Tests Convergents, "Distribution-Free" et Affine-Invariants des Hypothèses du Modèle à Composantes Indépendantes

Marc Hallin [*1] & Simos Meintanis [2] & Klaus Nordhausen [3]

[1] *Université libre de Bruxelles, Belgique, mhallin@ulb.be*
[2] *Université Nationale et Capodistrienne d'Athènes, Grèce, simosmei@econ.uoa.gr*
[3] *Université de Jyväskylä, Finlande, klaus.k.nordhausen@jyu.fi*

Résumé. Nous proposons une famille de tests de la validité des hypothèses sous-jacentes aux méthodes de l'analyse en composantes indépendantes. Ces tests reposent sur des statistiques de type L_2 pondérées et font intervenir les fonctions caractéristiques empiriques des composantes indépendantes estimées; pour un choix approprié des pondérations et de la méthode d'estimation des matrices de *"demixing"* ces tests sont convergents et invariants par transformation affine. La loi asymptotique sous l'hypothèse de la statistique de test est cependant trop complexe pour une implémentation pratique, et nous établissons la validité en échantillons finis de valeurs critiques permutationnelles ou de rééchantillonnage tirant parti des symétries du problème (en dimension p, un groupe de permutations de $(n!)^{p-1}$ éleéments). Ces dernières conduisent à des tests "distribution-free" en dépit du fait que permutations et rééchantillonnage sont effectués sur les composantes indépendantes *estimées* ou leurs rangs marginaux. Les simulations établissent les bonnes performances de ces tests et leur supériorité par rapport à leur unique compétiteur, basé sur la notion de *"distance covariance".*

Mots-clés. Fonctions caractéristique; indépendence totale; composantes indépendantes; tests de rangs.

Abstract. We propose a family of tests of the validity of the assumptions underlying independent component analysis methods. The tests are formulated as L2–type procedures based on characteristic functions and involve weights; a proper choice of these weights and the estimation method for the mixing matrix yields consistent and affine-invariant tests. Due to the complexity of the asymptotic null distribution of the resulting test statistics, implementation is based on permutational and resampling strategies. This leads to distribution-free procedures regardless of whether these procedures are performed on the estimated independent components themselves or the componentwise ranks of their components. A Monte Carlo study involving various estimation methods for the mixing matrix, various weights, and a competing test based on distance covariance is conducted under the null hypothesis as well as under alternatives.

Keywords. Characteristic function; total independence; independent component model; rank test.

*Marc Hallin gratefully acknowledges the support of the Czech Science Foundation grants GAČR22036365 and GA24-10078S.

1 Testing the validity of the independent component model

Consider the *independent component model* (ICM)

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{\Omega}\boldsymbol{Z} \tag{1}$$

whereby the p-dimensional random vector $\boldsymbol{X} \coloneqq (X_1, ..., X_p)^\top$, $p > 1$ is linearly associated with a latent random vector $\boldsymbol{Z} = \boldsymbol{Z}(\boldsymbol{\mu}, \boldsymbol{\Omega}) \coloneqq (Z_1(\boldsymbol{\mu}, \boldsymbol{\Omega}), ..., Z_p(\boldsymbol{\mu}, \boldsymbol{\Omega}))^\top = (Z_1, ..., Z_p)^\top$ of centered independent components, i.e., $\boldsymbol{Z}(\boldsymbol{\mu}, \boldsymbol{\Omega})$ is enjoying the property of *total independence*:[1] the p-dimensional distribution of $\boldsymbol{Z}(\boldsymbol{\mu}, \boldsymbol{\Omega})$ is the product of the (marginal) distributions of its components $Z_1(\boldsymbol{\mu}, \boldsymbol{\Omega}), \ldots, Z_p(\boldsymbol{\mu}, \boldsymbol{\Omega})$.

