Statistical Papers
https://doi.org/10.1007/s00362-023-01473-6

REGULAR ARTICLE

Improving the power of hypothesis tests in sparse contingency tables

Federica Nicolussi[1] · Manuela Cazzaro[2] · Tamás Rudas[3]

Received: 8 April 2022 / Revised: 15 May 2023

Abstract
When analyzing data in contingency tables it is frequent to deal with sparse data, particularly when the sample size is small relative to the number of cells. Most analyses of this kind are interpreted in an exploratory manner and even if tests are performed, little attention is paid to statistical power. This paper proposes a method we call redundant procedure, which is based on the union–intersection principle and increases test power by focusing on specific components of the hypothesis. This method is particularly helpful when the hypothesis to be tested can be expressed as the intersections of simpler models, such that at least some of them pertain to smaller table marginals. This situation leads to working on tables that are naturally denser. One advantage of this method is its direct application to (chain) graphical models. We illustrate the proposal through simulations and suggest strategies to increase the power of tests in sparse tables. Finally, we demonstrate an application to the EU-SILC dataset.

Keywords Categorical variables · MC simulation · Union intersection principle · Redundant test · Graphical model

1 Introduction

The standard approaches to testing the relationships among a set of categorical variables collected in a contingency table are to use the likelihood ratio test G^2 or the Pearson's X^2 test. However, sparse tables face the problem that conventional statis-

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 730998, InGRID-2 Integrating Research Infrastructure for European expertise on Inclusive Growth from data to policy.

✉ Federica Nicolussi
federica.nicolussi@polimi.it

[1] MOX, Department of Mathematics, Politecnico di Milano, Milan, Italy

[2] Department of Statistics and Quantitative Methods, University of Milano - Bicocca, Milan, Italy

[3] Department of Statistics, Eötvös Loránd University, Budapest, Hungary

Published online: 03 August 2023

tical tests have low power and are inaccurate for type I errors. Many works in the literature have dealt with this topic. For instance, Mehta and Patel (1983), in a study on sparsity, showed that Fisher's exact test and the asymptotic X^2 Pearson's test give contradictory results for high levels of sparseness. Rudas (1986) studied the goodness of fit of the X^2 test, likelihood ratio test G^2, and Cressie-Read statistics. Koehler (1986) showed that the Gaussian approximation of the likelihood ratio statistic G^2 is more accurate than the χ^2 approximation in sparse contingency tables. Cressie and Read (1989) report that G^2 is generally not well approximated by the chi-squared distribution in case of sparseness. In particular, in the presence of high sparsity the chi-squared approximation produces conservative significance levels. That is the likelihood ratio test G^2 in very sparse table tends to reject the null hypothesis less frequently.

The sparseness problem is becoming increasingly important. Contemporary studies, such as studies of gene expression, require a suitable methodology that takes account of sparseness. Procedures such as multiple testing, machine learning, and probabilistic graphical models can address the sparseness problem in different fields, and some existing algorithms take account of tests' loss of power for sparse tables; see, for instance, Mieth et al. (2016). However, the literature on this topic is limited.

On the one hand, some authors, such as Fienberg and Rinaldo (2012) and Dale (1986), have focused on the maximum likelihood estimation in log-linear models and the conditions under which the X^2 statistic and the likelihood ratio test G^2 can be normal, or chi-squared approximate when a table is sparse. On the other hand, other authors have proposed alternative statistics that give more accurate results than X^2 and G^2. For example, Kim et al. (2009) compared the statistic D^2 proposed by Zelterman (1987) and the statistic L_r introduced by Maydeu-Olivares and Joe (2005, 2006) with the classical X^2 Pearson's test.

All authors have concluded that the asymptotic efficiency of the statistical tests depends on the sparseness of the contingency table, and the statistics used in the exact methods, such as the Fisher score, are computationally expensive; see among others Kim et al. (2009).

The present study examines the power of different tests for data collected in contingency tables, highlighting the inverse relationships between sparseness and tests' power; that is, the power of tests evaluated on dense tables is higher than the power on sparse tables. Consequently, we consider including denser marginal contingency tables in the analysis. For sparse tables, it is appropriate to say that, by marginalizing over one or more variables, the resulting marginal contingency table reduces the sparsity because the number of observations n does not change, but the total number of cells d decreases.

Thus, if we have a marginal non-sparse contingency table on which we carry out an independence test, the likelihood ratio test's asymptotic properties hold and, as shown by Kim et al. (2009), for instance, the independence test does not lose power.

The idea of this paper is to use the advantage gained by the dense marginal contingency tables in the multiple testing procedure according to the logic of the union–intersection (UI) test; see Roy (1953), Gabriel (1969). In simple terms, the UI test states that if we express the null hypothesis as the intersection of several component hypotheses, we reject the global null hypothesis if we reject at least one single hypothesis. In this test, the global test's power is greater than or equal to that of the individual test with the highest power. Among others, Agresti and Gottard (2007) connected the UI logic to log-linear models. The UI approach also overcomes the limitations of the likelihood ratio test when the null or alternative hypotheses of a statistical testing problem are composed of finitely many regions of varying dimensionality, as illustrated by Perlman and Wu (2003).

In particular, where the (global) null hypothesis defined on a (global) sparse contingency table is the intersection of a small number of component null hypotheses defined on a lower-dimensional denser marginal table, the test of a component hypothesis may be more powerful than the global test. Accordingly, we propose a strategy called redundant procedure which consists of testing one hypothesis together with another explicitly or implicitly implied hypothesis.

These procedures become particularly relevant when relationships among a set of categorical variables are studied through graphical models.

We use graphical models that take advantage of chain graphs (CGs), also known as chain graph models (CGMs); see Drton (2009) for an overview. CGs are a natural generalization of directed acyclic graphs and undirected graphs. Directed acyclic graphs have proved useful in constructing expert systems, developing efficient updating algorithms, and in reasoning about causal relations. Further, they represent conditional independencies based on subsets of variables. Graphical models based on undirected graphs, in contrast, have been used in spatial statistics to analyze data from field trials, image processing, and many other applications. See Lauritzen and Richardson (2002) and references therein. CGs admit both directed and undirected arcs and they have been applied in many fields, Cox and Wermuth (1996).

We consider here CGs graphical models for the numerous marginal distributions they naturally involve. Indeed, the list of independencies read off these kinds of graphs typically involves subsets of variables. Among all the possible parameterizations of these models we choose the one based on marginal models, an extension of log-linear models proposed by Bergsma and Rudas (2002), that allows us to test hypotheses by constraining specific parameters to zero in certain marginal distributions. Indeed, these models and their generalizations are increasingly used to deal with problems associated with sparse data; see for instance Belilovsky et al. (2017), Sedgewick et al. (2016), Yoshida and West (2010), Henao and Winther (2009).

This paper is structured as follows. Section 2 explains procedures to avoid the sparsity problem. In particular, Sect. 2.1 presents the union–intersection logic with a focus on tests' power. In Sect. 2.2, as an original contribution, we suggest a redundant testing procedure, a strategy to deal with sparse contingency tables to increase tests' power. In Sect. 2.3 we present CGMs as a useful application for the redundant test. Section 2.4 develops our proposal further and suggests possible strategies to adopt. Section 3.1 reports a simulation study where we consider tests' power for different levels of sparsity and different sizes of tests. We present the results in Sect. 3.2 and discuss the occurrence of the different distributions under different situations in Sect. 3.3. An application to the EU-SILC dataset for 2016 is demonstrated in Sect. 4. Finally, Sect. 5 concludes.

2 Methodology

2.1 The Union–intersection principle

Given a fixed sample space, the family of joint probability distributions $\boldsymbol{p}_\theta$ of a set of categorical variables $\boldsymbol{X}$, depending on parameters θ and taking values in parameter space Θ, has the form

$$\mathcal{P}_\Theta = \{\boldsymbol{p}_\theta : \theta \in \Theta\}.$$

Let us consider $\theta = (\theta_A, \theta_B)$ as a vector valued so that (θ_A, θ_B) is a partition of θ and let $\Theta = \Theta_A \times \Theta_B$ be a partition of the parameter space. Model $\mathcal{M}_0$ describes a restriction on the parameter set $\theta_A \in \Theta_A^*$, with $\Theta_A^* \subset \Theta_A$, written in the form

$$\mathcal{M}_0 = \{\boldsymbol{p}_\theta : \theta_A \in \Theta_A^*\}.$$

Although some results of this paper apply more generally, it is assumed that the model of interest $\mathcal{M}_0$ is an open set in $\mathcal{P}_\Theta$.
Let a partition of vector θ_A be

$$\theta_A = (\theta_{A_1}, \theta_{A_2}, \ldots, \theta_{A_k}),$$

and for $i = 1, \ldots, k$, define

$$\mathcal{M}_{0_i} = \{\boldsymbol{p}_\theta : \theta_{A_i} \in \Theta_{A_i}^*\},$$

with an appropriate partition of Θ_A^*.
Then,

$$\mathcal{M}_0 = \cap_{i=1}^k \mathcal{M}_{0_i}. \tag{1}$$

In such a case, the union–intersection principle may offer advantages in testing over a direct test of the global hypothesis

$$H_0 : \boldsymbol{p}_\theta \in \mathcal{M}_0.$$

As will be seen, these advantages are particularly large when tests of some of the component hypotheses

$$H_{0_i} : \boldsymbol{p}_\theta \in \mathcal{M}_{0_i}$$

are more powerful than the test for H_0. For multivariate categorical data, this is the case when the component hypotheses are defined on lower dimensional marginal tables less affected by sparsity than the whole table on which $\mathcal{M}_0$ is defined.
The standard procedure to test H_0 is to perform a test T_0 and to reject H_0 if T_0 leads to rejection. In general, the union–intersection principle (Roy 1953; Roy and Mitra

1956) is to apply a testing procedure T_{0_i} to test H_{0_i} for all $i = 1, \ldots, k$ and to reject H_0 if any of T_{0_i} lead to rejection, or retain H_0 if none of the component tests leads to rejection. In this procedure, the global null hypothesis is the intersection of the component null hypotheses and the global alternative is the union of the component alternatives.

