Statistics and Probability Letters 193 (2023) 109732

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

A Bartlett-type correction for likelihood ratio tests with application to testing equality of Gaussian graphical models

Erika Banzato [a], Monica Chiogna [b], Vera Djordjilović [c], Davide Risso [a,*]

[a] Department of Statistical Sciences, University of Padua, via C. Battisti 241, Padua, Italy
[b] Department of Statistical Sciences, University of Bologna, Via Belle Arti, 41, Bologna, Italy
[c] Department of Economics, Ca' Foscari University of Venice, Cannaregio 873, Venice, Italy

ARTICLE INFO

Article history:
Received 22 April 2022
Received in revised form 12 September 2022
Accepted 24 October 2022
Available online 9 November 2022

MSC:
0000
1111

Keywords:
Likelihood ratio test
Hypothesis test
Multivariate normal distribution

ABSTRACT

This work defines a new correction for the likelihood ratio test for a two-sample problem within the multivariate normal context. This correction applies to decomposable graphical models, where testing equality of distributions can be decomposed into lower dimensional problems.

1. Introduction

Testing the equality of distributions in a two sample problem can conveniently be done resorting to the likelihood ratio test (LRT) statistic, $W_n = -2\log\Lambda_n$, where Λ_n is the likelihood ratio. In Wilks (1938), it is shown that for samples coming from p-variate normal distributions, W_n is asymptotically distributed as a chi-square with $f = p(p+3)/2$ degrees of freedom. It is well known (Muirhead, 1982) that the quality of the asymptotic approximation might be poor in finite sample problems, even at moderate sample sizes. However, convergence to the asymptotic distribution can be improved by multiplying the LRT statistic by a constant (Van der Vaart, 1998). Under the low-dimensional setting, where the number of variables p is considered fixed and n is large, the correction factor ρ proposed in Muirhead (1982) improves the convergence rate, but when the value of p is close to n or increases with it, this correction is unable to provide an improvement. In the high-dimensional setting, where p is assumed to increase with n, Jiang and Qi (2015) proposed a standardization of the LRT statistic that allows to resort to the central limit theorem and, therefore, to switch to a normal approximation. This solution, however, proves to be inaccurate for small p, given the asymmetry of the LRT statistic.

In a recent work, He et al. (2021) studied the *phase transition boundary*, d in what follows, which characterizes the approximation accuracy by establishing the necessary and sufficient condition for the chi-square approximation to hold when p increases with n. The authors showed that the chi-square approximation holds if and only if $p/n^d \to 0$, with $d = 1/2$ for the raw LRT statistic and $d = 2/3$ for its ρ-corrected version.

In this paper, we propose a new multiplicative correction factor, δ_n hereafter, defined to be the ratio between the degrees of freedom of the asymptotic chi-square approximation and an approximation of the expected value of the LRT

* Corresponding author.
E-mail address: davide.risso@unipd.it (D. Risso).

https://doi.org/10.1016/j.spl.2022.109732

statistic, under the null hypothesis, as a function of p and n. We prove that its phase transition boundary d is equal to 1, so that the chi-square approximation holds in all situations in which $p/n \to 0$. We show the usefulness of our proposal in the context of Gaussian graphical models (GGM). Here, the problem of testing equality of two distributions Markov with respect to a decomposable graph can be broken up into testing equality of lower dimensional Gaussian distributions. According to the structure of the graph, these lower dimensional problems can lead to very different values of the p/n ratio. Hence, it becomes crucial to rely on an approximation that guarantees a good finite sample accuracy even in extreme cases, where p is close to n.

