STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–15.
Published online in International Academic Press (www.IAPress.org)

A note on the test for complete independence in the high-dimensional setting

Filipe J. Marques*

Center for Mathematics and Applications (NOVA Math) and Department of Mathematics, NOVA School of Science and Technology (NOVA FCT), Caparica, Portugal

Abstract In this work, we propose a testing methodology for assessing complete independence in both low- and high-dimensional scenarios. The test statistic is constructed from likelihood ratio statistics used to test the independence of pairs of variables and is bounded in the interval $(0, 1)$. The properties of the test are explored through simulations in terms of power and bias, considering three different covariance structures: compound symmetric, first-order autoregressive, and band-Toeplitz-type covariance matrices. Furthermore, we propose asymptotic approximations based on a single Gamma distribution or a mixture of two Gamma distributions that are computationally efficient and straightforward to implement in practical applications. Numerical studies are provided to illustrate the accuracy of the proposed asymptotic approximations when compared with other models. An application to a high-dimensional financial asset returns dataset is also provided.

Keywords Asymptotic Approximations, Autoregressive, Complete Independence, Compound Symmetric, Likelihood Ratio Tests, Mixtures, Sample Correlations, Toeplitz

AMS 2010 subject classifications 62H15, 62E20, 62F03, 62F05

DOI: 10.19139/soic-2310-5070-3643

1. Introduction

In recent years, with the growth of data and its increasing complexity, the high-dimensional setting has attracted considerable attention, and several works have addressed different statistical problems under this framework. In this study, we focus on a fundamental test, namely the test for complete independence used to assess the independence of a set of variables of a multivariate normal vector both in the low and high dimension scenarios. Note that under the assumption of normality and of complete independence the covariance matrix of the random vector has a diagonal structure.

The test of complete independence is a particular case of a more general test, commonly referred to as block independence, which is used to test the independence of groups of variables. In the complete independence case, all groups consist of a single variable. In the low-dimensional setting, this test can be addressed using the standard likelihood ratio approach [1].

For the high dimensional setting, [19], proposed a test statistic based on the sum of the squares of the sample correlations and derived its limiting distribution as an univariate normal distribution. The same statistic was proposed in [20] together with results for other covariance structures. In [4], [11] and [16] the authors present alternative test statistics which are generalizations based on the statistic introduced in [19].

*Correspondence to: Filipe J. Marques (Email: fjm@fct.unl.pt). Center for Mathematics and Applications (NOVA Math) and Department of Mathematics, NOVA School of Science and Technology (NOVA FCT), Caparica, Portugal. Campus de Caparica, 2829 -516 Caparica, PORTUGAL.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

In the literature, it is also possible to find other methodologies, including a test statistic based on the Fisher's z transformation, originally proposed by [5], distance-based and kernel-based approaches [21, 10], rank-based and distribution-free methods [12, 3], and bootstrap and practical independence procedures [18, 2]. These alternative frameworks offer different trade-offs in terms of assumptions and finite-sample performance.

We propose a test statistic that is also based on sample correlations, that is bounded in the interval (0,1), and is motivated by the idea of testing the independence between all possible pairs of variables using the likelihood ratio test for independence and also by the work of [19]. We show that the distribution of this test statistic can be expressed as the product of Beta random variables, and we propose asymptotic approximations based on either a single Gamma distribution or on a mixture of two Gamma distributions to approximate the distribution of the negative logarithm of the test statistic. These approximations follow related approaches used to approximate the sum of the negative logarithms of Beta random variables (see for example [7] and [14]).

The derived results are of key methodological relevance, as they not only provide a new method for testing complete independence in a simple and precise manner, but also introduce what we call an Almost Likelihood Ratio Test (ALRT) statistic. Although based on likelihood ratio statistics, the ALRT statistic is not formally a likelihood ratio statistic; however, since it is bounded in the interval $(0,1)$, it can be combined with classical likelihood ratio statistics to construct tests for more general hypotheses, such as the block independence test, which allows for testing independence between groups of variables [14], or the sphericity test [7].

The paper is organized as follows. Section 2 introduces the null hypothesis for the test of complete independence, the proposed test statistic, and asymptotic approximations to its distribution. Section 3 presents numerical studies illustrating (i) the test properties with respect to power and bias, and (ii) the accuracy of the proposed approximations in both low- and high-dimensional scenarios. An application to high-dimensional financial asset returns data set is provided in Section 4. Finally, Section 5 summarizes the main conclusions.

2. The test statistic and its distribution

Let us consider a sample of size N extracted from a multivariate normal population of dimension p, with mean vector $\underline{\mu}$ and covariance matrix Σ, denoted $N_p(\underline{\mu}, \Sigma)$ with Σ given by

$$\Sigma = \left(\begin{array}{cccc} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{array} \right) .$$

Testing for complete independence corresponds to test the independence among all variables, which, under the assumption of multivariate normality, is equivalent to test the nullity of the covariances (or equivalently, the correlations) for all possible pairs of variables. Accordingly, the hypotheses can be formulated as follows

$$H_0 : \sigma_{ij} = 0 \text{ for all } i \neq j \quad \text{vs} \quad H_1 : \sigma_{ij} \neq 0 \text{ for some } i \neq j \, . \tag{1}$$

For $n > p$, for low dimensions, the expression of the likelihood ratio test statistic is given by [1, 17]

$$\Lambda_{\text{LRT}} = \left(\frac{|A|}{\prod_{i=1}^{p} a_{ii}} \right)^{\frac{N}{2}} = |R|^{\frac{N}{2}} \tag{2}$$

where R is the sample correlation matrix, A is the sample covariance matrix, and a_{ii} are the diagonal elements of A. The statistic in (2) was addressed, for example, in [1] and [17]. The exact distribution of Λ_{LRT} can be represented as the distribution of the product of independent Beta random variables; however, its complexity generally requires the use of approximations. The simplest approximation is based on Wilks' theorem [22], which states that the distribution of the negative logarithm of the likelihood ratio statistic can be approximated by a chi-squared distribution. In [1], a refined approximation based on a mixture of two χ^2 distributions is proposed. Other approaches have also been developed, including near-exact approximations, which exhibit high accuracy and good asymptotic properties [6, 7], although they may require greater computational effort for practical implementation.

In [19], the author proposes a statistic, which is based on the sum of the squares of the sample correlations, and shows that the statistic converges in distribution to a normal distribution. In [11], extensions of the test statistic proposed in [19] are discussed. In [15], the authors also addressed the complete independence test, both for low and high dimensions, considering some additional restrictions, and showed that the distribution of the test statistic was based on a single χ^2 distribution.