Denoting by φ_ℓ, $\ell = 1, ..., p$ and φ the characteristic functions (CFs) of $Z_\ell(\boldsymbol{\mu}, \boldsymbol{\Omega})$ and $\boldsymbol{Z}(\boldsymbol{\mu}, \boldsymbol{\Omega})$, respectively, the property of total independence is equivalent to

$$\varphi(\boldsymbol{t}) = \prod_{\ell=1}^{p} \varphi_\ell(t_\ell), \qquad \boldsymbol{t} = (t_1, ..., t_p)^\top \in \mathbb{R}^p. \tag{2}$$

The linear structure of (1) involves a mean vector $\boldsymbol{\mu} \in \mathbb{R}^p$ and a $(p \times p)$ matrix $\boldsymbol{\Omega}$, often termed the *mixing matrix*, which belongs to the space $\mathbb{M}^p$ of full-rank $(p \times p)$ matrices.

The independent component model (ICM) is underlying the broad class of methods known as *independent component analysis* (ICA). A widely used method—a Google search (google.com accessed on 05.03.2024) returns no less than 532,000,000 links!—ICA originally was developed in the engineering literature on signal processing, where it has countless applications. It recently also attracted much interest in finance and economics: see, for instance, Comon and Jutten [2010], Garcia-Ferrer et al. [2012], Matteson and Tsay [2017], Gourieroux et al. [2017], Hai [2020], or Miettinen et al. [2020], to quote only a few. All these applications crucially rely on the validity of the ICM assumptions (1)–(2) in some observed sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ of i.i.d. copies of $\boldsymbol{X}$. Somewhat surprisingly, though, no test of the validity of these assumptions, i.e., of the the null hypothesis

$$\mathcal{H}_0\colon \text{ the random vector } \boldsymbol{X} \text{ is such that (1) and (2) hold for some } (\boldsymbol{\mu}, \boldsymbol{\Omega}) \in \mathbb{R}^p \times \mathbb{M}^p \tag{3}$$

is available in the literature except for a procedure proposed in Matteson and Tsay [2017] who are using the notion of *distance covariance* and resampling-based critical values. Their test statistic, however, depends on the arbitrary ordering of the components of $\boldsymbol{Z}$, which is not a desirable property. Our objective here is to develop tests of $\mathcal{H}_0$ which, as we shall see, are universally consistent and outperform the Matteson and Tsay procedure.

Our tests are based on CFs and the fact that $\mathcal{H}_0$ holds if and only if there exists $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$ such that (2) is satisfied. More precisely, our test statistics are weighted measures of the discrepancy between the (empirical) joint CF and the product of marginal ones—a classical and quite natural measure of total dependence that can be traced back, e.g., to Csörgő and

[1] While *total independence* (2) implies *pairwise independence*—namely, $Z_j \perp\!\!\!\perp Z_k$ for all $j \neq k = 1, \ldots, p$ where $\perp\!\!\!\perp$, as usual, denotes stochastic independence—the converse, of course, is not true.

Hall [1982] and is also considered in Chen and Bickel [2005] in an estimation context. Nobody ever used these measures of total dependence as test statistics nor provided their asymptotic distributions under the assumption of total independence. The reason for this is that these asymptotic distributions (the distributions of infinite sums of chi-squared random variables weighted by the eigenvalues of a complicated integral operator) are hopelessly untractable: indeed, they not only depend on the actual distribution of the observations but also on the way (FOBI, JADE, FastICA, ...) the mixing matrix is estimated. They cannot be used in the construction of asymptotic critical values and hence are entirely useless in practice. This is probably the reason why these very natural characteristic-function-based statistics so far have not been used as test statistics.

The main and nontrivial theoretical novelty of this contribution is to show that these impracticable asymptotic critical values can be dispensed with and replaced with exact permutational ones. These permutational critical values, however, follow from a non-standard permutational scheme—not the usual $n!$ permutations of the observations, which leave the test statistics invariant! Moreover, the validity of our permutational approach is anything but straightforward, since it involves permutations (in dimension p, there are $(n!)^{p-1}$ of them) of the components of estimated residuals that are not i.i.d. under $\mathcal{H}_0$.

Our tests, unlike the Matteson and Tsay ones, are affine-invariant and have exact size α uniformly over $\mathcal{H}_0$, irrespective of the actual (absolutely continuous) distribution of the observations and the demixing method (FOBI, JADE, FastICA, ...) adopted for deriving the residuals $\widehat{\mathbf{Z}}$. They involve weights W; when choosing them, we pay special attention to their impact on the computation of the corresponding test statistics, a feature that is particularly important in case of high-dimensional ICMs.