For any test T_{0_i}, it is possible to choose a different level α_i. We denote this test $T_{0_i}(\alpha_i)$ to highlight the dependence with the α level. The vector $\boldsymbol{\alpha}^T = (\alpha_1, \ldots, \alpha_k)$ collects all the α_i levels that defines the UI test, $T_{UI}(\boldsymbol{\alpha})$. Note that, here $\boldsymbol{\alpha}$ is not the implied level of the UI test. Without any assumption of the real distribution, the rejection probability of the UI test satisfies the following inequalities:

$$\max_i P(T_{0_i}(\alpha_i) = rej) \leq P(T_{UI}(\boldsymbol{\alpha}) = rej) \leq \sum_i P(T_{0_i}(\alpha_i) = rej), \qquad (2)$$

where rej is the contraction of *reject*. These results are a generalization of the ones provided in Roy (1953) and Gabriel (1969) under the truthfulness of H_0.

Equality on the left hand side occurs if there is one rejection region of a component hypothesis containing all the other regions. Equality on the right hand side occurs if the rejection regions of the component tests are disjoint.

When H_0 holds, it is desirable that the rejection probability of $T_{UI}(\boldsymbol{\alpha})$ is not greater than a fixed level α^*. This is ensured if the levels α_i of the component tests T_{0_i} are such that $\sum_{i=1}^k \alpha_i = \alpha^*$; see the right hand side of (2). Bonferroni's correction is a popular choice to achieve this: $\alpha_i = \frac{\alpha^*}{k}$.

When H_0 does not hold, the rejection probabilities provide the power of the test. As implied by the left-hand side of (2), the power of the UI procedure is not less than the power of any component test.

A final aspect is that the component hypotheses may be defined not on the sample space of the global hypothesis, but instead on its lower-dimensional marginals. To apply these tests to the entire contingency table, a test of H_{0_i} is identified with a test of a hypothesis for distribution on the entire table, on the basis that its relevant marginal distribution possesses the characteristic formulated in $\mathcal{M}_{0_i}$.

The following theorem summarizes and generalizes results presented in Roy (1953) and Gabriel (1969) under the truthfulness of H_0. As shown in the proof in the Appendix, it comes easily from formula (2). Let $\mathcal{M}_1$ be a model such that $\mathcal{M}_1 = \left\{\boldsymbol{p}_\theta : \theta_A \notin \Theta_A^*\right\}$, or simply $\boldsymbol{p}_\theta \notin \mathcal{M}_0$, then the following theorem proposes a variant of (2).

Theorem 1 *Let*

$$\sum_{i=1}^{k} \alpha_i = \alpha^*. \qquad (3)$$

Then,

$$P(T_{UI}(\boldsymbol{\alpha}) = rej|\mathcal{M}_0) \leq \alpha^*. \qquad (4)$$

Further,

$$\max_i P(T_{0_i}(\alpha_i) = rej|\mathcal{M}_1) \leq P(T_{UI}(\boldsymbol{\alpha}) = rej|\mathcal{M}_1). \tag{5}$$

Thus, the UI test is more powerful than the component tests, and its probability of error type I is lower than α^*.
As (5) shows, if some of the component tests refer to hypotheses that apply to lower dimensional marginal tables, as in the case of the marginal definition and parametrization of CGMs considered in Sect. 2.3, the component tests are carried out in less (occasionally much less) sparse tables and, as illustrated in Sect. 3.2, have more power than testing H_0 directly. Thus, in this case, one can expect the *UI* procedure to have more power than a direct test of H_0.

2.2 A redundant testing procedure

The advantage of the UI test in the situations described in the previous subsection is clear. Theorem 1 gives comforting results on the power of the UI test, but not much can be concluded compared to the global test executed with a critical level of α^*. However, the method of gaining power by using less sparse marginal tables seems beneficial, as illustrated by the results presented in Sect. 3.2.
A combination of an overall test of H_0 with a test of one (or a few) of its component hypotheses, say H_{0_i}, offers advantages in such situations. This is the idea underlying the redundant procedure (R) that we propose in this section.

In order to test an overall hypothesis H_0, we test the hypothesis itself, which we call "global":

$$H_g : \boldsymbol{p}_\theta \in \mathcal{M}_0,$$

together with one or more component hypothesis(es) included in the global hypothesis:

$$H_i : \boldsymbol{p}_\theta \in \mathcal{M}_{0_i}.$$

For simplicity, we consider here only one hypothesis of the type H_i, but the results remain valid even if we add other composite hypotheses. We retain H_0 if neither a test of H_g nor a test of H_i suggest rejection. We reject H_0 if either suggest rejection.
Such a procedure is called 'redundant' because if testing procedures led to determining whether or not the relevant hypothesis was true, then testing both H_i and H_g could be redundant, as when H_g is true, then H_i is also true and when H_i is not true, then H_g is also not true. The procedure proposed also tests H_i, even if H_g was retained.
Then, for any true distribution and any level α_g and α_i at which T_g and T_i are carried out, for the test based on the redundant procedure, T_R, the following holds:

$$\max(P(T_g(\alpha_g) = rej),\ P(T_i(\alpha_i) = rej)) \leq P(T_R(\boldsymbol{\alpha}) = rej) \tag{6}$$

and

$$P(T_R(\boldsymbol{\alpha}) = rej) \leq P(T_g(\alpha_g) = rej) + P(T_i(\alpha_i) = rej), \tag{7}$$

where $\boldsymbol{\alpha}$ is a vector containing (α_g, α_i).

Generally speaking, the critical values of T_g and of T_i must be selected in such a way that under the null hypothesis H_0 the upper bound is small, and under the alternative that H_0 is not true the lower bound is large. Traditionally, in multiple testing procedures the critical values are selected in such a way that under H_0 the following inequality holds:

$$P(T_g(\alpha_g) = rej) + P(T_i(\alpha_i) = rej) \leq \alpha^*, \tag{8}$$

for some pre-selected value α^*, often equal to 0.05. The following results suggest a general strategy of choosing the levels so that the size is retained, but the redundant procedure's application leads to gain in power.

Hereafter, we use equally the critical value c and the error type I probability α to characterize test T. To provide all the information in a unique symbol, we define $c_\alpha^{\mathcal{M}}(T)$ as the critical value of test T such that

$$P(T(c_\alpha^{\mathcal{M}}(T)) = rej|\mathcal{M}) = \alpha \tag{9}$$

where $\mathcal{M}$ is the real model to which the real distribution belongs. Note that the value α in (9) is the probability of the error type I of test T when model $\mathcal{M}$ satisfies H_0. Otherwise, when model $\mathcal{M}$ does not satisfy H_0, α represents the test's power. When the critical value is inserted in a test, the part in brackets becomes superfluous.

Theorem 2 provides important information about the power of the redundant test procedure when particular conditions are satisfied. For clarity, Fig. 1 shows the conditions required in formulas (10) and (11). Note that coherent with $\mathcal{M}_1$, the model $\mathcal{M}_{1_i}$ is such that $\boldsymbol{p}_\theta \notin \mathcal{M}_{0_i}$.

Theorem 2 *Let α^* be the given maximal probability of error type I.*
Let β^* be such that the critical values for the test T_g are

$$c_{\alpha^*}^{\mathcal{M}_0}(T_g) \geq c_{\beta^*}^{\mathcal{M}_1}(T_g). \tag{10}$$

If the critical values for the test T_i also satisfy

$$c_{\alpha^*}^{\mathcal{M}_{0_i}}(T_i) < c_{\beta^*}^{\mathcal{M}_{1_i}}(T_i), \tag{11}$$

then there exists a range of α_g values for testing the global hypothesis H_g, and α_i values for testing the redundant hypothesis H_i, such that

$$P\Big\{T_R(\alpha_g, \alpha_i) = rej|\mathcal{M}_1\Big\} > P\Big\{T_0(\alpha^*) = rej|\mathcal{M}_1\Big\}, \tag{12}$$

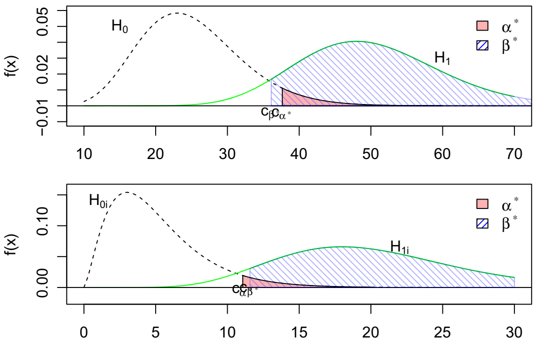

Fig. 1 *Top* Hypothesis testing where H_0 postulates that the χ^2 distribution of 25 degrees of freedom (df) (dashed line) is the true distribution and H_1 postulates that the χ^2 distribution has 50 df. The red area is α^* and the dashed blue area is β^* *Bottom* Hypothesis testing where H_{0_i} postulates that the χ^2 distribution has 5 df (dashed line) and H_{1_i} postulates that the χ^2 distribution has 20 df. The red area is α^* and the dashed blue area is β^*

and

$$P\Big\{T_R(\alpha_g, \alpha_i) = rej|\mathcal{M}_0\Big\} \leq P\Big\{T_0(\alpha^*) = rej|\mathcal{M}_0\Big\}. \tag{13}$$

Theorem 2 asserts that the redundant procedure is more powerful against $\mathcal{M}_1$ than the test for H_0 carried out with level α^*, and the size is not more than α^*.