We consider a different test statistic which can also be used in both low- and high-dimensional settings. Testing the null hypothesis in (1) can be done by testing the independence for each possible pair of variables, which can be done by the intersection of the $p(p-1)/2$ null hypotheses used to test the nullity of the covariance (or correlations) for each possible pair of variables, namely

$$H_0 = \bigcap_{i>j} H_0^{i,j} \tag{3}$$

for $i > j$ and $i, j = 1 \ldots, p$, where

$$H_0^{i,j} : \sigma_{ij} = 0 \tag{4}$$

is the null hypothesis to test nullity of the covariance between X_i and X_j, that is, under normality assumption, the null hypothesis of independence between X_i and X_j with $i > j$ and $i, j = 1 \ldots, p$. The power $2/N$ of the likelihood ratio test statistic to test the null hypothesis in (4) for some choice of i and j such that $i > j$ is

$$\Lambda_{ij} = \frac{|A_{ij}|}{a_{ii}a_{jj}} = \frac{a_{ii}a_{jj} - a_{ij}^2}{a_{ii}a_{jj}} = 1 - r_{ij}^2 \tag{5}$$

where A_{ij} is the matrix of order 2, built from the sample covariance matrix, A, as

$$A_{ij} = \begin{pmatrix} a_{ii} & a_{ij} \\ a_{ij} & a_{jj} \end{pmatrix}$$

and where a_{ii} and a_{jj} are the sample variances of the variables X_i and X_j, respectively, a_{ij} is the sample covariance between the variables X_i and X_j, and r_{ij} is the sample correlation between X_i and X_j. To test the null hypothesis in (3) we propose the test statistic

$$\Lambda = \prod_{i=2}^{p} \prod_{j=1}^{i-1} (1 - r_{ij}^2) \tag{6}$$

which is bounded in the interval $(0, 1)$ and for a sample of size N we will reject the null hypothesis in (3) if $\Lambda_{obs} < q_{0.05}$ where $q_{0.05}$ stands for the 0.05 quantile of the distribution of Λ in (6).

Although the statistic in (6) is the product of likelihood ratio statistics, we should note that it is not a likelihood ratio test statistic. The likelihood ratio test for complete independence given in (2) can be obtain from a standard maximum likelihood approach or may also be constructed from the decomposition of the null hypothesis into several hypotheses which state that each pair of variables is independent conditional on the remaining variables which, in the multivariate normal case, is naturally formulated in terms of partial correlations. For this reason, we may refer to this statistic as an Almost Likelihood Ratio Test statistic.

For $i, j = 1, \ldots, p$ such that $i > j$, under H_0, the density function of r_{ij} is given in [13] which can be used to show that the distribution of r_{ij}^2 is

$$r_{ij}^2 \sim \text{Beta}\left(\frac{1}{2}, \frac{N-2}{2}\right)$$

and thus, Λ_{ij} in (5) is distributed as

$$1 - r_{ij}^2 \sim \text{Beta}\left(\frac{N-2}{2}, \frac{1}{2}\right).$$

Then, the statistic Λ in (6) has the distribution of the product of $p(p-1)/2$ Beta random variables with parameters $(N-2)/2$ and $1/2$.

Since it is more convenient to work with sums of random variables, and following a similar procedure to the one used in related papers [7, 14] we consider the variable

$$T = -\log(\Lambda)$$

with Λ defined in (6). Thus,

$$T = -\log(\Lambda) = \sum_{i=2}^{p} \sum_{j=1}^{i-1} -\log(1 - r_{ij}^2) \tag{7}$$

which has the distribution of the sum of $p(p-1)/2$ Logbeta random variables, denoted by Logbeta$((N-2)/2, 1/2)$.

We should note that, although we can not ensure the independence of all the Beta random variables, since the sample correlations, r_{ij}, are not all mutually independent, the simulations studies show that, under H_0, this assumption can be made without significantly affecting either the properties of the test or the accuracy of the approximations provided.

Assuming the independence of the sample correlations, the first four moments of T can be expressed in terms of the polygamma function of order k, $\psi^{(k)}(\cdot)$, with

$$\psi^{(0)} = \psi \quad \text{(digamma)}, \qquad \psi^{(1)} = \psi', \qquad \psi^{(2)} = \psi'', \quad \ldots$$

See Appendix B for the expressions of the null moments of T.

Motivated by the results in [1] and [22] where a single χ^2 or a mixture of χ^2 distributions is used to approximate the distribution of sums of independent Logbeta random variables, and also based on approximations provided in related studies such as in [7] and [14], we propose two simple approximations for the distribution of T. Thus, we propose a single gamma distribution, $\Gamma(r, \lambda)$, matching 2 exact moments with density function given by

$$f_{\Gamma(r,\lambda)}(t) = \frac{\lambda^{-r}}{\Gamma(r)} t^{r-1} e^{-t/\lambda}, \; t > 0 \tag{8}$$

in this case we have

$$r = -\frac{m_1^2}{m_1^2 - m_2} \quad \text{and} \quad \lambda = \frac{m_2}{m_1} - m_1$$

where $m_i = E[T^i]$ for $i = 1, 2$ given respectively in (13) and (14) or a mixture of 2 gamma distributions, $M2\Gamma$, matching 4 exact moments, with density function given by

$$f_{M2\Gamma}(t) = \pi \frac{\lambda^{-r_1}}{\Gamma(r_1)} t^{r_1 - 1} e^{-t/\lambda} + (1 - \pi) \frac{\lambda^{-r_2}}{\Gamma(r_2)} t^{r_2 - 1} e^{-t/\lambda}, \; t > 0 \tag{9}$$

where the parameters r_1, r_2, λ and π in (9) are determined in such way that the approximating distribution matches the first 4 exact moments of T.

3. Numerical Studies

In this section, extensive simulations are used to study the properties of the test, namely with respect to its power and significance level, as well as to study the reliability and precision of the proposed approximations for the test statistic.

3.1. Test properties

In order to assess the test properties we simulate data from a multivariate Normal distribution with null mean vector and covariance matrix with the following structures i) a compound symmetric (CS), ii) a first order autoregressive (AR) or iii) a band-Toeplitz covariance matrix (bT) with bandwidth $k = 1$. The CS structure ensures that all the variables are equally correlated with each other and all have the same variance and may be represented as

$$\Sigma_{CS} = \sigma^2 \begin{pmatrix} 1 & \rho & \rho & \dots & \rho \\ \rho & 1 & \rho & \dots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \dots & 1 \end{pmatrix} = \sigma^2 \left[(1-\rho) I_p + \rho \mathbf{1}_p \mathbf{1}_p' \right] \tag{10}$$

where $\mathbf{1}_p$ is a vector of ones, I_p is the identity matrix of order p, $\sigma^2 > 0$ and $-1/(p-1) < \rho < 1$, while the AR structure ensures that the dependence between two random variables X_t and X_s is governed strictly by their temporal distance $|t - s|$ and may be represented as

$$\Sigma_{AR} = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{p-1} \\ \rho & 1 & \rho & \cdots & \rho^{p-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{p-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{p-1} & \rho^{p-2} & \rho^{p-3} & \cdots & 1 \end{pmatrix} = \sigma^2 \left(\rho^{|i-j|} \right)_{i,j} \text{, with } \; i,j = 1, \dots, p. \tag{11}$$

The band-Toeplitz (bT) with bandwidth $k = 1$ covariance matrix implies that the dependency is strictly local, that is any two variables separated by a distance greater than a specific threshold, in our case $k = 1$, are completely uncorrelated, resulting in a covariance of zero. This structure can be represented in the following way

$$\Sigma_{bT} = \sigma^2 \begin{pmatrix} 1 & \rho & 0 & \dots & 0 & 0 & 0 \\ \rho & 1 & \rho & \dots & 0 & 0 & 0 \\ 0 & \rho & 1 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & \rho & 0 \\ 0 & 0 & 0 & \dots & \rho & 1 & \rho \\ 0 & 0 & 0 & \dots & 0 & \rho & 1 \end{pmatrix}. \tag{12}$$

In the simulations presented in this subsection, we used 0.05 empirical quantiles of the distribution of T, and the range of values considered for ρ in each plot was the one that ensured that Σ_{CS}, Σ_{AR} or Σ_{bT}, defined respectively in (10), (11) and (12) were positive definite. Whenever $n > p$, we also consider the likelihood ratio test statistic for complete independence in (2). Thus, in Figures 1–3, the estimated power when using the statistic T in (6) is shown as a solid line, while the estimated power when using the likelihood ratio test statistic, Λ_{LRT}, in (2) is shown as a dashed line; the latter was only computed when $n > p$. The red line represents the 0.05 probability level, which is expected to be attained when $\rho = 0$.