2. A quick tour of the state of art

Consider two p-dimensional multivariate normal distributions, $N_p(\mu^{(j)}, \Sigma^{(j)})$, $j = 1, 2$, and the problem of testing their equality based on two independent random samples of size n_j. In detail, consider the hypothesis of equality of distributions

$$H_0 : \mu^{(1)} = \mu^{(2)},\ \Sigma^{(1)} = \Sigma^{(2)} \quad \text{vs.} \quad H_a : H_0 \quad \text{is not true.} \tag{1}$$

The LRT for testing (1), derived in Wilks (1938), can be written as

$$\Lambda_n = \frac{\prod_{j=1}^{2} \det\left(\hat{\Sigma}^{(j)}\right)^{n_j/2}}{\det\left(\hat{\Sigma}\right)^{n/2}},$$

where $n = n_1 + n_2$, $\hat{\Sigma}$ and $\hat{\Sigma}^{(j)}$, $j = 1, 2$ are the maximum likelihood estimates of the covariance matrices under the null and alternative hypotheses, respectively, and $\det(\hat{\Sigma})$ denotes the determinant of $\hat{\Sigma}$. Under the null hypothesis in (1), the LRT statistic $W_n = -2 \log \Lambda_n$, has an asymptotic chi-square distribution, with $f = p(p+3)/2$ degrees of freedom.

In settings where p is fixed and n is allowed to grow, a first correction of the statistic W_n was proposed by Bartlett (1937), based on a rescaling aimed at making its mean exactly equal to the mean of the asymptotic chi-square distribution, i.e., equal to f. The corrected statistic, W_n^B say, takes the following form

$$W_n^B = \frac{f}{E_{H_0}(W_n)} W_n, \tag{2}$$

where $E_{H_0}(W_n)$ is the expected value of W_n under the null hypothesis; see for example Van der Vaart (1998). Later, Muirhead (1982) proposed a version of Bartlett correction that leverages on an expansion of the correction factor, leading to the following correction

$$\rho = 1 - \frac{2p^2 + 9p + 11}{6(p+3)n} \left(\sum_{j=1}^{2} \frac{n}{n_j} - 1 \right). \tag{3}$$

The author showed that the resulting corrected statistic, W_n^ρ say, where $W_n^\rho = -2\rho \log \Lambda_n$, has a chi-square limit, with an improved approximation rate with respect to W_n. Both corrections, however, fail when p and n grow at comparable rates.

Recent studies have considered the problem when the dimension p changes with the sample size n. In these settings, Jiang and Yang (2013) and Jiang and Qi (2015) established the following result based on the central limit theorem (CLT):

$$\frac{\log \Lambda_n - \mu_n}{n\sigma_n} \xrightarrow{d} N(0, 1), \tag{4}$$

where μ_n and $\sigma_n > 0$ are functions of both n and p and are the asymptotic mean and standard deviation of $\log \Lambda_n$, respectively. The use of the central limit theorem has the advantage of being appropriate in a high dimensional setting; however, it is less accurate when p is small, due to the asymmetric shape of the LRT distribution.

3. Our proposal

In this section, we propose a Bartlett-type correction of the LRT statistic, under the assumption that p changes with the sample size n. This correction replaces the denominator of (2) with a function of the approximated mean given in Eq. (4). In a two sample problem, the term μ_n defined by Jiang and Qi (2015) is

$$\mu_n = \frac{1}{4} \left[-4p - \sum_{j=1}^{2} \frac{p}{n_j} + n r_n^2 (2p - 2n + 3) - \sum_{j=1}^{2} n_j r_{n_j'}^2 (2p - 2n_j + 3) \right], \tag{5}$$

where $n_j' = n_j - 1$ and $r_x = (-\log(1 - p/x))^{1/2}$, for $x > p$, and $n = n_1 + n_2$. Let $\mu_{w_n} = -2\mu_n$, we define the adjusted statistic T_n as

$$T_n = \delta_n W_n, \quad \delta_n = \frac{f}{\mu_{w_n}}, \tag{6}$$

where $f = p(p+3)/2$ are the degrees of freedom of the chi-square asymptotic null distribution of W_n. We now prove that T_n is asymptotically chi-square distributed.

Theorem 1. *Let $\mathbf{p} = (p_n)_{n\in N}$ be a sequence of integers $1 \le p_n < n_j - 1$. Under H_0, for T_n defined as in (6), $\min_{j=1,2} n_j \to \infty$ and $p/n \to 0$, we have that*

$$\sup_{-\infty<x<\infty} |P(T_n < x) - P(\chi^2_{f_n} < x)| \to 0$$

and the phase transition boundary of T_n is $d = 1$.