Now, beyond its ICM structure, $\mathcal{H}_0$ does not specify anything about $\boldsymbol{X}$; in particular, the distribution of $\boldsymbol{Z}$, hence that of $\boldsymbol{X}$, remains completely unspecified under $\mathcal{H}_0$ as well as under the alternative (which includes arbitrary dependencies between the components of $\boldsymbol{X}$). In principle, that distribution even could be discrete, although we are focusing on the absolutely continuous case for simplicity. The problem, therefore, is of a semiparametric nature, where the nuisance parameters are unspecified distributions. Ranks, in that context, naturally come into the picture, and we also propose rank-based tests of $\mathcal{H}_0$ which for the same reason as above, admit fully distribution-free exact critical values.

While the problem of testing bivariate independence has been treated quite extensively in the literature, the results on testing total independence are less abundant; see, however, Blum et al. [1961], Deheuvels [1981], Csörgő [1985], Kankainen [1995]. The related problem of testing vector independence—independence between vectors of finite dimension—recently attracted renewed interest: we refer to Roy et al. [2020], Genest et al. [2019], Herwartz and Maxand [2020], Shi et al. [2022a] for reviews of the existing literature. CF-based methods in this context date back to Csörgő and Hall [1982], Csörgő [1985], Feuerverger [1993], Kankainen and Ushakov [1998]; for more recent contributions, see Meintanis and Iliopoulos [2008], Fan et al. [2017], Pfister et al. [2017], Chakraborty and Zhang [2019], among others. In all these tests, an empirical contrast between the joint and the product of marginal CFs is considered. The advantages, in this context, of the CF-based approach over competing methods based, e.g., on distribution function are confirmed by the success of distance covariance methods

and their relation [Sejdinovic et al., 2013] to RKHS statistics; see Edelmann et al. [2019], Székely and Rizzo [2023], and Chen et al. [2019] for reviews on distance covariance and related approaches, and Eriksson and Koivunen [2003] and Chen and Bickel [2005] for a discussion of their advantages in the ICM context.

As for the advantages of rank-based procedures in the presence of unspecified distributions, we refer to Shi et al. [2022b], who show that pseudo-Gaussian methods such as Wilks' classical test for vector independence [Wilks, 1935], although asymptotically valid, severely over-reject under skewed distributions. Rank-based independence testing methods therefore are a natural solution and have been thoroughly investigated in the bivariate context—see, e.g., Chapters II.4.11 and III.6 of Hájek and Šidák [1967]. With the recent introduction of measure transportation-based concepts of multivariate ranks and signs [Chernozhukov et al., 2017, Hallin et al., 2021], this approach to bivariate independence testing has been extended to arbitrary dimensions [Shi et al., 2022a,b, 2023, Ghosal and Sen, 2022]. The problem we are facing here, again, is trickier. Independence, in these references, is to be tested between observable quantities which under the null are i.i.d. so that the distribution-freeness of rank-based statistics is straightforward while we are dealing with unobservable variables $Z_{j\ell}$, the values of which have to be estimated as $\widehat{Z}_{j\ell}$. These "estimated residuals" are no longer i.i.d. under the null, and handling them requires additional care.

Still in the ICM context, rank-based tests of vector independence have been proposed by Oja et al. [2016]. Their problem is quite different from ours, though: instead of the validity of the ICM assumptions, these authors are testing the independence of two given subvectors of $\boldsymbol{X}$ under maintained ICM assumptions.

2 The test statistics

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, with $\boldsymbol{X}_j := (X_{j1}, \ldots, X_{jp})^\top$, $j = 1, ..., n$ denote a sample of n independent copies of $\boldsymbol{X}$. To implement the test, since we only observe $\boldsymbol{X}_j$ and not $\boldsymbol{Z}_j$, the latent variables $(\boldsymbol{Z}_j,\ j = 1, ..., n)$ first need to be estimated from the data. Specifically, our test statistic is a function of the "estimated ICM residuals"