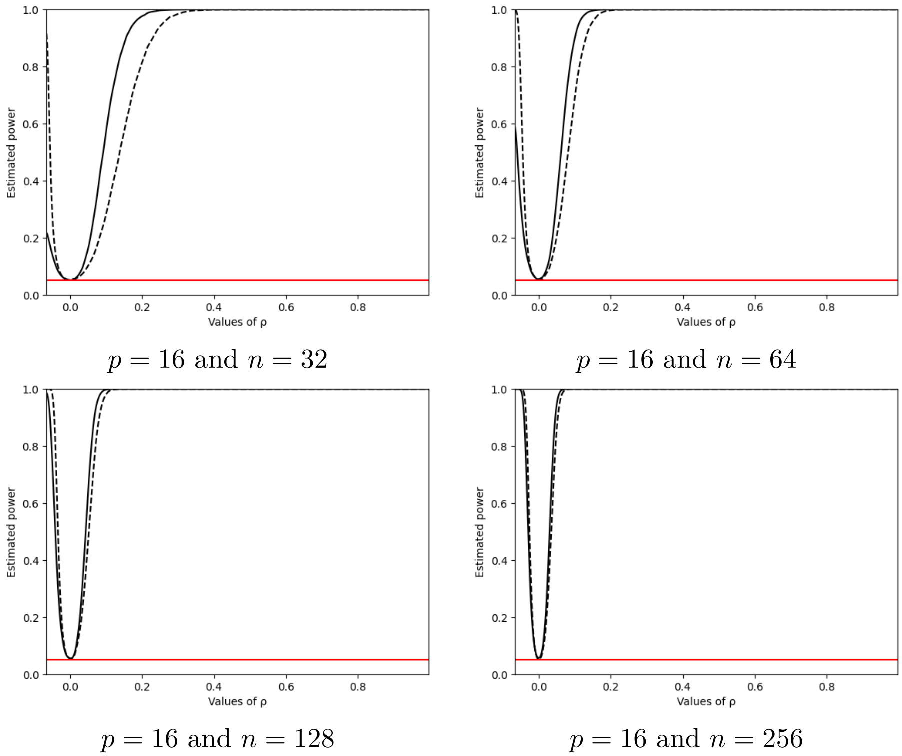


Figure 1. Estimated power for the complete independence test, for data simulated from a population with a CS covariance structure, in low-dimension scenarios using the likelihood ratio statistic in (2) (dashed line) and the proposed statistic in (6) (solid line)

From Figures 1–3 it is possible to see, for data simulated from a multivariate normal distribution with CS, AR or bT covariance structures, in low dimensions, for $p = 16$, that both the likelihood ratio test and the proposed testing methodology, seem to be unbiased, in the sense that the estimated power is 0.05 when $\rho = 0$, and that the power of the test based on the statistic T seems to converge to the power of the likelihood ratio test when n increases. It is also possible to see that the agreement between the estimated power for the two methodologies is more clear for the AR or bT covariances structures (see Figures 2–3), even when the dimension is low when compared with the number of variables. The results obtained for low dimension scenarios with $p = 8$ are presented in Figures 11-13 in Appendix A.

In high dimensional scenarios, $p > n$, in Figures 4-6, the estimated power for the new methodology proposed in this work is presented. Note that, the likelihood ratio test can not be applied in high dimensional scenarios. We may see from Figures 4 and 5, when the CS and AR structures are considered, that the power increases rapidly to 1 when the dimension increases, that is when the number of variables, p, increases even for small values of n, for example $n = 8$ (see the left plots in Figures 4 and 5). However, when the bT covariance structure is considered, from Figure 6, we see that when $n = 8$ (left plot of Figure 6) the power does not reaches 1, and just slightly increases with the number of variables. This behaviour may be explained by the low value of the sample size, since when we consider $n = 20$ the results show the power converging to 1 when ρ differs significantly from zero (see Figure 6, plot on the right).

3.2. Asymptotic approximations

To evaluate the accuracy of the proposed approximations in Section 2, we simulated 100 000 values of the statistic T in (7), under the null hypothesis, from a multivariate normal population with zero mean vector

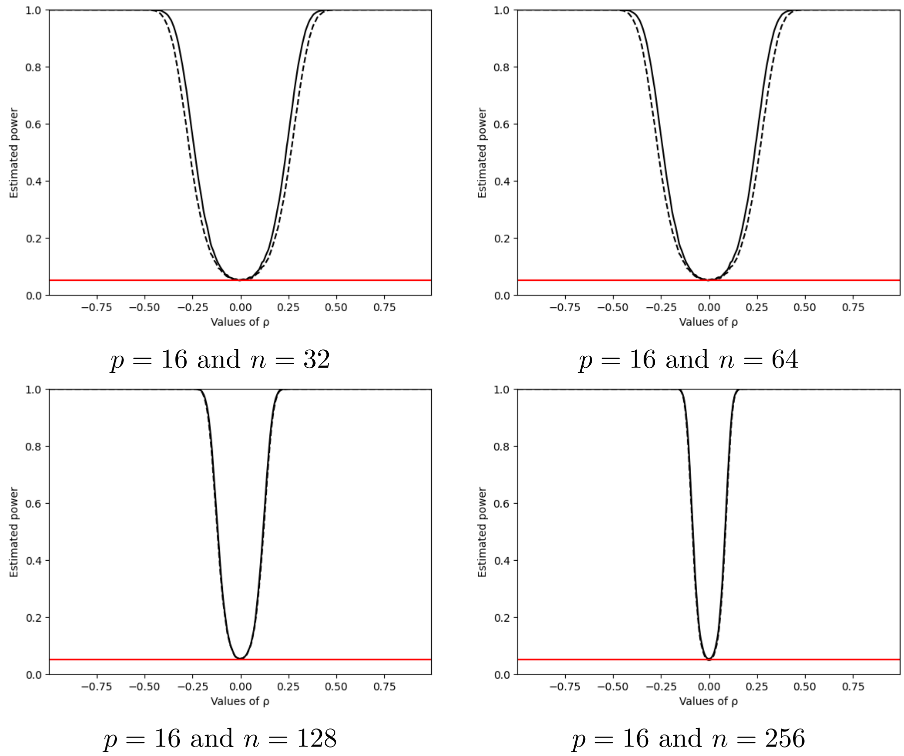


Figure 2. Estimated power for the complete independence test, for data simulated from a population with a AR covariance structure, in low-dimension scenarios using the likelihood ratio statistic in (2) (dashed line) and the proposed statistic in (6) (solid line)