Proof. See Appendix A. □

In Theorem 1, the condition $n_j > p+1$ is assumed to ensure the existence of the LRT. Moreover, the condition $p/n \to 0$ defines the phase transition of the adjusted statistic, as introduced in He et al. (2021), which represents the boundary in which the chi-square approximation starts to fail as p increases and characterizes the approximation accuracy. This boundary is an improvement over W_n and W_n^ρ, whose approximations hold for $p/n^d \to 0$, with $d = 1/2$ and $d = 2/3$, respectively.

4. Simulation study

In this section we present a simulation study to compare the performances of the LRT statistics based on four different approximations: the classic chi-square approximation, the ρ-adjusted approach of Muirhead (1982), the CLT approach of Jiang and Qi (2015) and our proposed δ-adjusted approach.

We study how the correction acts considering a fixed sample size and letting the dimension p change. Data are drawn from a multivariate normal distribution, with fixed covariance matrix and mean vector and we set $n_1 = n_2 = 50$ and $p = 2, 30, 40$. For each scenario, five thousand simulations are run. Results are shown in Fig. 1. For each value of p we plot the histograms of the empirical distribution of the four statistics, namely W_n, W_n^ρ, T_n and W_n^{clt}, and compare them with the chi-square distribution with $p(p+3)/2$ degrees of freedom in the first three cases and a standard normal in the last case. The top row of Fig. 1 shows how the statistic W_n departs from the theoretical χ^2 distribution as p grows. This is expected and motivates the need of an adjustment when dealing with testing problems in which the dimension grows with n. In fact, if 50 observations might be enough for testing a problem of dimension 2, this is not the case for other values of p, especially when p and n have comparable values. The second row shows the results for the statistic corrected with ρ. Note that, also in this case, the approximation to the χ^2 fails as p approaches the group sample size, n_j. With respect to the previous case, however, the departure from the chi-square distribution occurs for higher values of p. The third row highlights the problem of applying the CLT when p is small. For example, when $p = 2$ the approximation to the normal distribution fails, while it improves as p increases. This approach works well also for values of p very close to n_j. The bottom row shows the accuracy of the approximation of the proposed adjusted statistic T_n. Note that this correction leads to a good approximation regardless of the dimension of the testing problem, as long as $p/n \to 0$, and could be used as a unique tool for correcting W_n at different values of p and n.

Finally, we run some simulations to examine the phase transition boundary in Theorem 1, under the null hypothesis. We consider $p = \lfloor n_1^\varepsilon \rfloor$, $n_1 = n_2$, $n = \sum_{j=1}^2 n_j$ and $n_j \in \{100, 500, 1000\}$ and finally $\varepsilon \in \{6/24, \ldots, 23/24, 23.5/24\}$. $\lfloor\cdot\rfloor$ denotes the rounding to the nearest integer function. We plot the empirical type-I error rate (over 1000 simulations) versus ε, for each chi-square approximation: W_n, W_n^ρ and T_n. Results are plotted in Fig. 2. The first two panels confirm the results in He et al. (2021), while the one on the right hand side shows how the phase transition boundary of the adjusted statistic T_n is close to 1. The particular case with ε exactly equal to one is excluded, to ensure the identifiability of the covariance matrix.

5. Testing equality of distributions in Gaussian graphical models

In the remaining sections of the paper, we assume the reader is familiar with the basic theory of (decomposable) undirected graphical models, as presented for instance in Lauritzen (1996); see also Whittaker (1990). We adopt a standard terminology and a rather intuitive notation: we let $G = (V, E)$ denote an undirected graph, with V a finite set of nodes and $E = \{(v, t) : v \neq t;\ v, t \in V\}$ a finite set of edges between vertices. We denote its cliques, separators and residuals by C, S and R, respectively.

