




Bayesian network structural learning from complex survey data: a resampling based approach

Daniela Marella¹ · Paola Vicard² 

Accepted: 29 November 2021 / Published online: 21 January 2022
© Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract

Nowadays there is increasing availability of good quality official statistics data. The construction of multivariate statistical models possibly leading to the identification of causal relationships is of interest. In this context Bayesian networks play an important role. A crucial step consists in learning the structure of a Bayesian network. One of the most widely used procedures is the PC algorithm consisting in carrying out several independence tests on the available data set and in building a Bayesian network according to the tests results. The PC algorithm is based on the irremissible assumption that data are independent and identically distributed. Unfortunately, official statistics data are generally collected through complex sampling designs, then the aforementioned assumption is not met. In such a context the PC algorithm fails in learning the structure. To avoid this, the sample selection must be taken into account in the structural learning process. In this paper, a modified version of the PC algorithm is proposed for inferring causal structure from complex survey data. It is based on resampling techniques for finite populations. A simulation experiment showing the robustness with respect to departures from the assumptions and the good performance of the proposed algorithm is carried out.

Keywords Bayesian network · Complex survey data · Pseudo-population · Resampling · Structural learning

✉ Paola Vicard
paola.vicard@uniroma3.it

Daniela Marella
daniela.marella@uniroma1.it

¹ Department of Social and Economic Sciences, Sapienza Università di Roma, Piazzale Aldo Moro 5, 00185 Rome, Italy

² Department of Economics, Università Roma Tre, Via Silvio D'amico 77, 00145 Rome, Italy

1 Introduction

Nowadays there is increasing availability of good quality data produced by official statistics, and useful for both study purpose and decision making. Simple but important examples are the EU-SILC survey on income, the ISTAT survey on labour forces, the ISTAT survey on consumption expenses, etc.... Typically in those surveys data are collected with unequal first order inclusion probabilities. In all the above mentioned cases, as well as in many others, the construction of multivariate statistical models involving several variables is of primary interest. In this context Bayesian networks (BN) play a very important role because of their flexibility and easy-to-read representation of relationships among variables by means of edges connecting them. Bayesian networks are multivariate statistical models satisfying sets of conditional independence statements contained in a directed acyclic graph (DAG), see Cowell et al. (2007). The nodes of the graph correspond to random variables, while edges represent dependencies. In recent years BNs have been successfully applied to a large variety of contexts; among them official statistics. Data collected through a survey are typically affected by selection bias due to sampling design, nonresponse and measurement error. BNs appeared to be very useful in missing data imputation (Di Zio et al. (2004), Thibaudeau and Winkler (2002)), contingency table estimation for complex survey sampling (Ballin and Scanu (2010)) and measurement errors (Marella and Vicard (2013), Marella and Vicard (2015)). However, there are still serious obstacles complicating a wider application in official statistics contexts and in all those contexts where complex sampling designs are used. In fact, missing item imputation and measurement error correction can be performed once the BN structure is available (either known in advance or learned from data) since information has to be propagated throughout the network. Therefore, it is necessary to develop structural learning algorithms accounting for the sampling design complexity.

Learning BNs from a sample can be a time consuming task and a challenging issue even when data are independent and identically distributed (*i.i.d.*). For a survey on structural learning, see Drton and Maathuis (2017). In case of data driven learning, three broad classes of algorithms can be distinguished: constraint based algorithms, score-plus-search and hybrid algorithms. Constraint based algorithms carry out a series of independence tests and construct a graph satisfying the discovered independence statements. The second group of algorithms considers structure learning as a structural optimization problem, using a search strategy to select the structure optimizing a given score function which measures the fitting degree of network and data. Recently, the third group of algorithms, hybrid algorithms, has become widely used. They combine the ideas and the advantages of the previous two types of algorithms. One popular strategy is to use constraint-based algorithms (in the first learning phase) to determine the initial network structure, and then use score-based algorithms (in the second learning phase) to find the highest-scoring network structure, see Tsamardinos et al. (2006). The main constraint based algorithm is the PC algorithm, see Spirtes et al. (2000). It has several advantages, among which an intuitive basis. The PC algorithm is based on conditional