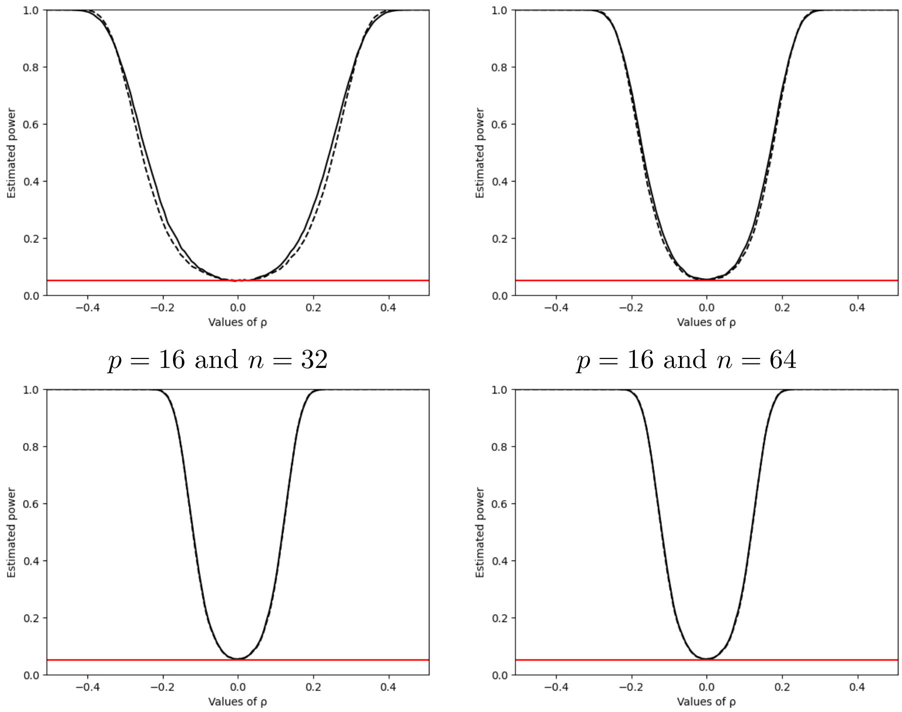
and a diagonal covariance matrix. We considered a mixture of two Gamma distributions with probability density function given in (9). The four parameters in (9) are obtained by matching the first four exact moments, which expressions are provided in Appendix B. We could also have considered a single Gamma distribution with probability density function given in (8), matching the first two exact moments of T, but the results would be indistinguishable from those presented in Figures 7–10.

In Figures 7–10, we present histograms constructed from the simulated data for different choices of p and n, together with the probability density function in (9).

In Figure 7, for low dimension scenarios, $n > p$, we can see that the mixture of two Gamma distributions has a perfect fit to the data. However, in high dimensional scenarios, mainly for small values of n, as in Figure 8, the fit exhibits a slight deviation, which appears to be corrected in Figures 9 and 10 as the sample size increases while remaining smaller than the number of variables. The mixture of Gamma distributions approximation is quite simple to use in practice, and Figures 7–10 show that the proposed approximation can be applied in both low- and high-dimensional scenarios. However, in high-dimensional settings, the sample size should be greater than 30.

In addition, to assess the precision of the approximations proposed and also to compared them with other existing models, we simulated 100 000 values of the statistic T defined in (7) and computed the empirical $\alpha = 0.95$ quantile, denoted by $\widetilde{q}_\alpha$. We then evaluated the approximate cumulative distribution function at this empirical quantile and compute the absolute difference between the obtained value and 0.95. Specifically, we considered the measure

$$\Delta = |F^*(\widetilde{q}_\alpha) - \alpha|$$



$p = 16$ and $n = 32$ $\quad$ $p = 16$ and $n = 64$

$p = 16$ and $n = 128$ $\quad$ $p = 16$ and $n = 256$

Figure 3. Estimated power for the complete independence test, for data simulated from a population with a bT covariance structure, in low-dimension scenarios using the likelihood ratio statistic in (2) (dashed line) and the proposed statistic in (6) (solid line)

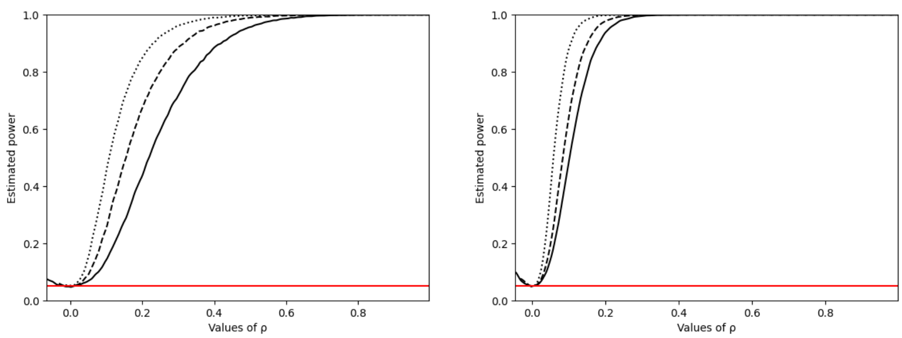


Figure 4. Estimated power for the complete independence test in high-dimension scenarios using the proposed statistic in (6), the CS covariance structure and for $n = 8$ and $p = 16$ (solid line), $p = 32$ (dashed line) and $p = 64$ (dotted line) (figure on the left), and for $n = 20$ and $p = 22$ (solid line), $p = 32$ (dashed line) and $p = 64$ (dotted line) (figure on the right)

where $F^*(.)$ denotes a given cumulative distribution function and $\widetilde{q}_\alpha$ the α quantile. We considered the following models: i) a Gamma distribution (GD), matching 2 of the exact moments, ii) a mixture of two Gamma distributions (M2G), matching 4 exact moments, iii) a shifted Weibull distribution (Wei), matching 3 exact moments, iv) a Log-normal distribution (LN), matching 2 exact moments and v) the Normal distribution (Norm) proposed in [19]. For the case of the Normal approximation proposed in

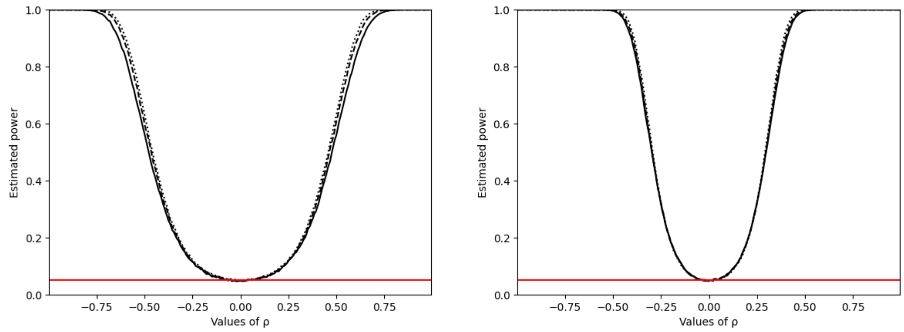


Figure 5. Estimated power for the complete independence test in high-dimension scenarios using the proposed statistic in (6), the AR covariance structure and for $n = 8$ and $p = 16$ (solid line), $p = 32$ (dashed line) and $p = 64$ (dotted line) (figure on the left), and for $n = 20$ and $p = 22$ (solid line), $p = 32$ (dashed line) and $p = 64$ (dotted line) (figure on the right)