Our proposal finds a natural application in the context of decomposable graphical models. One prominent advantage of decomposable graphs is that their cliques can be arranged so as to satisfy the running intersection property (RIP), and the joint probability distribution of the associated random vectors factorizes accordingly. In detail, if a graph $G = (V, E)$

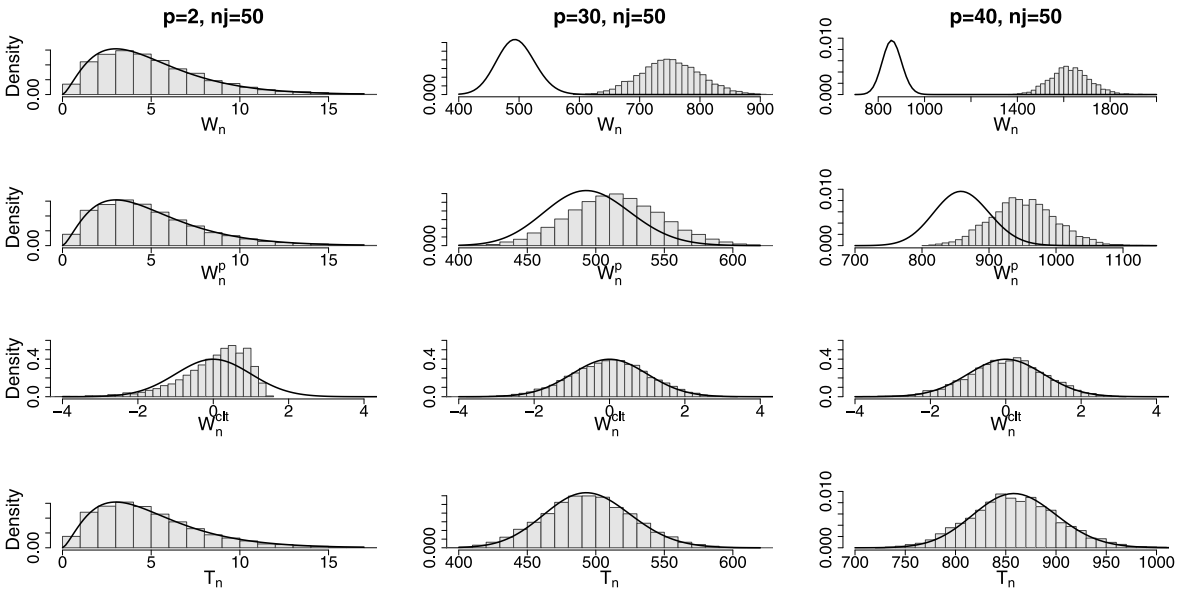

Fig. 1. Simulation results with $n_1 = n_2 = 50$ and $p = 2, 30, 40$. From the top to the bottom row: empirical distribution of W_n, W_n^ρ, W_n^{clt}, and T_n. The solid line in the first, second, and fourth rows shows the nominal χ^2 distribution, with 5, 495 and 860 degrees of freedom (from left to right) respectively. The solid line in the third row, corresponding to the W_n^{clt} statistic, shows the standard normal distribution.

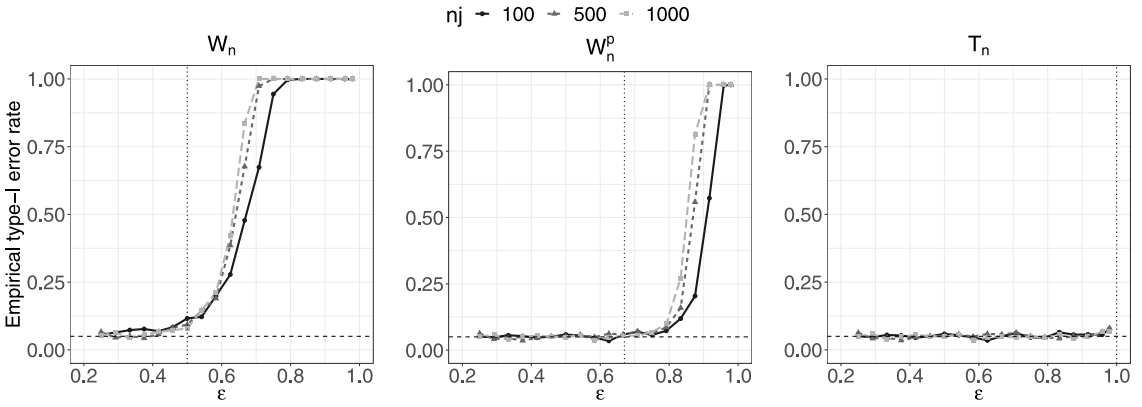

Fig. 2. Chi-square approximation of W_n, W_n^ρ and T_n. Empirical type-I error rate for $n_j \in \{100, 500, 1000\}$, $j = 1, 2$ over 1000 simulations. The vertical dotted lines represents the phase transition boundaries for the three statistics: 1/2, 2/3 and 1, respectively. The horizontal dashed line represents the nominal significance level, 0.05.