independence tests, usually performed by using the standard Pearson chi-square test statistic under *i.i.d.* assumption, which is equivalent to simple random sampling assumption. However, sample selection in surveys may involve more complex sampling designs based on stratification, different level of clustering and inclusion probabilities proportional to an appropriate size measure. Complex designs can severely impact on *i.i.d.* based method, as shown in Skinner et al. (1989). In such circumstances, the standard test procedure is not valid even asymptotically. As a consequence, in these cases the PC algorithm fails in correctly identifying the independence equivalence class of DAGs containing the true structure of the BN under examination.

Survey weights and design effects are appropriate tools by which complex sampling designs can be accommodated. As far as the chi-square statistic is concerned, corrections based on the design effects have been proposed in Rao and Scott (1981), Rao and Scott (1984). However, such corrections require design effect or, alternatively, full covariance matrix estimates of the cell proportions estimators that are not generally delivered and then need to be estimated by resampling.

In this paper, a novel approach for inferring causal structure from complex survey data is investigated requiring to combine two different statistical cultures: sampling theory and structural learning. A modified version of the PC algorithm (PC-cs algorithm, for short) that uses independence tests accounting for the sample selection mechanism is proposed. The sampling design complexity is accounted for *via* a design-based approach by including the sampling weights in the BN parameters estimates. In fact, such weights contain invaluable information about the relationship between the distribution of the sample data and the distribution in the population from which the sample is taken. After having estimated such parameters, a procedure based on the chi-square statistic for testing the association in a two-way table is proposed; its limiting sampling distribution is estimated by resorting to resampling techniques for finite populations. The new test procedure is applied to BN structural learning.

The paper is organized as follows. In Sect. 2 the PC algorithm for *i.i.d.* data and the basic assumptions on which it relies are briefly recalled. In Sect. 3 the modified version of the PC algorithm for complex survey data is introduced and described. A simulation study is performed in Sect. 4. Finally, advantages and disadvantages of the proposed approach are discussed in Sect. 5.

2 Discovering causal structure with the PC algorithm

2.1 Preliminary definitions

A DAG is a pair $G = (V, E)$ consisting of a set of vertices V and a set of directed edges between pairs of nodes.

A directed graph is acyclic in the sense that it is not possible to start from a node and go back to the same node following arrows directions. Each node represents a random variable, while a missing arrow between two nodes implies (un)conditional independence between the corresponding variables. Examples of DAG are shown in Fig. 1a, b. Consider Fig. 1a. In the arrow $X_2 \rightarrow X_3$, X_2 is said parent of X_3 and X_3 is

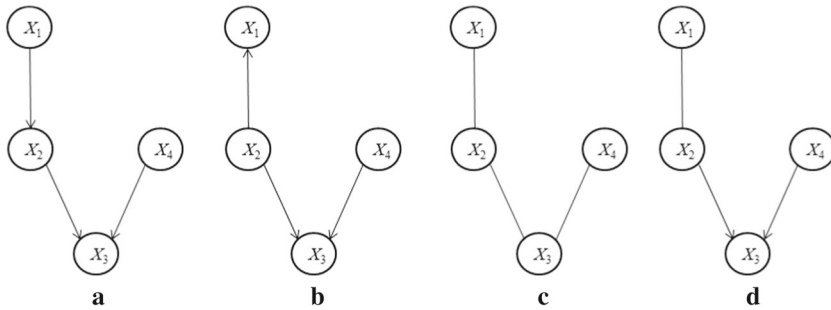


Fig. 1 **a** Example of DAG, **b** DAG Markov equivalent to **(a)**; **c** skeleton of DAGs in **(a)** and **(b)**; **d** completed partially DAG, CPDAG for DAGs **(a)**, **(b)**