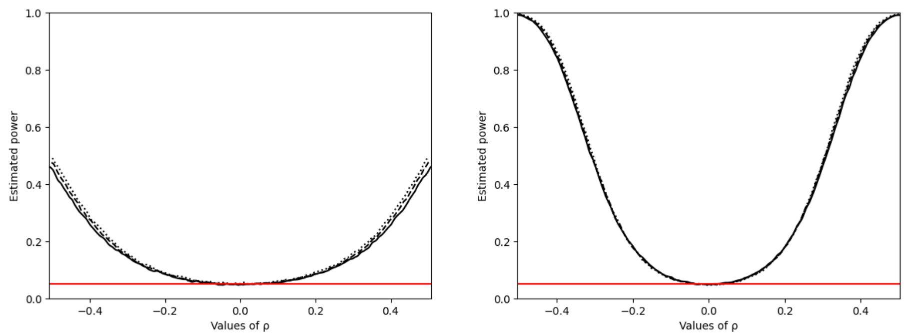


Figure 6. Estimated power for the complete independence test in high-dimension scenarios using the proposed statistic in (6), the bT covariance structure and for $n = 8$ and $p = 16$ (solid line), $p = 32$ (dashed line) and $p = 64$ (dotted line) (figure on the left), and for $n = 20$ and $p = 22$ (solid line), $p = 32$ (dashed line) and $p = 64$ (dotted line) (figure on the right)

[19], since the proposed statistic t_{mn} in [19] is different from the one proposed in this work, we have also simulated 100 000 values of the statistic t_{mn} in [19] and, following the same process, computed the empirical $\alpha = 0.95$ quantile and evaluated its value in the cumulative distribution function of the Normal distribution proposed in [19]. From Table 1 we may see that the Gamma distribution, in general, has the lowest value, and the results are identical to the ones obtained with a mixture of two Gamma distributions. The Normal distribution proposed in [19] presents improved values of Δ in high dimensional scenarios, namely for $n = 20$.

4. Application to High-Dimensional Financial Asset Returns

To demonstrate the applicability of the proposed test in empirical finance under a controlled high-dimensional setting ($p > n$), we evaluate our test statistic using industry portfolio returns from the Fama-French data library available in the following link. In financial economics, testing the hypothesis of complete independence across asset classes or sectors is fundamental to evaluating market integration, risk diversification, and factor-model specification. Under the complete independence assumption, the population covariance matrix of returns is diagonal, implying that asset-specific risks are uncorrelated and that there are no systemic cross-sectional risk drivers.

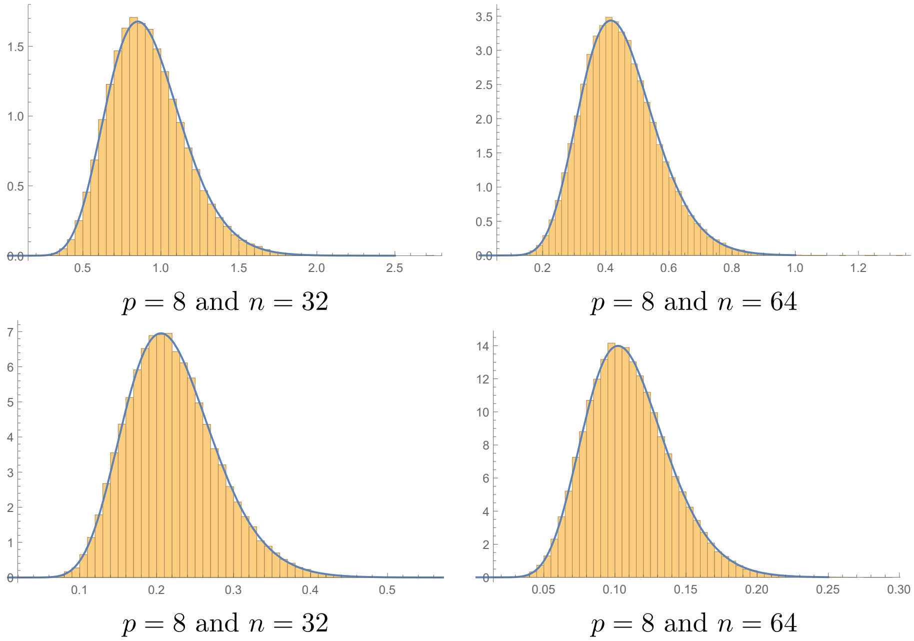


Figure 7. Histograms of the simulated data for the statistic T in (7) under the null hypothesis of complete independence in low-dimensional scenarios, together with the probability density function of the mixture of two Gamma distributions in (9)

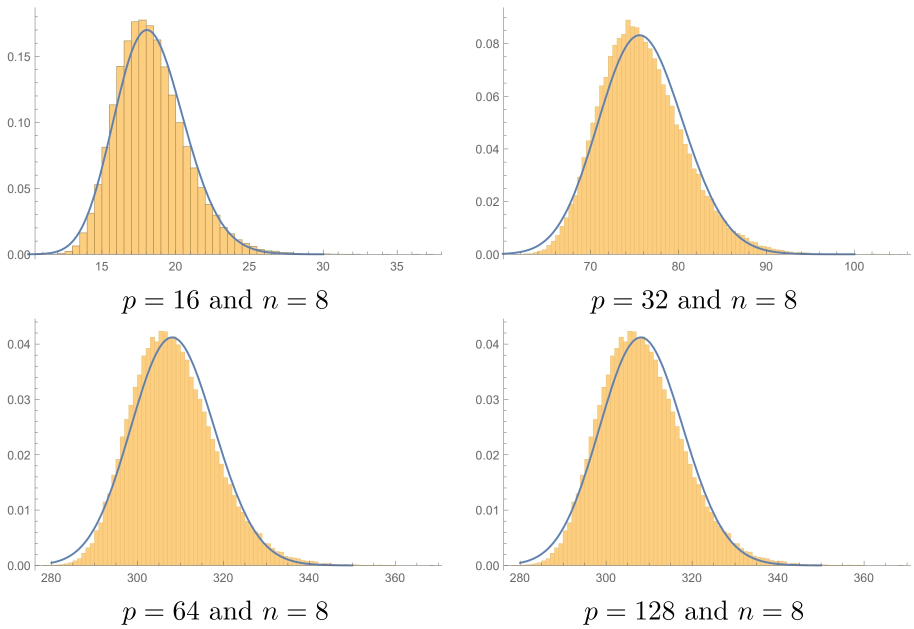


Figure 8. Histograms of the simulated data for the statistic T in (7) under the null hypothesis of complete independence in high-dimensional scenarios, together with the probability density function of the mixture of two Gamma distributions in (9)

We analyze daily returns from the Fama-French $p = 48$ Industry Portfolios. To construct a high-dimensional scenario where the number of variables exceeds the sample size, we select a recent window