decomposes into k, say, cliques, let C_i, $i = 1, \ldots, k$, be a sequence of cliques satisfying the RIP and $S_i = C_i \cap C_{i-1}$ and $R_i = C_i \setminus C_{i-1}$, $i = 2, \ldots, k$ the set of corresponding separators and residuals, respectively. Then, the probability distribution of the random vector X_V factorizes as $f(X_V) = f(X_{C_1}) f(X_{R_2}|X_{S_2}) \ldots f(X_{R_k}|X_{S_k})$. See Lauritzen (1996) for an exhaustive explanation. Such factorization renders tractable inference in the setting of large-scale graphical models, where the dimension p of the problem is higher that the available sample size n. Even when $p < n$, using the information on the graphical structure allows us both to improve the power of detecting a difference between the two distributions under study (the size of the model is reduced by constraints on the covariance matrix), and to localize that difference, thanks to the modular nature of graphical models (Djordjilović and Chiogna, 2022). This potential has fed the increasing prominence of graph–theoretic representations of probability distributions in fields such as statistical and quantum physics, bioinformatics, signal processing, econometrics and information theory. In our problem setting, this factorization assumes a crucial role as it allows to decompose the global problem of testing equality of distribution in two samples into a sequence of local tests of equality of distributions defined on a smaller set of variables, as follows

$$H = \bigcap_{i=1}^{k} H_i, \quad H_i : X_{R_i}^{(1)} | X_{S_i}^{(1)} \stackrel{d}{=} X_{R_i}^{(2)} | X_{S_i}^{(2)}, \quad i = 1, \ldots, k, \tag{7}$$

with $S_1 = \varnothing$ and $R_1 = C_1$. Hence, to test the global hypothesis H, one can test the k local hypotheses $\{H_i, i = 1, \ldots, k\}$ of equality of the conditional distributions of $X_{R_i}|X_{S_i}$. In the case of strong meta Markov models (Lauritzen, 1996; Edwards, 2000), as is the Gaussian case, Djordjilović and Chiogna (2022) showed that the local hypotheses $H_i, i = 1, \ldots, k$, are independent and that the LRT statistic for testing H also decomposes into k LRT statistics, one for testing each local hypothesis. Specifically, the LRT, W_n, factorizes as

$$W_n = \sum_{i=1}^{k} \left[W_n^{C_i} - W_n^{S_i} \right] = W_n^{C_1} + \sum_{i=2}^{k} W_n^{C_i|S_i}, \tag{8}$$

where W_n^A, $A \subseteq V$, represents the LRT for the hypothesis of equality of distributions for X_A, namely $H_{(A)} : \mu_A^{(1)} = \mu_A^{(2)}$, $\Sigma_A^{(1)} = \Sigma_A^{(2)}$, while $W_n^{A|B}$ is the LRT for the hypothesis of equality of distributions for $X_A|X_B$, $B \subseteq V \setminus A$, namely $H_{(A|B)} : \mu_{(A|B)}^{(1)} = \mu_{(A|B)}^{(2)}$, $\Sigma_{(A|B)}^{(1)} = \Sigma_{(A|B)}^{(2)}$, where $\mu_{(A|B)} = \mu_A - \Sigma_{AB}\Sigma_B^{-1}\mu_B$ and $\Sigma_{(A|B)} = \Sigma_A - \Sigma_{AB}\Sigma_B^{-1}\Sigma_{BA}$. As proved in Theorem 1 of Djordjilović and Chiogna (2022), the k statistics $W_n^{C_1}$ and $W_n^{C_i|S_i}$, $i = 2, \ldots, k$, in the right-hand side of (8) are all asymptotically independent and chi-square distributed, with f_{C_1} and $f_{C_i} - f_{S_i}$, $i = 2, \ldots, k$, degrees of freedom, respectively, being f_{C_i} and f_{S_i} the degrees of freedom associated to the marginal test on the cliques and the separators, respectively. It is worth noting that, since $W_n^{A|B} = W_n^A - W_n^B$, the only quantities needed to compute W_n are the observed values of the LRT on the marginal distributions defined over cliques and separators. It is easy to see that