$$\widehat{\boldsymbol{Z}}_j = \widehat{\boldsymbol{Z}}_j(\widehat{\boldsymbol{\mu}}_n, \widehat{\boldsymbol{\Omega}}_n) := \widehat{\boldsymbol{\Omega}}_n^{-1}\left(\boldsymbol{X}_j - \widehat{\boldsymbol{\mu}}_n\right) =: (\widehat{Z}_{j1}, ..., \widehat{Z}_{jp})^\top,\ j = 1, ..., n \tag{4}$$

obtained by plugging estimators $\widehat{\boldsymbol{\mu}}_n$ and $\widehat{\boldsymbol{\Omega}}_n$ of $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$ into (1). It is well known, however, that the matrix $\boldsymbol{\Omega}$ in (1) is not uniquely identified, and identification constraints need to be imposed which, without any loss of generality, actually define the parameter space for $\boldsymbol{\Omega}$. These constraints, which are closely related to the construction and statistical properties of $\widehat{\boldsymbol{\Omega}}_n$ are omitted here.

We now rapidly describe the proposed test statistics. Denote by

$$\varphi^{(n)}(\boldsymbol{t}) := \frac{1}{n}\sum_{j=1}^{n} \mathrm{e}^{\mathrm{i}\boldsymbol{t}^\top \widehat{\boldsymbol{Z}}_j}, \quad \boldsymbol{t} \in \mathbb{R}^p \tag{5}$$

and

$$\varphi_\ell^{(n)}(t_\ell) \coloneqq \frac{1}{n}\sum_{j=1}^{n} \mathrm{e}^{\imath t_\ell \widehat{Z}_{j\ell}}, \quad t_\ell \in \mathbb{R}, \quad \ell = 1, \ldots, p \tag{6}$$

(where $\imath$ stands for the imaginary root of -1) the joint and marginal empirical CFs, respectively, of the estimated ICM residuals $\widehat{\boldsymbol{Z}}_j$, $j = 1, \ldots, n$.

Similarly, letting $\mathbf{R}_j^{(n)} \coloneqq (R_{j1}^{(n)}, \ldots, R_{jp}^{(n)})^\top$ where $R_{j\ell}^{(n)}$, $j = 1, \ldots, n$, stands for the rank of $\widehat{Z}_{j\ell}$ among $\widehat{Z}_{1\ell}, \ldots, \widehat{Z}_{n\ell}$, $\ell = 1, \ldots, p$, define the joint and marginal rank-based statistics

$$\underset{\sim}{\varphi}_{\mathbf{J}}^{(n)}(\boldsymbol{t}) \coloneqq \frac{1}{n}\sum_{j=1}^{n} \mathrm{e}^{\imath \boldsymbol{t}^\top \mathbf{J}(\mathbf{R}_j^{(n)}/(n+1))}, \quad \boldsymbol{t} \in \mathbb{R}^p \tag{7}$$

and

$$\underset{\sim}{\varphi}_{\mathbf{J},\ell}^{(n)}(t_\ell) \coloneqq \frac{1}{n}\sum_{j=1}^{n} \mathrm{e}^{\imath t_\ell J_\ell(R_{j\ell}^{(n)}/(n+1))}, \quad t_\ell \in \mathbb{R}, \quad \ell = 1, \ldots, p \tag{8}$$

where $\mathbf{J}(\mathbf{R}_j^{(n)}/(n+1)) \coloneqq (J_1(R_{j1}^{(n)}/(n+1)), \ldots, J_p(R_{jp}^{(n)}/(n+1)))^\top$ and $\mathbf{J} \coloneqq (J_1, \ldots, J_p)^\top$ denotes a p-tuple of *score functions* $J_\ell : (0,1) \to \mathbb{R}$. Clearly, (7) and (8) are the joint and marginal empirical CFs, respectively, of the *scored ranks* $J_\ell(R_{j\ell}^{(n)}/(n+1))$ of the estimated ICM residuals $\widehat{\boldsymbol{Z}}_j$, $j = 1, \ldots, n$.