said child of X_2 . X_1 is called an ancestor of X_3 and X_3 a descendant of X_1 since X_1 is connected to X_3 by a directed path, *i.e.* a sequence of direction preserving arrows. Two nodes X_i and X_j are said adjacent if they are directly connected by a directed or undirected edge. A graph is complete if all vertices are joined by a directed or undirected edge. A v -structure is a triple of nodes (X_i, X_k, X_j) such that the arrows $X_i \rightarrow X_k$ and $X_j \rightarrow X_k$ are in the DAG. In such configuration, X_k is said unshielded collider. If X_i and X_j are connected then the collider X_k is termed shielded. For example, in Fig. 1a the triple (X_2, X_3, X_4) constitutes a v -structure where X_3 is an unshielded collider. The skeleton of a DAG G is the undirected graph obtained from G by replacing all arrows with lines (undirected edges).

For example, the undirected graph in Fig. 1c is the skeleton of the graphs in Fig. 1a, b.

We next recall the basic assumptions and then the main steps of the PC algorithm.

2.2 Assumptions

Let P be the joint probability distribution associated to G . The following assumptions are set when applying the PC algorithm (for details see Zhang and Spirtes (2008) and Uhler et al. (2013)):

1. *Sufficiency Condition.* The set of observed variables V is causally sufficient, that is every common direct cause relative to V of any pair of variables in V is also contained in V .
2. *Causal Markov Condition.* P is said to be Markov with respect to G if a node of G is probabilistically independent of its non descendants given its parents in G . For example, in Fig. 1a, X_3 is independent of X_1 given its parent X_2 . The causal Markov condition defines the set of conditional independence relations entailed by the DAG.
3. *Causal Faithfulness Condition.* The joint probability distribution P is faithful to G if all conditional independencies can be read off G . This means that if the true causal structure G does not entail a conditional independence relation according to the causal Markov condition, then the conditional independence relation does

not hold for the true probability distribution. In Zhang and Spirtes (2008) a decomposition of the faithfulness assumption useful for violation detection is proposed. The faithfulness condition implies adjacency-faithfulness and orientation-faithfulness. These do not constitute an exhaustive decomposition of the faithfulness assumption. However, the leftover part is irrelevant to the correctness of structural learning procedures such as PC algorithm.

2.3 PC algorithm for i.i.d. data

The PC algorithm starts with a complete undirected graph on the set V , and ends with a class of Markov equivalent DAGs where all the associated models encode the same conditional independence information.

Two DAGs are Markov equivalent if and only if they have the same skeleton and the same v -structures, Verma and Pearl (1990). For example DAGs a and b in Fig. 1 are Markov equivalent. In fact, they have the same skeleton (Fig. 1c) and the same v -structure $X_2 \rightarrow X_3 \leftarrow X_4$. A common tool for visualizing equivalence classes of DAGs are the completed partially directed acyclic graphs (CPDAG), see Spirtes et al. (2000). A CPDAG is a *summary* graph that has: a directed edge where all DAGs in the equivalence class have the same directed edge; an undirected edge between X_i and X_j if in the equivalence class there exist at least a DAG with $X_i \rightarrow X_j$ and a DAG with $X_i \leftarrow X_j$. An example of CPDAG is shown in Fig. 1d.

The PC algorithm proceeds according to the following two phases:

- Phase 1 *Skeleton estimation*. First, all pairs (X_i, X_j) are tested for marginal independence removing the edge between independent variables and saving the empty set as separating sets S_{ij} and S_{ji} . Then all the pairs, say (X_i, X_j) , still adjacent, are tested for independence conditionally on one single node adjacent to X_i . If X_i and X_j are judged to be independent given, say, X_k , the edge between X_i and X_j is removed and X_k is saved as separating sets S_{ij} and S_{ji} . The algorithm proceeds augmenting, one unit at a time, the conditioning set size until all adjacency sets are smaller than the conditioning set size. The resulting graph is the skeleton.
- Phase 2 *Arrows orientation*. The v -structures and their colliders are identified. A triple of vertices (X_i, X_k, X_j) in the skeleton such that the pairs (X_i, X_k) and (X_j, X_k) are adjacent but (X_i, X_j) is not, is oriented as a v -structure $X_i \rightarrow X_k \leftarrow X_j$ if X_k is not in $S_{ij} = S_{ji}$. Once all v -structures have been identified, it may be possible to orient some of the remaining edges, without introducing additional v -structures or directed cycles, thus leading to a CPDAG.