$$W_n^A = \sum_{j=1}^{2} n_j \log \frac{\det(\hat{\Sigma}_A)}{\det(\hat{\Sigma}_A^{(j)})} \tag{9}$$

for $A \in \{C_1, \ldots, C_k, S_1, \ldots, S_k\}$. Here, $\hat{\Sigma}_A$ is the maximum likelihood estimate of Σ_A, the block submatrix corresponding to the nodes in A in the null covariance matrix $\Sigma = \Sigma^{(1)} = \Sigma^{(2)}$; and $\hat{\Sigma}_A^{(j)}$ are the maximum likelihood estimates of $\Sigma_A^{(j)}$, the block submatrices corresponding to the nodes in A of $\Sigma^{(j)}$, $j = 1, 2$. Moreover, each W_n^A has a chi-square limit with $f_A = p_A(p_A + 3)/2$ degrees of freedom, where p_A is the cardinality of the set A. One remarkable side effect of the decomposition is that the dimension of each local problem is determined by the cardinality of the set of variables on which it is defined, so that, for a fixed sample size n, dimensionality regimes of local problems vary as a function of their cardinality. Local problems for which $p \ll n$ might coexist with problems for which $p \approx n$.

Our proposal naturally steps in this context, providing a convenient solution able to accommodate such variety of situations. The extension of our correction to the test statistics of the kind $W_n^{C|S}$ does not represent an obstacle, resulting indeed to be straightforward. In fact, being $E(W_n^{C|S}) = E(W_n^C) - E(W_n^S)$, it results $\mu_n^{C|S} = \mu_n^C - \mu_n^S$. The corrected statistics for the tests relative to the decomposition (7) simply become

$$T_n^{C_1} = \delta_n^{C_1} W_n^{C_1}, \quad \delta_n^{C_1} = \frac{f_{C_1}}{\mu_n^{C_1}} \tag{10}$$

$$T_n^{C_i|S_i} = \delta_n^{C_i|S_i} W_n^{C_i|S_i}, \quad \delta_n^{C_i|S_i} = \frac{f_{C_i|S_i}}{\mu_n^{C_i|S_i}}, \quad i = 2, \ldots, k. \tag{11}$$

6. Simulation in the graphical setting

In this section, we present a simulation study aimed at showing the performances of our corrected LRTs versus ordinary LRTs when working with Gaussian graphical models. For a real data application, see the Supplementary Material. We consider a p-variate Gaussian graphical model Markov with respect to a graph with $p = 14$ nodes and $k = 4$ cliques (see Supplementary Material for a representation of the graph). We consider a RIP-respecting sequence C_1, C_2, C_3, C_4 of cliques, with cardinalities $|C_1| = 8$, $|C_2| = 5$, $|C_3| = 3$, $|C_4| = 2$, giving rise to the following cardinalities for the corresponding sequence of separators: $|S_2| = 2$, $|S_3| = 1$, $|S_4| = 1$. We generate data assuming that differences between the two conditions are attributable to nodes 1 and 2, located in C_1. In particular, in one condition the means of the two elected nodes is set to be 1.5 times greater than the means of the same nodes in the other condition, while the variances are decreased by 50%. It follows that the null hypothesis of equality of distribution for X_{C_1} is false, since C_1 includes the two altered nodes. All remaining null hypotheses of equality of distribution for $X_{R_i}|X_{S_i}$, $i = 2, 3, 4$, are true, thanks to the Markov properties of the graph. We run 10,000 simulations assuming $n_1 = n_2 \in \{10, 50, 100, 250\}$. For each sample, we compute the following statistics: $W_n^{C_1}$, $W_n^{C_i|S_i}$, $T_n^{C_1}$, $T_n^{C_i|S_i}$, $i = 2, 3, 4$. The nominal Type I error rate is set to be $\alpha = 0.05$.