Our tests are rejecting the null hypothesis $\mathcal{H}_0$ in (3) for large values of the test statistics

$$T_{n,W} \coloneqq n\int_{\mathbb{R}^p} |D_n(\boldsymbol{t})|^2 \, W(\boldsymbol{t})\mathrm{d}\boldsymbol{t} \quad \text{and} \quad \underset{\sim}{T}_{n,\mathbf{J},W} \coloneqq n\int_{\mathbb{R}^p} |\underset{\sim}{D}_{n,\mathbf{J}}(\boldsymbol{t})|^2 \, W(\boldsymbol{t})\mathrm{d}\boldsymbol{t}, \tag{9}$$

(of the Cramér-von Mises type) where

$$D_n(\boldsymbol{t}) \coloneqq \varphi^{(n)}(\boldsymbol{t}) - \prod_{\ell=1}^{p} \varphi_\ell^{(n)}(t_\ell) \quad \text{and} \quad \underset{\sim}{D}_{n,\mathbf{J}}(\boldsymbol{t}) \coloneqq \underset{\sim}{\varphi}_{\mathbf{J}}^{(n)}(\boldsymbol{t}) - \prod_{\ell=1}^{p} \underset{\sim}{\varphi}_{\mathbf{J},\ell}^{(n)}(t_\ell) \tag{10}$$

with $\boldsymbol{t} \coloneqq (t_1, t_2, \ldots, t_p)^\top \in \mathbb{R}^p$ and some weight function $W : \boldsymbol{t} \mapsto W(\boldsymbol{t})$. Test statistics of the Kolmogorov-Smirnov type $S_n \coloneqq \sqrt{n}\sup_{\boldsymbol{t}\in\mathbb{R}^p} |D_n(\boldsymbol{t})|$ and $\underset{\sim}{S}_{n,\mathbf{J}} \coloneqq \sqrt{n}\sup_{\boldsymbol{t}\in\mathbb{R}^p} |\underset{\sim}{D}_{n,\mathbf{J}}(\boldsymbol{t})|$ could also be considered; see for instance Csörgő [1985]. However, computing a $\sup_{\boldsymbol{t}\in\mathbb{R}^p}$ runs into technical difficulties (already apparent in Csörgő [1985]) since the empirical CF converges weakly only over compact neighborhoods of $\mathbb{R}^p$ and this supremum has to be taken over some discrete grid. These drawbacks appear to have a direct impact on finite–sample powers, and Kolmogorov–Smirnov type tests typically are less powerful than their Cramér–von Mises counterparts (Henze et al. [2014]). We, therefore, only consider the latter.

References

J. R. Blum, J. Kiefer, and M. Rosenblatt. Distribution-free tests of independence based on the sample distribution function. *Annals of Mathematical Statistics*, 32:485–498, 1961.

S. Chakraborty and X. Zhang. Distance metrics for measuring joint dependence with application to causal inference. *Journal of the American Statistical Association*, 114:1638–1650, 2019.

A. Chen and P. Bickel. Consistent independent component analysis and prewhitening. *IEEE Transactions on Signal Processing*, 53:3625–3632, 2005.

F. Chen, S. G. Meintanis, and L. Zhu. On some characterizations and multidimensional criteria for testing homogeneity, symmetry and independence. *Journal of Multivariate Analysis*, 173:125–144, 2019.

V. Chernozhukov, A. Galichon, M. Hallin, and M. Henry. Monge–Kantorovich depth, quantiles, ranks and signs. *Annals of Statistics*, 45:223–256, 2017.

P. Comon and C. Jutten. *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Academic Press, Oxford, 2010.

S. Csörgő. Testing for independence by the empirical characteristic function. *Journal of Multivariate Analysis*, 16:290–299, 1985.

S. Csörgő and P. Hall. Estimable versions of Griffiths' measure of association. *Australian Journal of Statistics*, 24:296–308, 1982.

P. Deheuvels. An asymptotic decomposition for multivariate distribution-free tests of independence. *Journal of Multivariate Analysis*, 11:102–113, 1981.

D. Edelmann, K. Fokianos, and M. Pitsillou. An updated literature review of distance correlation and its applications to time series. *International Statistical Review*, 87:237–262, 2019.

J. Eriksson and V. Koivunen. Characteristic-function-based independent component analysis. *Signal Processing*, 83:2195–2208, 2003.

Y. Fan, P. Lafaye de Micheaux, S. Penev, and D. Salopek. Multivariate nonparametric test of independence. *Journal of Multivariate Analysis*, 153:189–210, 2017.