As far as the faithfulness assumption is concerned, adjacency-faithfulness condition is necessary to recover the skeleton of the true DAG and it is strictly related to Phase 1 of the PC algorithm. The orientation-faithfulness condition, necessary for finding the correct arrows orientation, is related to Phase 2 of the algorithm.

3 PC algorithm for complex survey data

3.1 The problem

Under the assumptions in Sect. 2.2 and if the sample size is large enough, the original PC algorithm is able to infer from data the Markov equivalence class the true causal DAG belongs to. This means that, if the input of the PC algorithm is a sample from a population distribution P faithful to some DAG, then in large sample limit, the algorithm can identify any probabilistic independence claim with perfect reliability. In practice, we do not have direct access to the true population distribution, and need to do statistical inference based on finite sample size. Hence, since the first phase of the PC algorithm consists in a series of conditional independence tests based on a finite sample size, it is possible that the original graph is not recovered even if the PC algorithm assumptions are verified at the population level. Therefore, it becomes very relevant to causal inference whether the population probability distribution, though faithful to the true causal structure, is far from or close to being unfaithful. As stressed in Zhang and Spirtes (2008), a population distribution is close-to-unfaithful to a causal structure, if the structure does not entail some conditional independence relation according to the Causal Markov Condition, but the conditional independence almost holds, or in other words, the conditional dependence is by some measure very weak in the population. Precisely, the dependence degree and the sample size determine how “weak” counts as “close to independence”. It is clear that at every finite sample size, there are distributions faithful to the true causal structure but so close to being unfaithful as to possibly cause troubles for inference at that sample size. For instance, two variables, though entailed to be dependent conditional on some variables, can be close to be conditionally independent. As a consequence, due to sample size, tests can fail to correctly identify such a dependence leading to errors in judgment about the properties of the population.

The situation worsens for complex survey data since the sampling design can modify independence and conditional independence relations associated with the population probability distribution P . Roughly speaking, complex sampling schemes lead to unequal selection probabilities; ignoring this can result in biased estimates of the population distribution P . As a consequence, even if the population distribution is faithful to some DAG G , the sample distribution may not be faithful to the same DAG because of the sample selection mechanism. Specifically, the selection probabilities are often unequal in at least some stages of the sample selection. When such probabilities are correlated with the survey variables of interest, the observed outcomes are no longer representative of the population outcomes and the model holding for the sample data is then different from the model holding in the population, see Pfeffermann (2001). In this case, conventional analysis, which ignores the sampling effects, may yield large bias and erroneous inference, as illustrated, for example, in the book edited by Skinner et al. (1989). In Marella and Pfeffermann (2019) the effect of alternative sampling designs on a trivariate normal population distribution has been evaluated. For instance, suppose

that (X, Y, Z) is a trivariate normal with mean vector μ and covariance matrix Σ , and that the sample inclusion probabilities have expectations

$$E_p(\pi_i|x_i, y_i, z_i) = \kappa \exp\{\gamma_X x_i + \gamma_Y y_i + \gamma_Z z_i\}, \tag{1}$$

where κ guarantees that the expectation is less or equal to one. It can be shown that the model holding for the sample outcomes is normal but with parameters μ_S and Σ , see Marella and Pfeffermann (2019). In this case, the sample model and the population model belong to the same family and differ only in some parameters. Clearly, other examples in Marella and Pfeffermann (2019) show that the population and the sample distributions can be either in the same family and differ only in some or all the parameters, or in different families.

In the sequel, we assume that the design variables used for sample selection are known for all the sample units so that the sufficiency condition is satisfied, as in Ballin and Scanu (2010). Nevertheless, the Markov and the faithfulness condition may fail if the sample is selected by a procedure that is biased towards two or more variables in the set V .

