pre-processing, was used on the principal components, and the coefficients of the mixing matrix were obtained; then they were applied as input to the PAM clustering algorithm. The results obtained from the clustering of time series showed that the pre-processing technique by the proposed algorithm can be advantageous in suitable clustering of different data that follow different distributions.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability

All data are included in the text.

Funding

This research received no specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] Pfister, N., Weichwald, S., Bühlmann, P., & Schölkopf, B. (2019). Robustifying independent component analysis by adjusting for group-wise stationary noise. *Journal of machine learning research*, *20*(147), 1–50. https://jmlr.org/papers/v20/18-399.html

[2] Hyvarinen, A., Karhunen, J., & Oja, E. (2002). Independent component analysis. *Studies in informatics and control*, *11*(2), 205–207. https://www.cs.helsinki.fi/u/ahyvarin/presentations/psykointroICA.pdf

[3] Comon, P. (1994). Independent component analysis, A new concept? *Signal processing*, *36*(3), 287–314. https://doi.org/10.1016/0165-1684(94)90029-9

[4] Comon, P., & Jutten, C. (2010). *Handbook of blind source separation: Independent component analysis and applications*. Academic press. https://www.amazon.com/Handbook-Blind-Source-Separation-Applications/dp/0123747260

[5] Asadi, F., Torabi, H., & Nadeb, H. (2025). A new approach for independent component analysis and its application for clustering the economic data. *International journal of computational economics and econometrics*, *15*(1–2), 147–171. https://doi.org/10.1504/IJCEE.2025.145019

[6] Sompairac, N., Nazarov, P. V, Czerwinska, U., Cantini, L., Biton, A., Molkenov, A. (2019). Independent component analysis for unraveling the complexity of cancer omics datasets. *International journal of molecular sciences*, *20*(18), 4414. https://www.mdpi.com/1422-0067/20/18/4414

[7] Tharwat, A. (2021). Independent component analysis: An introduction. *Applied computing and informatics*, *17*(2), 222–249. https://doi.org/10.1016/j.aci.2018.08.006

[8] Stone, J. V. (2004). Independent component analysis: A tutorial introduction. https://books.google.com/books?id=P0rROE-WFCwC&printsec=frontcover

[9] Gelle, G., Colas, M., & Serviere, C. (2001). Blind source separation: A tool for rotating machine monitoring by vibrations analysis? *Journal of sound and vibration*, *248*(5), 865-885. https://doi.org/10.1006/jsvi.2001.3819

[10] Pham, D. T. (1992). Separation of a mixture of independent sources through a maximum likelihood approach. *EUSIPCO'92 brussels, belgium*, 771–774. https://doi.org/10.1109/78.599941

[11] Pearlmutter, B., & Parra, L. (1996). Maximum likelihood blind source separation: A context-sensitive generalization of ICA. *Advances in neural information processing systems*, *9*. https://proceedings.neurips.cc/paper/1996/hash/dabd8d2ce74e782c65a973ef76fd540b-Abstract.html

[12] Langlois, D., Chartier, S., & Gosselin, D. (2010). An introduction to independent component analysis: Infomax and fast ICA algorithms. *Tutorials in quantitative methods for psychology*, *6*(1), 31–38. http://dx.doi.org/10.20982/tqmp.06.1.p031

[13] Asadi, F., & Torabi, H. (1402). Clustering of unemployment rate data using a new independent component analysis algorithm. *The national conference on analysis has given firsts.* Yasooj, Iran, Civilica. (**In Persian**). https://civilica.com/doc/1672257

[14] Asadi, F., Torabi, H., & Nadeb, H. (2025). An algorithm for independent component analysis using a general class of copula-based dependence criteria. *Journal of mahani mathematical research*, *14*(1), 527–550. https://doi.org/10.22103/jmmr.2024.23031.1591

[15] Hyvarinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE transactions on neural networks*, *10*(3), 626–634. https://doi.org/10.1109/72.761722

[16] Lee, T.W., Girolami, M., & Sejnowski, T. (1999). Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. *neural computation*, *11*, 417–441. https://doi.org/10.1162/089976699300016719

[17] Cardoso, J. F. (1999). High-order contrasts for independent component analysis. *Neural computation*, *11*(1), 157–192. https://doi.org/10.1162/089976699300016863

[18] Learned-Miller, E. G., & others. (2003). ICA using spacings estimates of entropy. *Journal of machine learning research*, *4*(12), 1271–1295. https://b2n.ir/zs2583

[19] Rahmanishamsi, J., Dolati, A., & Aghabozorgi, M. R. (2018). A copula based ICA algorithm and its application to time series clustering. *Journal of classification*, *35*, 230–249. https://link.springer.com/article/10.1007/s00357-018-9258-x

[20] Rahmani Shamsi, J., & Dolati, A. (2018). Rank based least-squares independent component analysis. *Journal of statistical research of Iran (JSRI)*, *14*(2), 247–266. https://b2n.ir/uh8615

[21] Amari, S., Cichocki, A., & Yang, H. (1995). A new learning algorithm for blind signal separation. *Advances in neural information processing systems*, *8*. https://proceedings.neurips.cc/paper/1995/hash/e19347e1c3ca0c0b97de5fb3b690855a-Abstract.html

[22] Calhoun, V. D., & Adali, T. (2012). Multisubject independent component analysis of fMRI: A decade of intrinsic networks, default mode, and neurodiagnostic discovery. *IEEE reviews in biomedical engineering*, *5*, 60–73. https://doi.org/10.1109/RBME.2012.2211076