2.3 The redundant testing procedure in chain graph models

To provide context and to introduce an important area where the procedures proposed in the paper may be useful, this subsection reviews graphical models associated with CGs. This model class includes, as special cases, graphical models associated with directed acyclic graphs and with undirected graphs. Graphical models are models where it is possible to read from a graph the independence structure between the variables involved. The use of graphical models has increased in recent decades because of their ability to display complex dependence structures intuitively. See, for example, Drton (2009). These models generally use the parameters proposed by Bergsma and Rudas (2002) and Bartolucci et al. (2007), who take advantage of a list of marginal distributions on which to define the parameters, according to certain properties listed by Bergsma and Rudas (2002). For different applications, see Nicolussi and Colombi (2017), Rudas et al. (2010), Marchetti and Lupparelli (2011), Nicolussi and Cazzaro (2021).

A brief description of these models is reported in this subsection. A CG is defined by a set of nodes and a set of arcs, which can either be directed or undirected. Two nodes are adjacent if an undirected arc links them. Two nodes are parent and child, respectively, where a directed arc starts from the parent and points to the child. Given a subset of nodes A, the parent set of A, $pa(A)$, is composed of the nodes that are parents of at least one node in A. A set of partially ordered chain components $\{C\}$

defines a partition of the nodes in the graph. There are only undirected arcs within each component, and between two nodes that belong to different components there are only directed arcs, all pointing in the same direction, in compliance with the partial order.[1] Given a component C_i, the set of components that have at least one node pointing to one node in C_i is described as set of parent components $pa(C_i)$. Intuitively, in a CGM, each node represents a variable, and the undirected (directed) arcs denote possible symmetric (asymmetric) dependence relationships between the variables. In contrast, the lack of an arc means no direct effect or association, that is, conditional independence. There are various precise interpretations associated with the same CG, see Drton (2009). This paper uses type IV CGMs, also known as multivariate regression graphical models. In this interpretation, a missing undirected arc between two nodes in the same component means that these two variables are conditionally independent given all variables in the parent components. If there is a missing arc between two variables in different components, then these are conditionally independent, given the parents of the variables in the child component. Further, variables in a component are conditionally independent of variables in a second component that contains no nodes into which a directed path leads from any variable in the first component, given the variables in the parents of the first component. The following definition presents these concepts formally.

Definition 1 Let us consider a set of categorical variables $\boldsymbol{X}$ with joint probability distribution $\mathcal{P}_\Theta = \left\{\boldsymbol{p}_\theta : \theta \in \Theta\right\}$. A probability distribution is said to be faithful to the graph if the following Markov properties hold.

(a) If in the CG, two subsets of variables, say $\boldsymbol{X}_A$ and $\boldsymbol{X}_B$ belonging to the same component C_i, are not connected, then

$$\boldsymbol{X}_A \perp\!\!\!\perp \boldsymbol{X}_B | pa(C_i).$$

(b) If in the CG, two subsets of variables, say $\boldsymbol{X}_A \subseteq C_i$ and $\boldsymbol{X}_B \subseteq C_j$, belong to different components such that arcs go from C_i to C_j, i.e., $C_i \in pa(C_j)$, but no arcs go from $\boldsymbol{X}_A$ to $\boldsymbol{X}_B$, i.e., $\boldsymbol{X}_A \notin pa(\boldsymbol{X}_B)$, then

$$\boldsymbol{X}_A \perp\!\!\!\perp \boldsymbol{X}_B | pa(\boldsymbol{X}_B).$$

(c) If in the CG, C_i and C_j are such components that no directed path leads from any variable in C_i to any variable in C_j, but C_j is not a parent of C_i, then

$$C_i \perp\!\!\!\perp C_j | pa(C_i).$$

Consequently, a CGM is defined by conditional independencies about various marginal tables and may be seen as the intersection of conditional independence models on these marginals; see Rudas et al. (2010). Some of these marginals may be much smaller than the table of the whole set of variables, because fewer variables are involved.

[1] In general, the components of a CG and the component hypotheses discussed in Sect. 2.1 are unrelated, but in the application discussed in this paper, the component hypotheses are the CGM assumptions for the components.

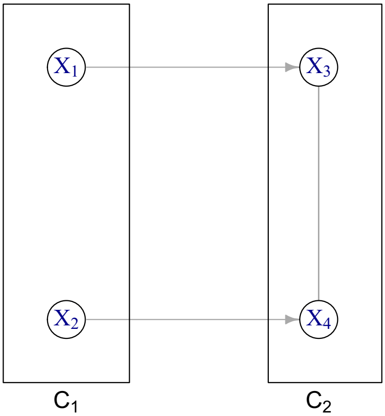

Fig. 2 A CGM with four variables X_1, X_2, X_3, and X_4 and two components $C_1 = (X_1, X_2)$ and $C_2 = (X_3, X_4)$

Example 1 Let X_1, X_2, X_3, and X_4 be four categorical variables, where X_1 and X_2 are explanatory for X_3 and X_4. The CGM in Fig. 2 describes one possible structure of relationships among these variables. This model posits three (conditional) independencies: $X_1 \perp\!\!\!\perp X_2$, that is, the explanatory variables are independent; $X_3 \perp\!\!\!\perp X_2|X_1$; and $X_4 \perp\!\!\!\perp X_1|X_2$, that is, only one explanatory variable has a direct effect on each response. All these independencies apply to lower-dimensional marginal tables of the 4−dimensional table. The CGM is the intersection of three models, each assuming one of the (conditional) independencies on the relevant marginal tables.

The redundant procedure finds an easy application in the CGM. In this case, the global hypothesis can consider the family of distributions that satisfies the whole set of conditional independencies represented by the graph. Thus, it involves the joint distribution defined on a table with d_0 cells. In order to define H_{0_i} we can select, if any, one conditional independence defined on a smaller table with d_{0_i} cells, where $d_0 > d_{0_i}$. By considering Theorem 2, to what extent (11), depicted in the bottom of Fig. 1, may be achieved depends on two contradicting factors. One is that the family of probabilities satisfying the two hypotheses belongs to $\mathcal{M}_0 \subseteq \mathcal{M}_{0_i}$, and thus distributions in the alternative of H_0 may not be in the alternative of H_{0_i}. Even those that are in the alternatives of both hypotheses may be closer to H_{0_i} than to H_0. In general, this suggests that in the power of T_{0_i} is less than the power of T_0 against distributions in the alternative of H_0 for tests conducted at the same level of significance. This makes the redundant procedure, which makes use of the H_0 test, have more power than the *UI* procedure, which makes use of a H_{0_i} hypothesis test instead. The other fact is that d_{0_i} is lower (occasionally much lower) than d_0. Therefore, the former refers to a marginal table that has fewer (occasionally much fewer) cells, and thus is less sparse. This implies that the test of H_{0_i} is more powerful than the test of H_0, against alternatives in their respective alternative hypotheses.
When the first effect dominates, i.e. the smooth manifold $\mathcal{M}_{0_i}$, defined in d_{0_i}−dimensional space, contains the alternative distribution of H_0, the joint testing of H_0 in the redundant procedure allows to not loose too much power. On the contrary, when the second effect dominates, we can gain a lot of power by the density of the table. Indeed, as illustrated in Sect. 3.1, it is the second of these two contradicting effects that domi-

nates. Thus in most cases, the redundant testing procedure has greater power than the simple test of H_0, even if the two procedures have the same size.

2.4 Strategies

So far, we have highlighted the advantages of the UI and R procedures when at least one component hypothesis concerns a small subset of variables. This subsection goes further and suggests separately testing not a component hypothesis, but rather an implication. However, possible strategies depend on the data, the variables, and the hypotheses to test. Consider the case of testing a global hypothesis involving several variables, defined on a huge contingency table that is likely sparse. An independence based on a smaller subset of variables could be implied from the global hypothesis by applying the *weak-union* or *decomposition* properties of the independence statements, see Maathuis et al. (2018). In this case, if there is at least one independence statement that can be defined on a less sparse marginal, the redundant procedure can consider the original global hypothesis and this implied independence.

When this scenario is not possible, we can investigate whether there is a plausible hypothesis in a smaller and denser marginal distribution.

Notice that Theorem 2 holds also in this case.

3 Simulations

3.1 A simulation study

We show two scenarios from which sparsity can originate. In the first case we have a small set of variables, one of which has a large number of categories. The second scenario considers a large number of variables. In both cases, the overall hypothesis H_0 to be tested consists of two sub-hypotheses (H_{0_1} and H_{0_2}). We set H_{0_1} as the hypothesis in the dense marginal table, while H_{0_2} as the one in the sparse contingency table. The simulation presented supports the results obtained for the UI and R procedures in terms of performance, because both have higher power, or at least similar power, to the global test.