In the PC algorithm for complex survey data the skeleton learning step (Phase 1) of the PC algorithm is modified by introducing a procedure for testing association in a two-way table for data coming from complex sample surveys. Such a procedure is introduced in Sect. 3.2 where the existence of a limiting distribution of the test statistic under the independence null hypothesis is proved. In paragraph 3.3 such a distribution is estimated by resampling methods for finite population.

3.2 Independence test for complex sample surveys

In this section a test for independence based on the chi-square statistic in a two-way table for complex survey data is described. The test relies on some results in Conti et al. (2018) where a resampling technique allowing to make inference on the superpopulation parameters in a finite population setting is proposed.

Denote by A and B two characters of interest with H (A^1, \dots, A^H) and K (B^1, \dots, B^K) categories, respectively. Furthermore, let the superpopulation parameters p^{hk}, p^h, p^k be defined as

$$\begin{aligned} p^{hk} &= \text{Prob}(A = A^h, B = B^k), & p^h &= \text{Prob}(A = A^h) \\ p^k &= \text{Prob}(B = B^k) \end{aligned} \tag{2}$$

for $h = 1, \dots, H$ and $k = 1, \dots, K$. Let \mathcal{U}_N be a finite population of size N generated from the superpopulation model (2), labeled by integers $1, \dots, N$. For each unit i , let Y_i^h (Y_i^k) be the indicator variable taking value 1 if the unit i assumes the modality A^h (B^k) and 0 otherwise, for $h = 1, \dots, H$ ($k = 1, \dots, K$). Let $Y_i^{hk} = Y_i^h Y_i^k$ so that for each unit i the following equalities hold

$$\sum_{h=1}^H Y_i^h = \sum_{k=1}^K Y_i^k = 1, \quad \sum_{h=1}^H Y_i^{hk} = Y_i^k, \quad \sum_{k=1}^K Y_i^{hk} = Y_i^h. \quad (3)$$

For each unit i in the population \mathcal{U}_N , let D_i be the sample membership indicator, *i.e.* a Bernoulli random variable taking value 1 whenever i is in the sample, and 0 otherwise, and let $\mathbf{D}_N = (D_1, \dots, D_N)$ be the sample membership random vector for the population units. An unordered, without replacement sampling design S is the probability distribution of \mathbf{D}_N . In particular $\pi_i = E_S[D_i]$ is the first order inclusion probability of unit i . The suffix S denotes the sampling design used to select population units. The sample selection, and therefore π_i s, depends on the values of design variables (like strata, cluster indicator variables or size measures) that are statistically related to Y_i^h and Y_i^k but are not included in the inference model. Consequently, the distribution holding for the sample data may be very different from the distribution in the population. In other words, even if the Y_i^h , Y_i^k , $i = 1, \dots, N$, are *i.i.d.* at a population level, they are not *i.i.d.* at a sample level due to the sampling design (Pfeffermann (1993)).

The effective sample size is the r.v. $n_s = D_1 + \dots + D_N$. In the sequel we will confine ourselves to fixed size sampling designs, such that $n_s \equiv n$. Assumptions on the sampling design according to which the sample is drawn, are similar to those used in Conti (2014), Conti and Marella (2015a) and Conti et al. (2019) (assumptions A1-A6). From now on we will assume that the sampling design possesses asymptotically maximal entropy. More precisely, consider the Poisson sampling design where the random variables D_i s are independent with $E_S[D_i] = \pi_i$, and $n = \pi_1 + \dots + \pi_N$. The *rejective sampling* (or *normalized conditional Poisson sampling*) is obtained by conditioning the Poisson design to $n_s = n$. Denote by $H(S) = E_S[\log Pr_S(\mathbf{D}_N)]$ the entropy of sampling S . In particular, the rejective design possesses maximal entropy among all sampling designs with fixed sample size and fixed first order inclusion probabilities. Let $Pr_R(\cdot)$ be the sample probabilities for the rejective sampling, the Kullback-Leibler divergence of Pr_S from Pr_R can be written as,

$$\Delta_{KL}(Pr_S||Pr_R) = H(Pr_R) - H(Pr_S). \quad (4)$$

In the sequel we assume that $\Delta_{KL}(Pr_S||Pr_R) \rightarrow 0$ as $n, N \rightarrow \infty$. This means that the sampling design S possesses high (asymptotically maximal) entropy.