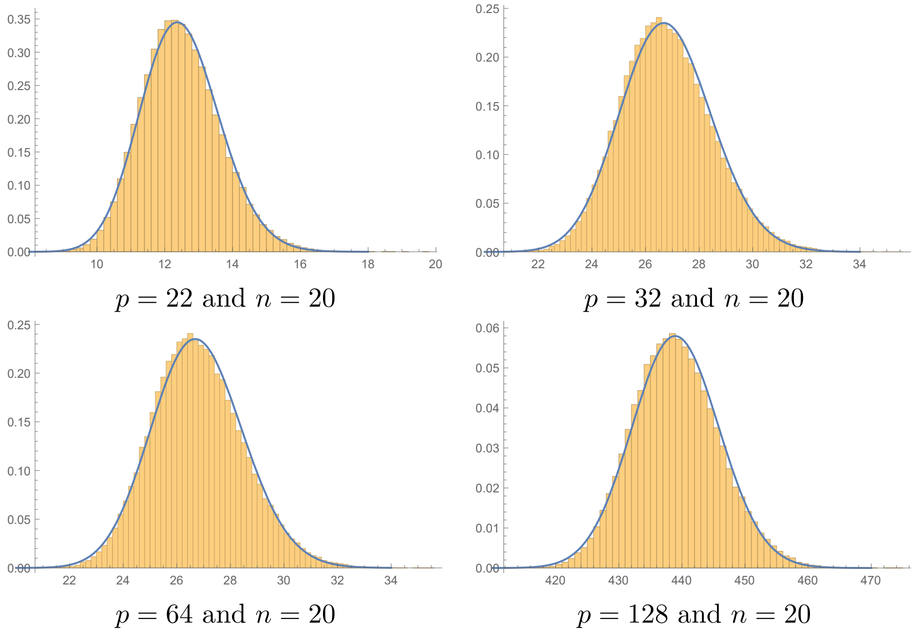


Figure 9. Histograms of the simulated data for the statistic T in (7) under the null hypothesis of complete independence in high-dimensional scenarios, together with the probability density function of the mixture of two Gamma distributions in (9)

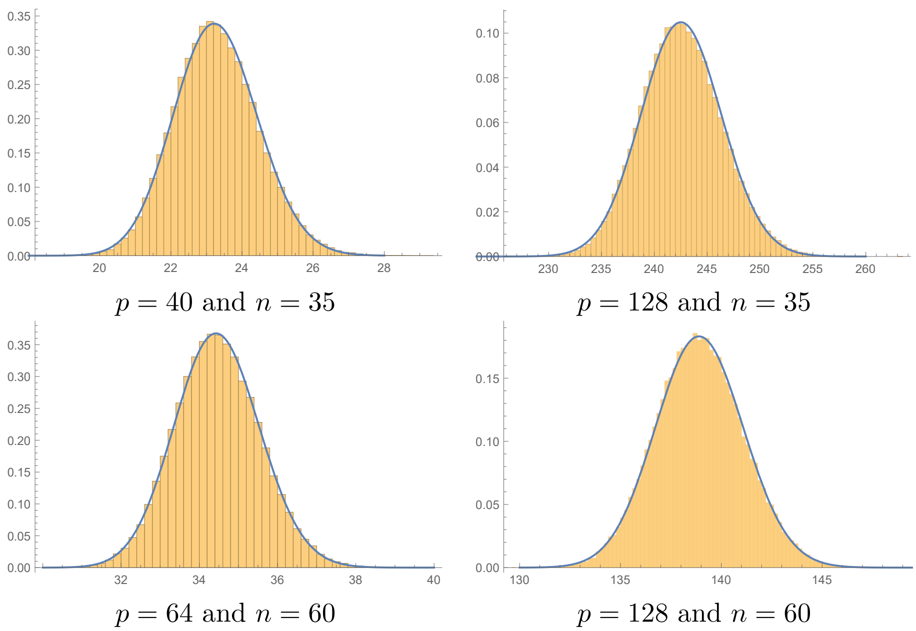


Figure 10. Histograms of the simulated data for the statistic T in (7) under the null hypothesis of complete independence in high-dimensional scenarios, together with the probability density function of the mixture of two Gamma distributions in (9)

of $n = 30$ consecutive trading days. Let $R_{i,t}$ denote the daily return of industry i at time t, where $i = 1, \dots, 48$ and $t = 1, \dots, 30$. The return series are standardized, and we compute the $p \times p$ sample

Table 1. Values of the measure Δ for the five models considered (GD, SGD, Wei, LN, Norm) and different choices of the values of n and p.

Scenarios	GD	M2G	Wei	LN	Norm
$p=8,\ n=32$	3.3×10^{-4}	3.4×10^{-4}	9.6×10^{-4}	2.9×10^{-3}	1.2×10^{-2}
$p=8,\ n=64$	5.6×10^{-4}	5.6×10^{-4}	7.2×10^{-4}	2.7×10^{-3}	1.3×10^{-2}
$p=8,\ n=128$	1.1×10^{-3}	1.1×10^{-3}	1.1×10^{-4}	2.2×10^{-3}	1.2×10^{-2}
$p=8,\ n=256$	6.7×10^{-4}	6.7×10^{-4}	2.0×10^{-3}	3.8×10^{-3}	1.3×10^{-2}
$p=16,\ n=8$	5.5×10^{-3}	5.6×10^{-3}	5.7×10^{-3}	2.8×10^{-3}	1.1×10^{-2}
$p=32,\ n=8$	8.2×10^{-3}	8.2×10^{-3}	9.1×10^{-3}	6.6×10^{-3}	1.1×10^{-2}
$p=64,\ n=8$	9.2×10^{-3}	9.2×10^{-3}	1.0×10^{-2}	8.4×10^{-3}	1.1×10^{-2}
$p=128,\ n=8$	1.1×10^{-2}	1.1×10^{-2}	1.2×10^{-2}	1.1×10^{-2}	1.1×10^{-2}
$p=22,\ n=20$	3.6×10^{-3}	3.6×10^{-3}	3.9×10^{-3}	8.7×10^{-3}	5.6×10^{-3}
$p=32,\ n=20$	4.4×10^{-3}	4.4×10^{-3}	5.0×10^{-3}	6.7×10^{-3}	3.6×10^{-3}
$p=64,\ n=20$	5.3×10^{-3}	5.3×10^{-3}	6.2×10^{-3}	5.6×10^{-3}	5.6×10^{-3}
$p=128,\ n=20$	5.3×10^{-3}	5.3×10^{-3}	6.2×10^{-3}	5.6×10^{-3}	5.6×10^{-3}

correlation matrix $R=(r_{ij})$. Because $p=48>n=30$, the sample correlation matrix is singular with rank at most $n-1=29$, rendering classical likelihood ratio tests inapplicable.

We test the null hypothesis of complete independence, $H_0:\sigma_{ij}=0$ for all $i\neq j$, against the alternative $H_1:\sigma_{ij}\neq 0$ for some $i\neq j$, where σ_{ij} is the population covariance.

We evaluate the test statistic T defined in (7) for $p=48$, the sum evaluates across $m=\frac{p(p-1)}{2}=1\,128$ distinct industry pair correlations. For this data set the observed value of the test statistic T in (7) was equal to 202.821 and with $p-value<0.001$.

The empirical test statistic overwhelmingly rejects the null hypothesis of complete independence ($p-value<0.001$). The observed value of T drastically exceeds its expected theoretical value under H_0, reflecting strong dependence effect.

The python code for for the implementation of the test and for the replication of this application is provided in the file `Application_GIT.ipynb` available in the Github Repository CompleteIndependence.

5. Conclusions

In this work, we propose a new test for complete independence among a set of variables, based on the intersection of null hypotheses of pairwise independence and on their likelihood ratio test statistics. The distribution of the test statistic can be represented as a product of Beta random variables and is bounded in the interval (0,1), with values close to zero providing evidence against the null hypothesis of complete independence.