We use marginal models (Bergsma and Rudas 2002) that allow to impose constraints on log-linear parameters defined on marginal distributions. However, alternative parameterizations of the joint and marginal probability distributions are still valid; see for instance the log-mean parameterization of Roverato (2015). The likelihood ratio test is used to infer the previous hypotheses by comparing the likelihood of the model to test with the likelihood of the unconstrained model. When the null hypothesis holds, the test statistic asymptotically follows the χ^2 distribution with degrees of freedom equal to the difference in the number of free parameters in the two models. Note that, here we are considering situations where the asymptotic theory, in general, does not hold, thus we will evaluate the power of the tests according to the Monte Carlo (MC) procedure which we will detail shortly.

We test the global hypothesis $H_0 : \boldsymbol{p}_\theta \in \mathcal{M}_0 = \cap_{i=1}^{2} \mathcal{M}_{0_i}$ by following three different approaches.

- $\boldsymbol{T_0}$: the likelihood ratio test for model $\mathcal{M}_0$. The corresponding test statistic is denoted G_0^2.
- $\boldsymbol{T_{UI}}$: the likelihood ratio test for model $\mathcal{M}_{0_1}$ and the likelihood ratio test for model $\mathcal{M}_{0_2}$ taking advantage of the union–intersection procedure. In this case, the two test statistics are $G_{0_1}^2$ for the independence H_{0_1} and $G_{0_2}^2$ for H_{0_2}.
- $\boldsymbol{T_R}$: the likelihood ratio test for model $\mathcal{M}_{0_1}$ and the likelihood ratio test for model $\mathcal{M}_0$ taking advantage of the redundant procedure. In this case, the two test statistics are $G_{0_1}^2$ for the independence H_{0_1} and G_0^2 for H_0.

The simulation aims to compare the power of the test of H_0, $\pi(\boldsymbol{T_0})$, with the test performed using the *UI* procedure, $\pi(\boldsymbol{T_{UI}})$, and with the test performed using the redundant procedure, $\pi(\boldsymbol{T_R})$.
We build four joint distributions, where the real distribution falls surely:

- pH_0: where H_0 is true;
- pH_{1_1}: where the component hypothesis H_{0_1} does not hold;
- pH_{1_2}: where the component hypothesis H_{0_2} does not hold;
- $pH_{1_{12}}$: where neither hypothesis holds.

Two scenarios are presented.
Scenario 1 Let X_1, X_2, and X_3 be three random variables, taking values in a table $\mathcal{I}_{123}$ of dimension $(2 \times 2 \times 50)$ and let the joint probability distribution defined as $p_{123}(\boldsymbol{i}_{123})$ with $\boldsymbol{i}_{123} \in \mathcal{I}_{123}$. Further, let the overall hypothesis be H_0, expressed as the intersection of

$$\begin{array}{l} H_{0_1} : X_1 \perp\!\!\!\perp X_2 \\ H_{0_2} : X_2 \perp\!\!\!\perp X_3 | X_1, \end{array} \tag{14}$$

against the alternative H_1 where at least one independence does not hold.

Thus, the two models corresponding to the two sub-hypotheses impose constraints on $\mathcal{I}_{12}$ and $\mathcal{I}_{123}$ marginal tables of dimension 2×2 and $2 \times 2 \times 50$, respectively.
To construct the theoretical distributions where H_{0_1}, H_{0_2}, both or either holds we use the following schemes:

$$p_{123}(\boldsymbol{i}_{123}) = p_1(i_1) p_{2|1}(i_2|i_1) p_{3|12}(i_3|\boldsymbol{i}_{12}) \tag{15}$$

for all $i_1 \in \mathcal{I}_1$, $i_2 \in \mathcal{I}_2$ and $i_3 \in \mathcal{I}_3$, where the notation $p_{\circ|\star}$ stays for the conditioning distribution of $X_\circ$ given $X_\star$. Further, under the different hypotheses, we apply the following simplification:

$$\begin{cases} p_{2|1}(i_2|i_1) = p_2(i_2) & \text{under } H_{0_1} \\ p_{3|12}(i_3|\boldsymbol{i}_{12}) = p_{3|1}(i_3|i_1) & \text{under } H_{0_2}. \end{cases} \tag{16}$$

Let $\boldsymbol{p}_1$ be the vector containing the marginal distribution of X_1 and $\boldsymbol{p}_2$ accordingly. We set $\boldsymbol{p}_1 = [0.3, 0.7]$ and $\boldsymbol{p}_2 = [0.75, 0.25]$. To cover several situations we

Table 1 Odds ratios evaluated in the marginal table X_1, X_2 for the three cases considered in pH_{1_1}. Here, *la* is "*low-association*", *ma* is "*moderate-association*", and *sa* is "*strong-association*"

Cases	*la*	*ma*	*sa*
Odds-ratio	0.545	0.304	6.851

Table 2 Mean and standard deviation of the 98 odds ratios evaluated in the conditional tables $X_2, X_3|X_1 = 0$ and $X_2, X_3|X_1 = 1$ for the two cases considered in pH_{1_2}. Here, *mv* is "*moderate-variability*" and *pv* is "*pronounced-variability*"

Cases	*mv*	*pv*
Mean	1.04	1.10
Standard deviation	0.379	0.595

contemplated three possible marginal distributions of (X_1, X_2), with increasing levels of association. The distributions are obtained by fixing the value of the odds-ratio as described in Table 1.

From $p_{12}(\boldsymbol{i}_{12})$ we obtain $p_{2|1}(i_2|i_1)$ to use in formula (15).

The conditional distribution $p_{3|1}(i_3|i_1)$ for all $i_3 \in \mathcal{I}_3$ is obtained by sampling from a Uniform(0,1) and then normalizing to ensure they sum to 1 for all $i_1 \in \mathcal{I}_1$. Finally, we build the conditional distribution $p_{3|12}(i_3|\boldsymbol{i}_{12})$ for all $i_1 \in \mathcal{I}_1$, $i_2 \in \mathcal{I}_2$ and $i_3 \in \mathcal{I}_3$, by considering different association cases in $p_{23|1}(\boldsymbol{i}_{23}|i_1) = p_{3|12}(i_3|\boldsymbol{i}_{12})p_{2|1}(i_2|i_1)$. In particular, for all $i_1 \in \mathcal{I}_1$, we evaluated $(2-1) \times (50-1)$ odds-ratios in $p_{23|1}(\boldsymbol{i}_{23}|i_1)$, with $\boldsymbol{i}_{23} \in \mathcal{I}_{23}$. We considered two cases of different associations as described in Table 2.

Scenario 2 Let $\boldsymbol{X}$ be a random vector composed of 8 dichotomous variables where X_1 and X_2 are two dependent variables, X_3 and X_4 are two explicative variables for X_1, X_2. Finally, X_5, X_6, X_7 and X_8 are explicative for all the variables. We wish to test the overall hypothesis H_0, composed of

$$\begin{aligned} H_{0_1} &: X_5 \perp\!\!\!\perp X_6|\boldsymbol{X}_{78} \\ H_{0_2} &: X_1 \perp\!\!\!\perp X_2|\boldsymbol{X}_{345678}. \end{aligned} \tag{17}$$

The hypothesis H_{0_1} implies one constraint on a 2^4 marginal table, and H_{0_2} implies constraints on a 2^8 table. The previous interpretation in terms of a regression model is a possible situation in which we can meet hypothesis H_0, but not the only one. We use it because it is well represented by a CG composed of three components: $\boldsymbol{X}_{5678}$, $\boldsymbol{X}_{34}$, and $\boldsymbol{X}_{12}$, and where the only missing arcs are $X_5 - X_6$ and $X_1 - X_2$.

We proceed as in scenario 1 in defining the different distributions under the possible combinations of assumptions. Let the joint probability distribution be factorized as

$$\begin{aligned} p_{\boldsymbol{X}}(\boldsymbol{i}) = p_{78}(\boldsymbol{i}_{78})p_{56|78}(\boldsymbol{i}_{56}|\boldsymbol{i}_{78})p_{34|5678}(\boldsymbol{i}_{34}|\boldsymbol{i}_{5678}) \\ p_{2|345678}(i_2|\boldsymbol{i}_{345678})p_{1|2345678}(i_1|\boldsymbol{i}_{2345678}) \end{aligned} \tag{18}$$

Further, under the different hypotheses, we apply the following simplification:

$$\begin{cases} p_{56|78}(\boldsymbol{i}_{56}|\boldsymbol{i}_{78}) = p_{5|78}(i_5|\boldsymbol{i}_{78})p_{6|78}(i_6|\boldsymbol{i}_{78}) & \text{under } H_{0_1} \\ p_{1|2345678}(i_1|\boldsymbol{i}_{2345678}) = p_{1|345678}(i_1|\boldsymbol{i}_{345678}) & \text{under } H_{0_2}. \end{cases} \tag{19}$$

Here, we fixed the marginal distribution of X_1 as $\boldsymbol{p}_1 = [0.2, 0.8]$ and we generate any conditional distribution by sampling from a Uniform(0,1) and adjusting the results ensure that the sum comes to 1.

When the probability distributions of both scenarios are ready, we estimate the power of the test. As we have already mentioned, we evaluate the power of the tests according to the MC procedure detailed below.

Step 1 We simulate a sample of size n from a multinomial distribution with the joint probability distribution pH_0.

Step 2 We calculate the test statistics G_0^2, $G_{0_1}^2$, and $G_{0_2}^2$.

Step 3 We repeat *step 1* and *step 2* m times obtaining the three MC distributions of G_0^2, $G_{0_1}^2$, and $G_{0_2}^2$.