The properties of high entropy sampling designs are discussed in Berger (2011), Grafström (2010) and references therein. As a matter of fact, asymptotics for high entropy sampling designs only depend on their first order inclusion probabilities. In other terms, if two high entropy designs have the same first order inclusion probabilities, then they have the same asymptotic behaviour. High entropy sampling designs generate highly randomized samples, which in turn make the design more robust. A discussion about high entropy designs and their relationship with robustness can be found in Grafström (2010). Examples of maximal asymptotic entropy sampling designs are simple random sampling, Rao-Sampford design, Chao design, stratified design, etc., see Berger (2011).

3.2.1 Parameters estimators and their distributions

Let p_N^{hk}, p_N^h, p_N^k be the finite population parameters defined as

$$\begin{aligned}
 p_N^{hk} &= \frac{1}{N} \sum_{i=1}^N Y_i^{hk}, p_N^h = \frac{1}{N} \sum_{i=1}^N Y_i^h = \sum_{k=1}^K p_N^{hk}, \\
 p_N^k &= \frac{1}{N} \sum_{i=1}^N Y_i^k = \sum_{h=1}^H p_N^{hk}
 \end{aligned}
 \tag{5}$$

where $h = 1, \dots, H, k = 1, \dots, K$. In complex sample design sample proportions would result in inconsistent estimators of the population parameters, Pfeffermann (1993). Parameters (5) can be estimated using the classical Hájek estimators (Hájek (1964))

$$\begin{aligned}
 \hat{p}^{hk} &= \frac{\sum_{i=1}^N \frac{D_i Y_i^{hk}}{\pi_i}}{\sum_{i=1}^N \frac{D_i}{\pi_i}}, \quad \hat{p}^h = \frac{\sum_{i=1}^N \frac{D_i Y_i^h}{\pi_i}}{\sum_{i=1}^N \frac{D_i}{\pi_i}} = \sum_{k=1}^K \hat{p}^{hk}, \\
 \hat{p}^k &= \frac{\sum_{i=1}^N \frac{D_i Y_i^k}{\pi_i}}{\sum_{i=1}^N \frac{D_i}{\pi_i}} = \sum_{h=1}^H \hat{p}^{hk}
 \end{aligned}
 \tag{6}$$

for $h = 1, \dots, H, k = 1, \dots, K$, where $1/\pi_i$ is the sampling weight, that is the reciprocal of the probability that the unit i is included in the sample. Note that, in the simple random sampling case, the estimators (6) reduce to the proportion of units in the sample belonging to categories (h, k) , h and k , respectively.

The existence of the limiting distribution of the Hájek estimators (6) is studied in Conti et al. (2018), under both the sample selection and the population generation randomness sources. Similar results are in Boistard et al. (2017) where no resampling schemes allowing to recover the limiting distribution of the chi-square test statistic (14) defined in Sect. 3.2.2 are proposed. Here, similarly to Conti et al. (2018), we analyze the behaviour of the stochastic processes

$$W^{HK} = \{\sqrt{n}(\hat{p}^{hk} - p^{hk}), h = 1, \dots, H, k = 1, \dots, K\}, \tag{7}$$

$$W^H = \{\sqrt{n}(\hat{p}^h - p^h), h = 1, \dots, H\}, \tag{8}$$

$$W^K = \{\sqrt{n}(\hat{p}^k - p^k), k = 1, \dots, K\}. \tag{9}$$

Each of these processes can be partitioned as follows

$$\begin{aligned}
 W^{HK} &= \{W_n^{HK} + W_N^{HK}, h = 1, \dots, H, k = 1, \dots, K\} \\
 &= \{\sqrt{n}(\hat{p}^{hk} - p_N^{hk}) + \sqrt{f}\sqrt{N}(p_N^{hk} - p^{hk})\},
 \end{aligned}
 \tag{10}$$