The simulation studies on power and bias demonstrate that the proposed test performs well. In particular, it behaves as an unbiased test for samples generated from a multivariate normal distribution with a CS, AR and bT covariance structures. Moreover, in low-dimensional settings, the power obtained using the proposed statistic appears to converge to that achieved by the likelihood ratio test for complete independence. In high-dimensional settings, the test exhibits sensitivity to departures from the null hypothesis, with a natural convergence of the power to one as the magnitude of the departure increases.

The histograms presented in the numerical studies show a reasonable fit between the empirical distribution of T in (7) and the probability density function of a mixture of two gamma distributions matching the first four exact moments in (9), for data simulated under the null hypothesis of complete independence. We note that, in high dimensions, the sample size n should be at least 30 to achieve a reasonable fit. The results suggest that the approximation is accurate and sufficiently simple to be used in practical applications. Based on the similarity of the results displayed in Table 1 for the single Gamma distribution and the mixture of two Gamma distributions, for the sake of simplicity and practical use, we suggest using a single Gamma distribution.

Since the statistic proposed in this work, in (6) and (7), referred to as an almost likelihood ratio statistic, is bounded in the interval $(0, 1)$ and is based on likelihood ratio test statistics, it can be naturally combined with classical likelihood ratio statistics, providing a flexible framework that opens the way for the development of methods for testing more complex covariance structures in high-dimensional scenarios. A natural example is the test for block independence, which is used to test the independence of groups of variables. In this test, under the null hypothesis, it is assumed that the covariance matrix is block diagonal, so the objective is to test whether all off-block-diagonal matrices are null. This may be achieved using a similar approach to the one used in this work, where the test statistic would be constructed based on the product of the likelihood ratio statistics used to test the nullity of the off-block-diagonal matrices. Future work may also include simulations to investigate the properties of the test under departures from the multivariate normality assumption and when considering more complex structures for the covariance matrix.

The results presented in this paper are easy to implement and apply in practice, and therefore require very little computational effort. The code necessary for the implementation of the proposed test and for the computation of the associated $p-value$ is made available in the file `Application_GIT.ipynb` which can be used to replicate the application provided in this paper. The file is available in the Github Repository CompleteIndependence

Acknowledgement

This work is funded by national funds through the FCT – Fundação para a Ciência e a Tecnologia, I.P., under the scope of the projects UID/00297/2025 (https://doi.org/10.54499/UID/00297/2025), UID/PRR/00297/2025 (https://doi.org/10.54499/UID/PRR/00297/2025), and UID/PRR2/00297/2025 (https://doi.org/10.54499/UID/PRR2/00297/2025) (Center for Mathematics and Applications – NOVA Math)”

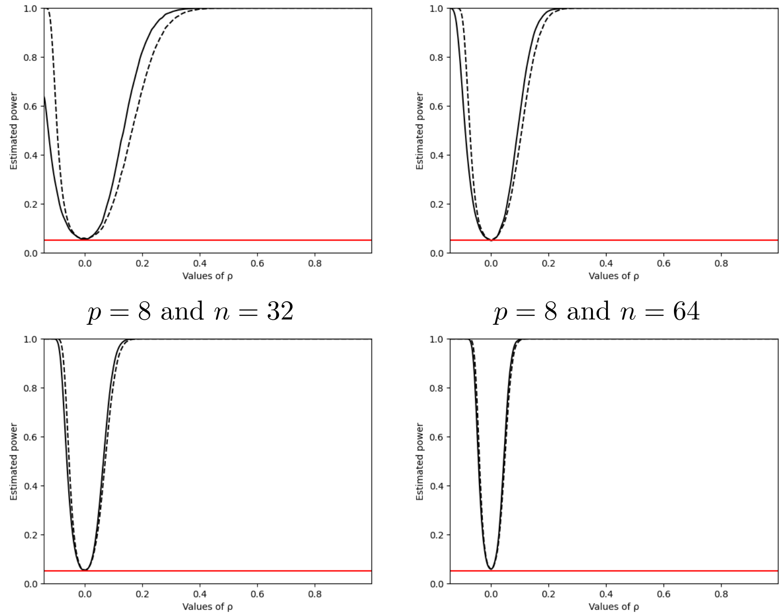
A. Plots of the estimated power for $p = 8$

B. First four exact null moments of T in (7)

Under the assumption of independence of the sample correlations, the first four null moments of T in (7) are given, respectively, by

$$E[T] = m\left[\psi\left(\frac{N-1}{2}\right) - \psi\left(\frac{N-2}{2}\right)\right]; \tag{13}$$

$$E[T^2] = m\left[\psi'\left(\frac{N-2}{2}\right) - \psi'\left(\frac{N-1}{2}\right)\right] + m^2\left[\psi\left(\frac{N-1}{2}\right) - \psi\left(\frac{N-2}{2}\right)\right]^2; \tag{14}$$



$p = 8$ and $n = 32$ $p = 8$ and $n = 64$

$p = 8$ and $n = 128$ $p = 8$ and $n = 256$

Figure 11. Estimated power for the complete independence test, for data simulated from a population with a CS covariance structure, in low-dimension scenarios using the likelihood ratio statistic in (2) (dashed line) and the proposed statistic in (6) (solid line)

$$\begin{aligned}
\mathbb{E}[T^3] &= m\left[\psi^{(2)}\left(\frac{N-1}{2}\right) - \psi^{(2)}\left(\frac{N-2}{2}\right)\right] \\
&\quad +3m^2\left[\psi^{(1)}\left(\frac{N-2}{2}\right) - \psi^{(1)}\left(\frac{N-1}{2}\right)\right]\left[\psi\left(\frac{N-1}{2}\right) - \psi\left(\frac{N-2}{2}\right)\right] \\
&\quad +m^3\left[\psi\left(\frac{N-1}{2}\right) - \psi\left(\frac{N-2}{2}\right)\right]^3;
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[T^4] &= m\left[\psi^{(3)}\left(\frac{N-2}{2}\right) - \psi^{(3)}\left(\frac{N-1}{2}\right)\right] \\
&\quad +4m^2\left[\psi^{(2)}\left(\frac{N-1}{2}\right) - \psi^{(2)}\left(\frac{N-2}{2}\right)\right]\left[\psi\left(\frac{N-1}{2}\right) - \psi\left(\frac{N-2}{2}\right)\right] \\
&\quad +3m^2\left[\psi^{(1)}\left(\frac{N-2}{2}\right) - \psi^{(1)}\left(\frac{N-1}{2}\right)\right]^2 \\
&\quad +6m^3\left[\psi^{(1)}\left(\frac{N-2}{2}\right) - \psi^{(1)}\left(\frac{N-1}{2}\right)\right]\left[\psi\left(\frac{N-1}{2}\right) - \psi\left(\frac{N-2}{2}\right)\right]^2 \\
&\quad +m^4\left[\psi\left(\frac{N-1}{2}\right) - \psi\left(\frac{N-2}{2}\right)\right]^4
\end{aligned}$$

where $m = \frac{p(p-1)}{2}$, and where $\psi^{(k)}(\cdot)$ is the polygamma function of order k.