Step 4 For fixed α_1 and α_2, and $\alpha^* = \alpha_1 + \alpha_2$, we calculate the quantile of order $1-\alpha_1$ of the distribution $G_{0_1}^2$, denoted by $X_{0_1;(1-\alpha_1)}^2$, the quantile of order $1-\alpha_2$ of the distribution $G_{0_2}^2$, denoted by $X_{0_2;(1-\alpha_2)}^2$ and the quantiles of orders $1-\alpha_2$ and $1-\alpha^*$ of the distribution G_0^2, denoted by $X_{0;(1-\alpha_2)}^2$ and $X_{0;(1-\alpha^*)}^2$.

Step 5 We repeat steps 1 to 3 replacing the distribution pH_0 with pH_{1_1}, pH_{1_2}, and $pH_{1_{12}}$ in all the scenarios considered. For each of these we evaluate the distributions of the test statistics $\bar{G}_0^2$, $\bar{G}_{0_1}^2$, and $\bar{G}_{0_2}^2$.

Step 6 We evaluate the rejection rate r as the proportion of test statistics in the alternative distributions exceeding the corresponding quantile. Below, we report the formulas used for all cases.

The estimated rejection rate of the $\boldsymbol{T_0}$ test, with level α^*, $r(H_0;\alpha^*)_{T_0}$, is

$$Mean(\mathbb{1}_{\bar{G}_0^2 \geq X_{0;(1-\alpha^*)}^2}), \tag{20}$$

where $\mathbb{1}$ is the indicator function.
The estimated rejection rate for hypothesis H_{0_1}, with level α_1, $r(H_{0_1};\alpha_1)$, is

$$Mean(\mathbb{1}_{\bar{G}_{0_1}^2 \geq X_{0_1;(1-\alpha_1)}^2}). \tag{21}$$

Analogously, we define $r(H_{0_2};\alpha_2)$.
The estimated rejection rate obtained using the union–intersection testing procedure $\boldsymbol{T_{UI}}$, at level $\boldsymbol{\alpha}^T = (\alpha_1, \alpha_2)$, $r(H_0, \boldsymbol{\alpha})_{T_{UI}}$, is

$$r(H_{0_1};\alpha_1) + r(H_{0_2};\alpha_2) - Mean(\mathbb{1}_{(\bar{G}_{0_1}^2 \geq X_{0_1;(1-\alpha_1)}^2 \cap \bar{G}_{0_2}^2 \geq X_{0_2;(1-\alpha_2)}^2)}). \tag{22}$$

Finally, the estimated rejection rate obtained using the redundant testing procedure $\boldsymbol{T_R}$, $r(H_0, \boldsymbol{\alpha})_{T_R}$, considering the two critical values fixed at level α_1 and α_2, is

$$r(H_{0_1}; \alpha_1) + r(H_0; \alpha_2) - Mean(\mathbb{1}_{(\bar{G}^2_{0_1} \geq X^2_{0_1;(1-\alpha_1)} \cap \bar{G}^2_0 \geq X^2_{0;(1-\alpha_2)})}). \tag{23}$$

All the analyses are carried out with the statistical software `R` (R Core Team 2016) with the packages `hmmm` (Colombi et al. 2014) and `MASS` (Venables and Ripley 2002) to estimate the (marginal) log-linear models and the package `igraph` (Csardi and Nepusz 2006) for the graphical representations.

3.2 Results of the simulation

Table 3 shows the detailed results for scenario 1, where native distributions pH_{1_1}, pH_{1_2}, and $pH_{1_{12}}$ are the closest to independence of all the cases considered (case *low-association* for pH_{1_1} and case *moderate-variation* for pH_{1_2}). This is the most critical case because we add the difficulty of discriminating between situations of independence and situations that are close to independence. Note that more details on critical values are reported in the Supplementary materials for both scenarios. The rejection rates in Table 3 are based on the quantiles $X^2_{0;(1-\alpha^*)}$, $X^2_{0;(1-\alpha_2)}$, $X^2_{0_1;(1-\alpha_1)}$, and $X^2_{0_2;(1-\alpha_2)}$ where the levels are $\alpha_1 = \alpha_2 = 0.025$ and $\alpha^* = 0.05$. We evaluate the rejection rates for the three alternative distributions: pH_{1_1}, pH_{1_2}, $pH_{1_{12}}$, and the null distribution pH_0. The rejection rate assumes different meanings according to whether or not the underlying distribution satisfies the null (sub)-hypothesis relative to the test. On the one hand, when the (sub)-hypothesis holds in the alternative distribution (step 5), the rejection rate is the simulated size of the test. On the other hand, when the hypothesis does not hold in the alternative distribution, the rejection rate is the simulated power (π). Thus, in Table 3, the rejection rates in shaded cells represent the simulated values of the level of the corresponding tests, whereas the rejection rates in other cells are simulated powers.

By looking at Table 3 the following considerations hold.

1. The rejection rates of $\boldsymbol{T_{UI}}$, $r(H_0, \boldsymbol{\alpha})_{T_{UI}}$, are always greater than or equal to the highest value of the rejection rates of the two component tests $r(H_{0_1}, \alpha_1)$ and $r(H_{0_2}, \alpha_2)$, according to the first inequality in formula (2).
2. The same holds for the rejection rate of $\boldsymbol{T_R}$, which is always greater than or equal to the highest value of the rejection rates of the two component tests $r(H_{0_1}, \alpha_1)$ and $r(H_0, \alpha_2)$, according to formula (6).
3. The simulated power of the test on sparse tables is always low (see the non-shaded cells of $r(H_{0_2}, \alpha_2)$, $r(H_0, \alpha_2)$, and $r(H_0, \alpha^*)_{T_0}$). This is due to the lack of information (low number of observations) needed for correct rejection of the null hypothesis.
4. The simulated power of the test for a dense table is always greater than the power of the test for a sparse table (first panel, comparison between $r(H_{0_1}, \alpha_1)$ (dense) and $r(H_0, \alpha_2)$ or $r(H_0, \alpha^*)_{T_0}$; third panel, comparison between $r(H_{0_1}, \alpha_1)$ (dense) and $r(H_{0_2}, \alpha_2)$, $r(H_0, \alpha_2)$, or $r(H_0, \alpha^*)_{T_0}$).

Table 3 Rejection rates for **scenario 1** evaluated on the distributions pH_{1_1} (*low-association*, Table 1), pH_{1_2} (*moderate variability*, Table 2), $pH_{1_{12}}$ (*low-association* and *moderate variability*) and pH_0, respectively

n	m	$r(H_{0_1}, \alpha_1)$	$r(H_{0_2}, \alpha_2)$	$r(H_0, \alpha_2)$	$r(H_0, \boldsymbol{\alpha})_{T_{UI}}$	$r(H_0, \boldsymbol{\alpha})_{T_R}$	$r(H_0, \alpha^*)_{T_0}$
$pH_{1_1} : X_1 \not\perp\!\!\!\perp X_2$							
500	1000	0.711	**0.028**	0.125	0.719	0.725	0.189
	10000	0.736	**0.036**	0.135	0.744	0.746	0.209
200	1000	0.354	**0.034**	0.074	0.377	0.380	0.123
	10000	0.377	**0.029**	0.070	0.394	0.398	0.125
$pH_{1_2} : X_2 \not\perp\!\!\!\perp X_3 \vert X_1$							
500	1000	**0.024**	0.030	0.032	0.052	0.054	0.061
	10000	**0.026**	0.036	0.035	0.060	0.059	0.068
200	1000	**0.018**	0.035	0.032	0.052	0.048	0.063
	10000	**0.027**	0.037	0.036	0.063	0.061	0.067
$pH_{1_{12}} : X_1 \not\perp\!\!\!\perp X_2$ and $X_2 \not\perp\!\!\!\perp X_3 \vert X_1$							
500	1000	0.725	0.034	0.147	0.735	0.738	0.211
	10000	0.726	0.044	0.151	0.737	0.741	0.232
200	1000	0.346	0.040	0.082	0.374	0.377	0.139
	10000	0.370	0.040	0.089	0.395	0.402	0.149
$pH_0 : X_1 \perp\!\!\!\perp X_2$ and $X_2 \perp\!\!\!\perp X_3 \vert X_1$							
500	1000	**0.021**	**0.014**	**0.016**	**0.035**	**0.036**	**0.034**
	10000	**0.024**	**0.026**	**0.027**	**0.048**	**0.049**	**0.050**
200	1000	**0.021**	**0.022**	**0.019**	**0.042**	**0.038**	**0.037**
	10000	**0.026**	**0.023**	**0.021**	**0.048**	**0.046**	**0.044**

The boldface show the simulated levels of the tests and non-shaded cells show the simulated power

n: number of observations; m: number of simulated elements in the MC distributions; $r(H, \alpha)$: rejection rate of the component tests H_{0_1}, H_{0_2}, and H_0, with test size equal to $\alpha_1 = \alpha_2 = 0.025$, $\boldsymbol{\alpha}^T = (0.025, 0.025)$, and $\alpha^* = 0.05$

5. The simulated size of $\boldsymbol{T_{UI}}$ and $\boldsymbol{T_R}$ suggests that by using $\alpha_1 = \alpha_2 = 0.025$ for the single tests, as Bonferroni's correction recommends, the real size of both procedures is closed to 0.05, which is the same as the global test $\boldsymbol{T_0}$ (fourth panel). Thus, in this table, the first three columns are comparable and the results are closed to 0.025. Analogously, the last three columns are comparable and the results are closed to 0.05. As simulated samples increase, these values are more precisely.
6. Both $\boldsymbol{T_{UI}}$ and $\boldsymbol{T_R}$ testing procedures have considerably more power against alternatives than the $\boldsymbol{T_0}$ test when H_{0_1} does not hold (first and third panels) and only a slight loss of power when H_{0_1} holds but H_{0_2} does not hold (second panel). This is because both procedures gain power from the test on the dense marginal table, and in the second panel the rejection rate for H_{0_1} represents the simulated value of α_1 (shaded cells). However, the power reduction is minimal compared to the power gain in the other situations. Further, as shown in Sect. 3.3, the situation described in pH_{1_2} is hard to occur.
7. In the first and third panels, $\boldsymbol{T_R}$ better performs than or is equivalent to $\boldsymbol{T_{UI}}$. Ultimately, the results of both tests are similar in all panels. In the first and third panels, both tests gain power from the test on the dense marginal table, which contributes almost completely to the power of the two tests. Indeed, $r(H_{0_1}, \alpha_1) >> r(H_{0_2}, \alpha_2)$ for the $\boldsymbol{T_{UI}}$ and $r(H_{0_1}, \alpha_1) >> r(H_0, \alpha_2)$ for the $\boldsymbol{T_R}$. In the second panel, the major contribution to power is made by the characterizing test for the two procedures ($r(H_{0_2}, \alpha_2)$ for $\boldsymbol{T_{UI}}$ and $r(H_0, \alpha_2)$ for $\boldsymbol{T_R}$). In all panels, the two characterizing tests have similar rejection rates. This has different justifications.