$$\begin{aligned} W^H &= \{W_n^H + W_N^H, h = 1, \dots, H\} \\ &= \{\sqrt{n}(\widehat{p}^{h\cdot} - p_N^{h\cdot}) + \sqrt{f}\sqrt{N}(p_N^{h\cdot} - p^{h\cdot})\}, \end{aligned} \quad (11)$$

$$\begin{aligned} W^K &= \{W_n^K + W_N^K, k = 1, \dots, K\} \\ &= \{\sqrt{n}(\widehat{p}^{\cdot k} - p_N^{\cdot k}) + \sqrt{f}\sqrt{N}(p_N^{\cdot k} - p^{\cdot k})\} \end{aligned} \quad (12)$$

for $h = 1, \dots, H, k = 1, \dots, K$ and $f = \frac{n}{N}$. The components W_n^{HK}, W_n^H, W_n^K depend on the sample selection randomness while the components W_N^{HK}, W_N^H, W_N^K depend on the superpopulation randomness. Proposition 1 establishes the convergence of processes (7) – (9) to Gaussian distributions.

Proposition 1 *Under the assumptions A1-A6 of Proposition 1 in Conti et al. (2018), as n and N increase, the sequences:*

1. W_n^{HK} and W_N^{HK} converge in distribution to degenerate multivariate normal distributions with zero mean vector and singular covariance matrices Σ_1^{HK} and Σ_2^{HK} of order HK , respectively. As a consequence, the whole process W^{HK} converges to a degenerate multivariate normal distribution with zero mean vector and singular covariance matrix $\Sigma^{HK} = \Sigma_1^{HK} + f\Sigma_2^{HK}$.
2. W_n^H and W_N^H converge in distribution to degenerate multivariate normal distributions with zero mean vector and singular covariance matrices Σ_1^H and Σ_2^H of order H , respectively. As a consequence, the whole process W^H converges to a degenerate multivariate normal distribution with zero mean vector and singular covariance matrix $\Sigma^H = \Sigma_1^H + f\Sigma_2^H$.
3. W_n^K and W_N^K converge in distribution to degenerate multivariate normal distributions with zero mean vector and singular covariance matrices Σ_1^K and Σ_2^K of order K , respectively. As a consequence, the whole process W^K converges to a degenerate multivariate normal distribution with zero mean vector and singular covariance matrix $\Sigma^K = \Sigma_1^K + f\Sigma_2^K$.

The proof rests on the same ideas as the proof of Proposition 1 in Conti et al. (2018) and it can be seen as a direct consequence. The proof guidelines are reported in the Appendix.

3.2.2 The independence test statistic

Suppose to test the null hypothesis that the two categorical variables A and B are independent, against the alternative hypothesis that they are associated. Formally

$$\mathcal{H}_0 : p^{hk} = p^h \cdot p^k \quad \text{against} \quad \mathcal{H}_1 : p^{hk} \neq p^h \cdot p^k. \quad (13)$$

The used test statistic is

$$\chi_H^2 = n \sum_{h=1}^H \sum_{k=1}^K \frac{(\hat{p}^{hk} - \hat{p}^h \cdot \hat{p}^k)^2}{\hat{p}^h \cdot \hat{p}^k} \tag{14}$$

where the sampling weights in the Hájek estimators \hat{p}^{hk} , \hat{p}^h and \hat{p}^k (6) compensate for different selection probabilities. As stressed in Proposition 3, for complex survey data: (i) the statistic (14) does have a limiting distribution; (ii) the limiting distribution does not necessarily approach a chi-square distribution due to the singularity of the covariance matrices Σ^{HK} , Σ^H and Σ^K in Proposition 1. From Proposition 1, the following result holds.