A. Feuerverger. A consistent test for bivariate dependence. *International Statistical Review*, 61:419–433, 1993.

A. Garcia-Ferrer, E. Gonzalez-Prieto, and D. Pena. A conditionally heteroskedastic independent factor model with an application to financial stock returns. *International Journal of Forecasting*, 28:70–93, 2012.

C. Genest, J. G. Nešlehová, B. Rémillard, and O. A. Murphy. Testing for independence in arbitrary distributions. *Biometrika*, 106:47–68, 2019.

P. Ghosal and B. Sen. Multivariate ranks and quantiles using optimal transport: Consistency, rates and nonparametric testing. *Annals of Statistics*, 50:1012–1037, 2022.

C. Gourieroux, A. Monfort, and J.-P. Renne. Statistical inference for independent component analysis: Application to structural VAR models. *Journal of Econometrics*, 196:111–126, 2017.

T. Hai. Estimation of volatility causality in structural autoregressions with heteroskedasticity using independent component analysis. *Statistical Papers*, 61:1–16, 2020.

J. Hájek and Z. Šidák. *Theory of Rank Tests*. Academic Press, New York, 1967.

M. Hallin, E. del Barrio, J. Cuesta-Albertos, and C. Matrán. Distribution and quantile functions, ranks and signs in dimension d: A measure transportation approach. *Annals of Statistics*, 49:1139–1165, 2021.

N. Henze, Z. Hlávka, and S. Meintanis. Testing for spherical symmetry via the empirical characteristic function. *Statistics*, 48:1282–1296, 2014.

H. Herwartz and S. Maxand. Nonparametric tests for independence: a review and comparative simulation study with an application to malnutrition in India. *Statistical Papers*, 61: 2175–2201, 2020.

A. Kankainen. *Consistent testing of total independence based on the empirical characteristic function*. PhD thesis, University of Jyväskylä, 1995.

A. Kankainen and N. Ushakov. A consistent modification of a test for independence based on the empirical characteristic function. *Journal of Mathematical Sciences*, 89:1486–1494, 1998.

D. S. Matteson and R. S. Tsay. Independent component analysis via distance covariance. *Journal of the American Statistical Association*, 112:623–637, 2017.

S. G. Meintanis and G. Iliopoulos. Fourier methods for testing multivariate independence. *Computational Statistics & Data Analysis*, 52:1884–1895, 2008.

J. Miettinen, M. Matilainen, K. Nordhausen, and S. Taskinen. Extracting conditionally heteroskedastic components using independent component analysis. *Journal of Time Series Analysis*, 41:293–311, 2020.

H. Oja, D. Paindaveine, and S. Taskinen. Affine-invariant rank tests for multivariate independence in independent component models. *Electronic Journal of Statistics*, 10:2372–2419, 2016.

N. Pfister, P. Bühlmann, B. Schölkopf, and J. Peters. Kernel-based tests for joint independence. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80:5–31, 2017.

A. Roy, S. Sarkar, A. K. Ghosh, and A. Goswami. On some consistent tests of mutual independence among several random vectors of arbitrary dimensions. *Statistics and Computing*, 30:1707–1723, 2020.

D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS–based statistics in hypothesis testing. *Annals of Statistics*, 41:2263–2291, 2013.

H. Shi, M. Drton, and F. Han. Distribution-free consistent independence tests via center-outward ranks and signs. *Journal of the American Statistical Association*, 117:395–410, 2022a.

H. Shi, M. Hallin, M. Drton, and F. Han. On universally consistent and fully distribution-free rank tests of vector independence. *Annals of Statistics*, 50:1933–1959, 2022b.

H. Shi, M. Drton, M. Hallin, and F. Han. Distribution-free tests of multivariate independence based on center-outward quadrant, Spearman, and Kendall statistics. *Bernoulli,* to appear, 2023.

G. J. Székely and M. L. Rizzo. *The Energy of Data and Distance Correlation.* CRC Press, Florida, USA, 2023.

S. S. Wilks. On the independence of k sets of normally distributed statistical variables. *Econometrica*, 3:309–326, 1935.