   - **Panel 1**: $r(H_{0_2}, \alpha_2)$ is about 0.025 because it represents the level of the simulated test (shaded cells). While $r(H_0, \alpha_2)$ is a simulated power that is not high because of the sparsity of the table, but still it is greater than $r(H_{0_2}, \alpha_2)$. Indeed, when we go from $n = 200$ to $n = 500$ we see that $r(H_0, \alpha_2)$ increases.
   - **Panel 2**: In both tests, the rejection rates represent the simulated power. In this case, since H_{0_1} holds, the two tests H_{0_2} and H_0 tend to reject in the same cases, that is, when there is sufficient evidence to reject H_{0_2}.
   - **Panel 3**: We are in a situation similar to panel 2, but $r(H_0, \alpha_2) > r(H_{0_2}, \alpha_2)$ because the hypothesis H_0 is more restrictive than H_{0_2} and H_{0_1} (contained in H_0) does not hold. However, the similarity between the two simulated powers can be explained by the fact that the tests for H_{0_2} and for H_0 differ in the constraint on the only parameter describing the independence between X_1 and X_2. Likely, by increasing the number of parameters to be constrained to test H_{0_1} the difference between the two tests will be more pronounced.

Table 4 compares the three procedures two by two in terms of power, for all cases considered in scenario 1 and scenario 2.

The results reported in Table 4 are in line with those in Table 3. In particular, for all cases and both scenarios, the redundant procedures perform better than, or exactly the same as, the global procedure when the alternative distribution is pH_{1_1} or $pH_{1_{12}}$. The union–intersection procedure has similar results for the first scenario, in the second looks a little worse. However, in these cases the discrepancy is minimal, and by looking at the plots in the Supplementary materials it can be seen that by changing the level of

Table 4 Paired comparison of tests for scenario 1 and scenario 2

Dist	K	T_{UI} vs T_0	T_R vs T_0	T_R vs T_{UI}
Scenario 1				
pH_{1_1}	12	12	12	12
pH_{1_2}	8	1	1	2
$pH_{1_{12}}$	24	24	24	24
Scenario 2				
pH_{1_1}	4	4	4	4
pH_{1_2}	4	1	1	0
$pH_{1_{12}}$	4	4	4	4

K: number of cases considered
Columns from 3–5: number of times the power of the first test is greater than or equal to the power of the second test

α_1 and α_2 more satisfactory results can be obtained. Indeed, in pH_{1_1} and $pH_{1_{12}}$ the two procedures gain power from hypothesis H_{0_1} (test in a dense table).
Further, in almost all cases the redundant procedure has the same power as or outperforms the union–intersection procedure. As argued in the comment of Table 3 (item 7), the difference between the simulated powers in $\boldsymbol{T_{UI}}$ and $\boldsymbol{T_R}$ in scenario 1 is small because of the little difference (only 1) in the number of constrained parameters to be tested. Weaknesses arise when pH_{1_2} holds. However, none of the three tests has high power (all under 10%). In scenario 2, the tests of H_0 and H_{0_2} differ in 4 constrained parameters and the difference between $\boldsymbol{T_R}$ and $\boldsymbol{T_{UI}}$ is more clear. For more details on scenario 2 see Fig. 6 in Supplementary materials.
A necessary consideration should be made regarding the satisfaction of conditions (10) and (11) of Theorem 2. In both scenarios, the conditions are met in all cases for pH_{1_1} and $pH_{1_{12}}$. Indeed, in these two cases we have clear evidence of the power improvement of the redundant test, whereas in pH_{1_2} we have already identified weaknesses. Detailed tables are provided in the Supplementary materials.

Finally, we want to study how the power of the test changes as the α_1 level of the H_{0_1} test changes.

Figure 3 shows the distribution of the power of the scenario 1, in the most critical situation of *low-association* and *moderate variability*. The figures concerning all the other cases in details are in the Supplementary materials together with the plots concerning scenario 2. The underline distributions are pH_{1_1}, pH_{1_2}, and $pH_{1_{12}}$ for the three rows, respectively. In the columns the n and m changes. The two dotted lines represent the UI (green) and the redundant (light blue) procedures, while the continue red line represent the constant power of the global test.
The redundant and the UI procedures show results very similar, although, in correspondence of low values of α_1, in the first and third rows, the redundant procedure seems to be slight better. The reason lies in the fact that when α_1 decreases, the contribution of the test on the dense table decreases in both procedures. However, also when the major contribution to the test is due to testing H_0 and H_{0_2} in redundant and UI procedure, respectively, the difference is small. This is due to the fact that H_0 and H_{0_2} have a similar parameter space (they differ in only one parameter constrained).

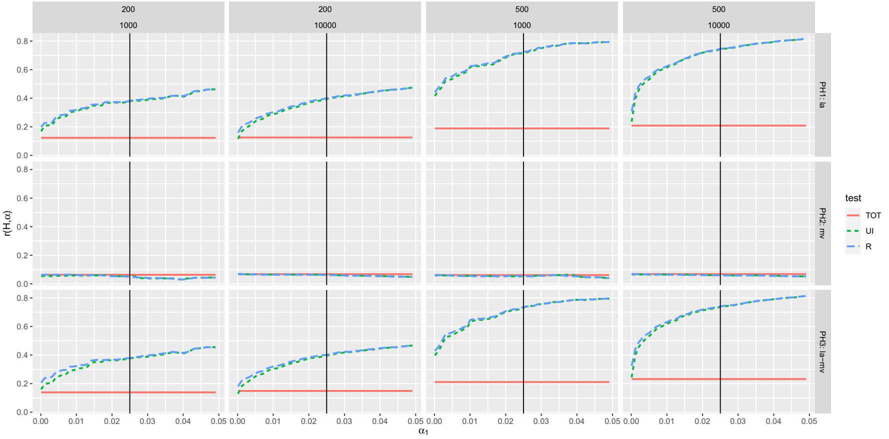

Fig. 3 *Scenario 1*. Power distribution of the test H_0 (case *low-association* Table 1 and *moderate-variability* Table 2), as size α_1 of test H_{0_1} increases. The three underlying distributions are displayed by row. In red (continuous line) is the constant power of the global test at fixed level $\alpha^* = 0.05$. The black vertical line fix the value of $\alpha_1 = 0.025$, that represents the data displayed in Table 3

This is confirmed also in scenario 2, where the parameter space of the two hypotheses differ in 4 constrained parameters and the rejection rates of the two procedure have a slightly difference, especially when the significance level of H_{0_1} test is small.

It is evident that when Theorem 2 holds, (first and third rows), the redundant procedure and the union intersection procedure have more power whatever the value of α_1 (see Fig. 6 in Supplementary materials).

Furthermore, in the first and third line, by imputing large α_1, the power increases. On the other hand, in the second line, it seems that the power of UI and redundant procedure is slightly worst by increasing α_1. For this reason, a middle ground with $\alpha_1 = 0.025$ is chosen (black vertical line).

3.3 Occurrence of the different distributions under the component hypothesis

As we have seen, in both scenarios the only situation where $\boldsymbol{T_{UI}}$ and $\boldsymbol{T_R}$ do not perform as well as the test $\boldsymbol{T_0}$ in terms of power is when the underlying distribution is pH_{1_2}. Namely, when the distribution in the alternative of H_0 satisfies H_{0_1}, although the three procedures do not perform very differently from each other. However, we next illustrate, by assuming a particular data generating mechanism of data perturbation, that this case is not so frequent.

We implement the following algorithm.

1. We generate a random contingency table with fixed size $n = 200, 500$ using the underlying distribution pH_0.
2. We increase c cells randomly selected from the simulated joint distribution under H_0 by f unit(s) and decrease c random cells (with starting frequency at least equal to f) by f unit(s); we used different values of c and f to perturb about the 5%, 10% and 15% of total frequencies, respectively.