Results are reported in Table 1 (see also the Supplementary Material for a simulation under the global null). Row 1 of Table 1 shows the empirical power of the test, while rows 2–4 show the empirical Type I error rates. For what concerns W_n, note that for small sample sizes, the empirical Type I error rate is significantly higher than the nominal one, due to a large number of false rejections. This happens for all the local problems, but, for a fixed sample size, the number of false rejections largely depends on the dimension of the problem. As expected, this behavior decreases as the sample size increases, and asymptotically, the distribution of W_n can be approximated with a chi-square. On the other hand, the adjusted statistic T_n reaches the nominal size of the test for each considered sample size, regardless of the dimension of the local problems. The power of the test based on the adjusted statistic T_n on the clique C_1 increases with the sample

Table 1
Power and Type I error computed for each term of the decomposition. Proportion of rejected tests out of 10 thousand simulations, for different sample sizes, with significance level $\alpha = 0.05$.

n_j	W_n				T_n			
	10	50	100	250	10	50	100	250
C_1	0.985	0.730	0.970	1.000	0.066	0.535	0.946	1.000
$C_2 \vert S_2$	0.445	0.082	0.065	0.056	0.048	0.051	0.050	0.049
$C_3 \vert S_3$	0.167	0.061	0.056	0.051	0.049	0.044	0.048	0.049
$C_4 \vert S_4$	0.109	0.060	0.051	0.057	0.047	0.052	0.048	0.055

size. The high power observed for W_n should not be misleading, as it highly depends on the false rejections due to the approximation issues already highlighted in Section 4. The adjusted statistic meets the expectations, being able to identify the altered clique, while controlling the Type I error of the remaining local tests.

7. Conclusions

In this paper, we proposed an adjusted LRT, which leads to valid inference at different dimensionality regimes. Our proposal overcomes some weaknesses of alternative corrections reported in the literature, that occur at small sample sizes and, in particular, when the dimension p is close to n. We showed that the phase transition boundary of the LRT statistic corrected following our proposal is $d = 1$, indicating that the only condition needed to work is $p/n \to 0$. Simulations confirmed that the adjusted test statistic is well approximated by a chi-square distribution both for small and large values of p.

In the context of decomposable Gaussian graphical models, where the problem of testing equality of two networks breaks down into a sequence of problems defined on smaller sets of variables, our correction can help tackling the possibly high heterogeneity resulting from the decomposition in terms of dimensionality regimes. Our simulation study showed that the size of the test was reached for different configurations of p and n and, in the presence of a difference in two conditions, the adjusted statistic is able to detect it, still controlling the Type I error in the other cliques.

Data availability

The data are publicly available in the R package ALL.

Acknowledgments

DR was supported by the National Cancer Institute of the National Institutes of Health (U24CA180996).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.spl.2022.109732.

References

Bartlett, M.S., 1937. Properties of sufficiency and statistical tests. Proc. R. Soc. Lond. A Math. Phys. Sci. 160 (901), 268–282.
Djordjilović, V., Chiogna, M., 2022. Searching for a source of difference in graphical models. J. Multivariate Anal. 190, 1–13.
Edwards, D.I., 2000. Introduction to Graphical Modelling. Springer.
He, Y., Meng, B., Zeng, Z., Xu, G., 2021. On the phase transition of Wilks' phenomenon. Biometrika 108 (3), 741–748.
Jiang, T., Qi, Y., 2015. Likelihood ratio tests for high-dimensional normal distributions. Scand. Stat. Theory Appl. 42 (4), 988–1009.
Jiang, T., Yang, F., 2013. Central limit theorems for classical likelihood ratio tests for high-dimensional normal distributions. Ann. Statist. 41 (4), 2029–2074.
Lauritzen, S.L., 1996. Graphical Models. Clarendon Press.
Muirhead, R.J., 1982. Aspects of Multivariate Statistical Analysis. Wiley & Sons.
Van der Vaart, A.W., 1998. Asymptotic Statistics. Cambridge University Press.
Whittaker, J., 1990. Graphical Models in Applied Multivariate Statistics. Wiley.
Wilks, S.S., 1938. The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann. Math. Stat. 9 (1), 60–62.