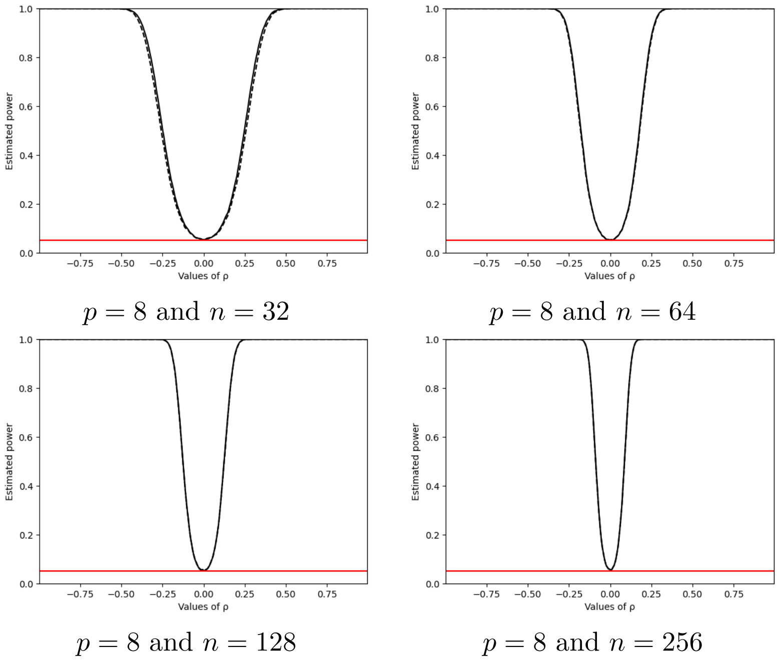


Figure 12. Estimated power for the complete independence test, for data simulated from a population with a AR covariance structure, in low-dimension scenarios using the likelihood ratio statistic in (2) (dashed line) and the proposed statistic in (6) (solid line)

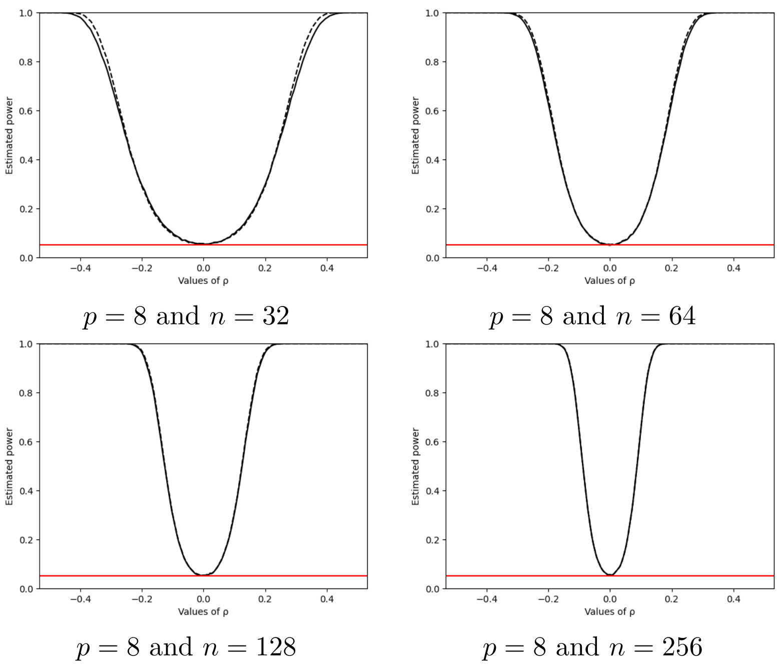


Figure 13. Estimated power for the complete independence test, for data simulated from a population with a bT covariance structure, in low-dimension scenarios using the likelihood ratio statistic in (2) (dashed line) and the proposed statistic in (6) (solid line)

REFERENCES

1. T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, J. Wiley & Sons, New York, 2003.
2. P. J. Bickel, and D. A. Freedman, *Some asymptotic theory for the bootstrap*, The Annals of Statistics, vol. 9, no. 6, pp. 1196–1217, 1981.
3. J. R. Blum, J. Kiefer, and M. Rosenblatt, *Distribution-free tests of independence based on the sample distribution function*, The Annals of Mathematical Statistics, vol. 32, no. 2, pp. 485–498, 1961.
4. S. Chang, and Y. Qi, *On tests for complete independence of normal random vectors*, arXiv preprint, 2017.
5. S. Chen, and G. S. Mudholkar, *Null distribution of the sum of squared z-transforms in testing complete independence*, Ann Inst Stat Math, vol. 42, pp. 149–155, 1990.
6. C. A. Coelho, *The Generalized Near-Integer Gamma Distribution: A Basis for 'Near-Exact' Approximations to the Distribution of Statistics which are the Product of an Odd Number of Independent Beta Random Variables*, Journal of Multivariate Analysis, vol. 89, pp. 191–218, 2004.
7. C. A. Coelho, and F. J. Marques, *Near-exact distributions for the independence and sphericity likelihood ratio test statistics*, Journal of Multivariate Analysis, vol. 101, pp. 583–593, 2010.
8. E. F. Fama and K. R. French, *Industry costs of equity*, Journal of Financial Economics, vol. 43, no. 2, pp. 153–193, 1997.
9. T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, *et al.*, *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring*, Science, vol. 286, no. 5439, pp. 531–537, 1999.
10. A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. Smola, *A kernel statistical test of independence*, Advances in Neural Information Processing Systems, vol. 20, pp. 585–592, 2008.
11. D. He, H. Liu, K. Xu, and M. Cao, *Generalized Schott type tests for complete independence in high dimensions*, Journal of Multivariate Analysis, vol. 183, pp. 104731, 2021.
12. W. Hoeffding, *A non-parametric test of independence*, The Annals of Mathematical Statistics, vol. 19, no. 4, pp. 546–557, 1948.
13. K. O. Kymn, *The Distribution of the Sample Correlation Coefficient Under the Null Hypothesis*, Econometrica, vol. 36, no. 1, pp. 187–189, 1968.
14. F. J. Marques, C. A. Coelho, and B. C. Arnold, *A general near-exact distribution theory for the most common likelihood ratio test statistics used in Multivariate Analysis*, Test, vol. 20, pp. 180–203, 2011.
15. F. J. Marques, J. Diogo, M. Norouzirad, and R. Bispo, *Testing the independence of variables for specific covariance structures: A simulation study*, Math. Meth. Appl. Sci., vol. 46, pp. 10421–10434, 2023.
16. G. Mao, *A new test of independence for high-dimensional data*, Statistics & Probability Letters, vol. 93, pp. 14–18, 2014.
17. R. J. Muirhead, *Aspects of Multivariate Statistical Theory*, John Wiley & Sons, New York, 1982.
18. J. P. Romano, *Bootstrap and randomization tests of some nonparametric hypotheses*, The Annals of Statistics, vol. 17, no. 1, pp. 141–159, 1989.
19. J. R. Schott, *Testing for Complete Independence in High Dimensions*, Biometrika, vol. 92, no. 4, pp. 951–956, 2005.
20. M. S. Srivastava, *Some tests concerning the covariance matrix in high dimensional data*, J. Japan Statist. Soc., vol. 35, no. 2, pp. 251–272, 2005.
21. G. J. Székely, and M. L. Rizzo, *Brownian distance covariance*, The Annals of Applied Statistics, vol. 3, no. 4, pp. 1236–1265, 2009.
22. S. S. Wilks, *The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses*, The Annals of Mathematical Statistics, vol. 9, pp. 60–62, 1938.