Table 5 *Scenario 1*. Percentage in which the 10,000 perturbed distributions are closest to the column hypotheses, according to the odds-ratios

		$H_{0_1}\&H_{0_2}$	$H_{1_1}\&H_{0_2}$	$H_{0_1}\&H_{1_2}$	$H_{1_1}\&H_{1_2}$
c	f	$n = 200$			
5	1	43.94	33.57	12.86	9.61
5	2	42.35	31.54	15.13	10.98
5	3	41.92	29.27	17.61	11.20
c	f	$n = 500$			
6	3	72.78	22.98	3.07	1.17
5	5	73.03	21.68	4.25	1.04
7	5	72.96	21.57	4.20	1.27

f: amount of frequencies that is shifted from one cell to another; $2c$: number of cells randomly selected to shift the frequency. The combination of f and c selected are such to perturb about the 5%, 10%, 15% of the total frequencies

3. We evaluate the odds ratio(s) on the dense marginal (conditional) distribution(s) concerning the variables involved in H_{0_1} and the odds-ratios on the sparse conditional distributions of the variables involved in H_{0_2}. We consider situations close to independencies the ones where more than the 50% of the odds-ratios stays in the mean interquartile range of the odd-ratios evaluated under the assumption of H_0.
4. We repeat step 1–3 10,000 times and count the number of times the simulated tables are closest to the independencies H_{0_1} and H_{0_2}.

The results for different n, c and f, in scenario 1, were displayed in Table 5.

The results in Table 5 show that any perturbation of the joint data has a strong affection also on the smallest marginal table of (X_1, X_2). Thus, by perturbing data, there is more propensity to reject H_{0_1} instead of H_{0_2} or both. Indeed, in sparse tables, the ones considered for testing H_{0_2}, often we have few information to reject the null hypothesis, and the results in Table 5 reflect this tendency. Similar results occur also in scenario 2, see Supplementary materials for details.

4 Application to real data

The EU-SILC (Statistics on Income and Living Conditions) of 2016 (Eurostat 2017) is one of the main data sources for the European Union's periodic reports on social circumstances and the prevalence of poverty in member countries. Here, we suggest exploiting the potentiality of the CGM to study how a set of pre-selected factors affect the poverty indicator. Specifically, we analyze the work force by considering individuals whose *self-defined current economic status* (variable PL031 in the survey) is (i) *employee working full-time*, (ii) *employee working part-time*, (iii) *self-employed working full-time*, (iv) *self-employed working part-time*, (v) *unemployed*, (vi) *permanently disabled or/and unfit to work*, or (vii) *fulfilling domestic tasks and care responsibilities*. In accordance with relevant literature (see, for instance, Molina and Rao (2010)) we select the following seven variables:

G Gender (1 = male, 2 = female);
A Age, categorized into four groups (1 = 16 $\vdash$ 36; 2 = 36 $\vdash$ 46; 3 = 46 $\vdash$ 55; 4 = 55 $\vdash$ 81);
E Educational level, as the highest ISCED level attained (000 = less than primary education; 100 = primary education, 200 = lower secondary education, 300 = upper secondary education, 400 = post-secondary education);
W Status in employment (1 = self-employed with employees, 2 = self-employed without employees, 3 = employee, 4 = family worker, 5 = unemployed);
M Marital status (1 = never married, 2 = married, 3 = separated, 4 = widowed, 5 = divorced);
H General health (1 = very good, 2 = good, 3 = fair, 4 = bad, 5 = very bad);
P Poverty indicator (0 = equivalized disposable income $\geq$ at risk of poverty threshold, 1 = equivalized disposable income $<$ at risk of poverty threshold).

The contingency table formed by these seven variables is composed of d =10,000 cells. Note that the survey involves different orders of magnitude of individuals across European countries and this gives rise to contingency tables, involving the variables under investigation, with different levels of sparsity. As an example, mosaic plots concerning the variables *gender*, *age* and the *poverty index* with respect to the countries Austria and Spain is given in the Supplementary materials as cases of high and low sparsity, respectively.

We can group the seven variables into three components according to their meanings and nature. Gender, G, and age, A, are biographical data and we thus collect these in the first component, C_1; the educational level, E, status in employment, W, marital status, M, and general health, H, all involve socio-economic aspects, and thus we collect them in the second component, C_2. Finally, the poverty index, P, is assigned to the third component, C_3. According to the relationships highlighted by previous sociological works, it may be interesting to study the underlying relationships among the poverty index and other variables, and to identify which factors have the greatest impact on P. In this case, a suitable model could include independence among the variables in the first group and the poverty status, conditionally to the variables in the second group. Unfortunately, this hypothesis implies constraints on a sparse joint distribution. Further, the independence statements implied by this hypothesis are all defined on sparse distributions. We can bypass the problem of the low power of this test by adding a further hypothesis in a smaller and denser marginal distribution, for instance, the plausible independence among the age and gender variables. This system of relationships is well represented by the graph in Fig. 4, where the features in the second group affect each other and all the variables within each component affect the variables in the next component.

The independencies implied by this CGM can be summarized with the following hypotheses:

$$\begin{aligned} H_{0_1} &: G \perp\!\!\!\perp A \\ H_{0_2} &: P \perp\!\!\!\perp GA|EWMH. \end{aligned} \tag{24}$$

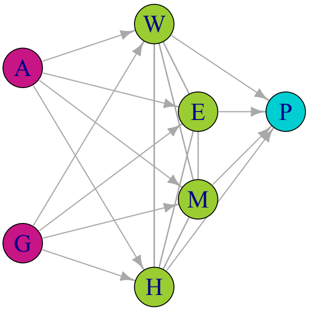

Fig. 4 CGM representing the two independencies $G \perp\!\!\!\perp A$ and $P \perp\!\!\!\perp GA|EWMH$. The three components are colored differently

One possible list of marginal sets that is compatible with the graph in Fig. 4 is $\{(GA); (GAEWMH); (GAEWMHP)\}$. In this way, we obtain marginal contingency tables of 8, 5000, and 10,000 cells, respectively.

We implement the redundant procedure to be more confident of the result of the test of H_0. However, in Sect. 3.1, we showed that when the global test $\boldsymbol{T_0}$ is applied to sparse tables, such as in this case, it has substantial limitations because of the low information that we can extract from the data, due to the sparseness. Indeed, to guarantee the set level of α^*, there is a low rejection rate when the hypothesis is not true. In light of this, we are more confident using the redundant testing procedure. To test the hypotheses we use both LR test and the generalization of the exact Fisher test, Agresti (2012). The results were concordant and are displayed in Table 6.

The second and third columns of Table 6 lists the behaviour of the countries with respect to the single component hypotheses H_{0_1} and H_{0_2}. The last three, the conclusion according to the three procedures discussed. In particular we reject the global hypothesis for p-values lower than 0.01 and the component hypotheses for p-values lower than 0.005.

Thus, according the *UI* and the *R* procedures, for the countries AT, CY, DE, DK, FI, NL, NO, RO, SE and SI we are more confident choosing the graph in Fig. 4 to represent the whole dataset.

In the other cases, with the exception of EL, the results of the global test contrast with those obtained by the *R* and *UI* procedures. Concerning the sparseness, we are more confident in the $\boldsymbol{T_R}$ approach. After the graph structure is identified, it is possible to analyse the marginal parameters, make opportune considerations, and draw conclusions. However, this analysis is outside the present paper's scope and is thus omitted.

5 Conclusions

The low power of tests involving sparse data is a well-known problem. This study proposes confronting this issue by exploiting the power of tests carried out on dense distributions and combining multiple tests. To do this, we used the UI procedure, and further proposed a new procedure called the redundant procedure. Undoubtedly, the redundant procedure brings similar results to the Union Intersection procedure

Table 6 List of countries where the hypotheses H_{0_1} and H_{0_2} have been retained/rejected and the consequent decision according to the three procedures

Decision	H_{0_1} $\alpha_1 = 0.005$	H_{0_2} $\alpha_2 = 0.005$	T_0 $\alpha = 0.01$	T_R $\alpha = (0.005, 0.005)$	T_{UI} $\alpha^* = (0.005, 0.005)$
Retain	AT, CY, DE, DK, FI, NL, NO, RO, SE, SI	AT, BE, BG, CY, CZ, DE, DK, EE, FI, FR, HR, HU, LT, LV, NL, NO, PL, PT, RO, RS, SE, SI, SK, UK	AT, BE, BG, CY, CZ, DE, DK, EE, ES, FI, FR, HR, HU, LT, LV, NL, NO, PL, PT, RO, RS, SE, SI, SK, UK	AT, CY, DE, DK, FI, NL, NO, RO, SE, SI	AT, CY, DE, DK, FI, NL, NO, RO, SE, SI
Reject	BE, BG, CZ, EE, EL, ES, FR, HR, HU, LT, LV, PL, PT, RS, SK, UK	EL, ES	EL	BE, BG, CZ, EE, EL, ES, FR, HR, HU, LT, LV, PL, PT, RS, SK, UK	BE, BG, CZ, EE, EL, ES, FR, HR, HU, LT, LV, PL, PT, RS, SK, UK

since both aim to gain power by exploiting the same component hypothesis, defined in the denser table. However, the crucial difference lies in damage control when things do not work well. In fact, the trick used in the redundant procedure ensures that we do not have minimal performance that far off from the global test performed with a smaller level of confidence. This aspect, while seemingly minor, ensures that we can enunciate the conditions under which the redundant procedure out-performs standard inference, in terms of test power, which is not possible for the *UI* procedure. In addition, the redundant procedure is characterized by a more restrictive assumption than the *UI* procedure. Obviously, when the hypotheses characterizing the tests define the parameters on very similar spaces, the performance of the two procedures does not differ much. We have deepened simulative studies to investigate the performance of the two procedures under consideration. In the first case, where the sparsity of the table seems to greatly affect the performance of the classical ML test, we have the best results for the two procedures examined. In the second case, however, the three tests have very similar not so bad performance and there is no strong evidence of improvement. Thus, there seem to be two factors affecting the good performance of the methods discussed here. First, the penalizing sparsity on classical inference and second, the marginal choice to gain power in the view of these composite tests. Finally, we apply the proposed procedure to analyze the EU-SILC dataset. Our application stops after identifying plausible independencies and demonstrating the strategy we have developed.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00362-023-01473-6.

Funding Open access funding provided by Politecnico di Milano within the CRUI-CARE Agreement.

Appendix

Proof of Theorem 1 By assuming a distribution belonging to $\mathcal{M}_0$, the right-hand side of (2) together with (3) imply the level result in (4). The power result in (5) is implied similarly by the left-hand side of (2) and it applies to the entire alternative hypothesis. □

Proof of Theorem 2 To prove the inequality in formula (12), according to (11), we need to consider a probability type I error (α_i) for the test T_i, such that the corresponding critical value (c_i) satisfies the following inequality:

$$c_{\alpha^*}^{\mathcal{M}_{0_i}}(T_i) < c_i < c_{\beta^*}^{\mathcal{M}_{1_i}}(T_i). \tag{25}$$

This situation is represented in Fig. 1 (bottom). Remember that, given two critical values c_1 and c_2, such that $c_1 < c_2$, then

$$P(T(c_1) = rej|\mathcal{M}) > P(T(c_2) = rej|\mathcal{M}). \tag{26}$$

In force of the (26), we obtain

$$P(T_i(c_i^{\mathcal{M}_{1_i}}) = rej|\mathcal{M}_{1_i}) > P(T_i(c_{\beta^*}^{\mathcal{M}_{1_i}}) = rej|\mathcal{M}_{1_i}) = \beta^*. \tag{27}$$

This inequality easily comes by looking at Fig. 1 (bottom) where the dashed blue area increases for any values lower than $c_{\beta^*}^{\mathcal{M}_{1_i}}(T_i)$.
According to the definition, $c_{\beta^*}^{\mathcal{M}_1}(T_g)$ is the critical value such that $P(T_g(c_{\beta^*}^{\mathcal{M}_1}) = rej|\mathcal{M}_1) = \beta^*$, thus

$$P(T_i(c_i^{\mathcal{M}_{1_i}}) = rej|\mathcal{M}_{1_i}) > P(T_g(c_{\beta^*}^{\mathcal{M}_1}) = rej|\mathcal{M}_1). \tag{28}$$

Further, by comparison of the critical values, one obtains from (10) that

$$P\left\{T_g(c_{\beta^*}^{\mathcal{M}_1}) = rej|\mathcal{M}_1\right\} \geq P\left\{T_g(c_{\alpha^*}^{\mathcal{M}_0}) = rej|\mathcal{M}_1\right\}, \tag{29}$$

that is, the power (β^*) of the test T_g is greater than or equal to the power of the test T_0 carried out with a probability of type I error equal to α^*.
The (27) and (29) imply that

$$P(T_i(c_i^{\mathcal{M}_{1_i}}) = rej|\mathcal{M}_{1_i}) > P\left\{T_g(c_{\alpha^*}^{\mathcal{M}_0}) = rej|\mathcal{M}_1\right\} = P\left\{T_0(c_{\alpha^*}^{\mathcal{M}_0}) = rej|\mathcal{M}_1\right\},$$

because T_g tests exactly the null hypothesis H_0.
According to (6), the power of the redundant test $T_R(\alpha_g, \alpha_i)$ is greater or equal to the power of $T_i(\alpha_i)$. As a consequence, the power result in (12) holds, irrespective of the choice of α_g.
To prove the inequality in formula (13), let us start by considering the inequality in (7) when the H_0 hypothesis is true:

$$P(T_R(\alpha_g, \alpha_i) = rej|\mathcal{M}_0) \leq P(T_g(\alpha_g) = rej|\mathcal{M}_0) + P(T_i(\alpha_i) = rej|\mathcal{M}_0).$$

The two addends in the r.h.s., by definition, are α_g and α_i, then their sum is less or equal to α^*, according to (8).
Finally, the α^* is the probability of type I error that we impose in the likelihood ratio test we use to test H_0, which definition is $P(T_0(\alpha^*) = rej|\mathcal{M}_0)$. □

References

Agresti A (2012) Categorical data analysis, vol 792. Wiley, New York

Agresti A, Gottard A (2007) Independence in multi-way contingency tables: S.N. Roy's breakthroughs and later developments. J Stat Plan Inference 137(11):3216–3226

Bartolucci F, Colombi R, Forcina A (2007) An extended class of marginal link functions for modelling contingency tables by equality and inequality constraints. Stat Sin 17(2):691–711

Belilovsky E, Kastner K, Varoquaux G, Blaschko MB (2017) Learning to discover sparse graphical models. In: International conference on machine learning, pp 440–448

Bergsma WP, Rudas T (2002) Marginal models for categorical data. Ann Stat 30(1):140–159

Colombi R, Giordano S, Cazzaro M (2014) hmmm: an R package for hierarchical multinomial marginal models. J Stat Softw 59(11):1–25

Cox DR, Wermuth N (1996) Multivariate dependencies: models, analysis and interpretation, vol 67. CRC Press, Boca Raton

Cressie N, Read TR (1989) Pearson's χ^2 and the loglikelihood ratio statistic G^2: a comparative review. Int Stat Rev 57(1):19–43

Csardi G, Nepusz T (2006) The igraph software package for complex network research. InterJournal Complex Syst 1695:1–9

Dale JR (1986) Asymptotic normality of goodness-of-fit statistics for sparse product multinomials. J R Stat Soc Ser B 48(1):48–59

Drton M (2009) Discrete chain graph models. Bernoulli 15(3):736–753

Eurostat (2017) Eu-silc user database, description version 2016

Fienberg SE, Rinaldo A (2012) Maximum likelihood estimation in log-linear models. Ann Stat 40(2):996–1023

Gabriel KR (1969) Simultaneous test procedures-some theory of multiple comparisons. Ann Math Stat 40(1):224–250

Henao R, Winther O (2009) Bayesian sparse factor models and dags inference and comparison. In: Advances in neural information processing systems, pp 736–744

Kim S-H, Choi H, Lee S (2009) Estimate-based goodness-of-fit test for large sparse multinomial distributions. Comput Stat Data Anal 53(4):1122–1131

Koehler KJ (1986) Goodness-of-fit tests for log-linear models in sparse contingency tables. J Am Stat Assoc 81(394):483–493

Lauritzen SL, Richardson TS (2002) Chain graph models and their causal interpretations. J R Stat Soc Ser B 64(3):321–348

Maathuis M, Drton M, Lauritzen S, Wainwright M (2018) Handbook of graphical models. CRC Press, Boca Raton

Marchetti GM, Lupparelli M (2011) Chain graph models of multivariate regression type for categorical data. Bernoulli 17(3):827–844

Maydeu-Olivares A, Joe H (2005) Limited-and full-information estimation and goodness-of-fit testing in 2^n contingency tables: a unified framework. J Am Stat Assoc 100(471):1009–1020

Maydeu-Olivares A, Joe H (2006) Limited information goodness-of-fit testing in multidimensional contingency tables. Psychometrika 71(4):713

Mehta CR, Patel NR (1983) A network algorithm for performing Fisher's exact test in r×c contingency tables. J Am Stat Assoc 78(382):427–434

Mieth B, Kloft M, Rodríguez JA, Sonnenburg S, Vobruba R, Morcillo-Suárez C, Farré X, Marigorta UM, Fehr E, Dickhaus T et al (2016) Combining multiple hypothesis testing with machine learning increases the statistical power of genome-wide association studies. Sci Rep 6:36671

Molina I, Rao JNK (2010) Small area estimation of poverty indicators. Can J Stat 38(3):369–385

Nicolussi F, Cazzaro M (2021) Context-specific independencies in stratified chain regression graphical models. Bernoulli 27(3):2091–2116

Nicolussi F, Colombi R (2017) Type ii chain graph models for categorical data: a smooth subclass. Bernoulli 23(2):863–883

Perlman MD, Wu L (2003) On the validity of the likelihood ratio and maximum likelihood methods. J Stat Plan Inference 117(1):59–81

R Core Team (2016) R: a Language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Roverato A (2015) Log-mean linear parameterization for discrete graphical models of marginal independence and the analysis of dichotomizations. Scand J Stat 42(2):627–648

Roy SN (1953) On a heuristic method of test construction and its use in multivariate analysis. Ann Math Stat 24(2):220–238

Roy SN, Mitra SK (1956) An introduction to some non-parametric generalizations of analysis of variance and multivariate analysis. Biometrika 43(3–4):361–376

Rudas T (1986) A Monte Carlo comparison of the small sample behaviour of the Pearson, the likelihood ratio and the Cressie-Read statistics. J Stat Comput Simul 24(2):107–120

Rudas T, Bergsma WP, Németh R (2010) Marginal log-linear parameterization of conditional independence models. Biometrika 97(4):1006–1012

Sedgewick AJ, Shi I, Donovan RM, Benos PV (2016) Learning mixed graphical models with separate sparsity parameters and stability-based model selection. BMC Bioinform 17(S5):S175

Venables WN, Ripley BD (2002) Modern applied statistics with S, 4th ed. Springer, New York. ISBN 0-387-95457-0

Yoshida R, West M (2010) Bayesian learning in sparse graphical factor models via variational mean-field annealing. J Mach Learn Res 11:1771–1798

Zelterman D (1987) Goodness-of-fit tests for large sparse multinomial distributions. J Am Stat Assoc 82(398):624–629