Proposition 2 *Let $\mathbf{p}^{HK} = (p^{11}, \dots, p^{1K}, \dots, p^{H1}, \dots, p^{HK})$, $\mathbf{p}^{H\cdot} = (p^1, \dots, p^{H\cdot})$, $\mathbf{p}^{\cdot K} = (p^{\cdot 1}, \dots, p^{\cdot K})$ be vectors of length HK , H and K , respectively, and denote by $\hat{\mathbf{p}}^{HK}$, $\hat{\mathbf{p}}^H$ and $\hat{\mathbf{p}}^{\cdot K}$ their estimates. Furthermore, let*

$$\hat{\mathbf{p}} = (\hat{\mathbf{p}}^{HK}, \hat{\mathbf{p}}^H, \hat{\mathbf{p}}^{\cdot K})$$

and

$$\mathbf{p}^{\mathcal{H}_0} = (p^1 \cdot p^1, \dots, p^H \cdot p^K, p^1, \dots, p^H, p^1, \dots, p^K)$$

be two vectors of length $T = HK + H + K$. Under the null hypothesis \mathcal{H}_0 the statistic

$$\mathbf{Z} = \sqrt{n}(\hat{\mathbf{p}} - \mathbf{p}^{\mathcal{H}_0}) \tag{15}$$

converges, as n and N go to infinity, to a degenerate multivariate normal distribution with T components having zero mean vector and singular covariance matrix Σ^T .

The proof of Proposition 2 is a consequence of Proposition 1 and is reported in the Appendix.

Proposition 3 *Let f be the continuous, differentiable and with continuous derivatives function defined as follows*

$$f : \mathbf{Z} \rightarrow \chi_H^2 = n \sum_{h=1}^H \sum_{k=1}^K \frac{(\hat{p}^{hk} - \hat{p}^h \cdot \hat{p}^k)^2}{\hat{p}^h \cdot \hat{p}^k}. \tag{16}$$

From Proposition 2, it follows that χ_H^2 tends in distribution to a quadratic form of a degenerate multivariate distribution.

The proof of Proposition 3 is reported in the Appendix.

3.3 Estimation of test statistic limiting distribution under \mathcal{H}_0 via resampling

Resampling methods are computer-based methods for assigning measures of accuracy to statistics estimates and performing statistical inference. In the Bayesian Network learning process, the Efron’s bootstrap Efron (1979) has been proposed as a computationally efficient approach for answering questions about specific network features. In Friedman et al. (1999) the aim is the increase in the learning procedure

performance with information from the bootstrap estimates, even when the amount of data is not enough to induce a high scoring network. Then, the bootstrap method is used to answer questions as: (i) Is the existence of an edge between two nodes warranted? (ii) Is the Markov blanket of a given node robust? (iii) Can we say anything about the ordering of the variables?

Here resampling methods are used to recover the limiting sampling distribution of the test statistic (14) under the independence null hypothesis. With this regard, in sampling from finite populations, the original Efron's bootstrap can lead to biased results since it does not take into account the dependence among units due to sampling design. Adaptations taking into account the non *i.i.d.* nature of the data have been proposed in the literature, which is mainly devoted to estimate variances of estimators; cfr. Mashreghi et al. (2016). The main approaches are essentially two: ad hoc approaches and plug in approaches (cfr. Ranalli and Mecatti (2012), Chauvet (2007) and references therein). The basic idea of ad hoc approaches consists in resampling from the original sample through a special design accounting for the dependence among units. This approach is pursued in McCarthy and Snowden (1985), Rao and Wu (1988), where the re-sampled data produced by the *i.i.d.* bootstrap are properly rescaled, as well as in Sitter (1992), Beaumont and Patak (2012), Chatterjee (2011), Conti and Marella (2015b), where a rescaled bootstrap process based on asymptotic results is proposed. Among the ad hoc approaches we also quote the paper by Antal and Tillé (2011), where a mixed resampling design is proposed. Plug-in approaches are based on the idea of expanding the sample to a pseudo-population that plays the role of a prediction of the original one. Then, bootstrap samples are drawn from such a pseudo-population according to some appropriate resampling design. The most intuitive choice consists in using the same sampling design used to draw the original sample from the population; cfr. Gross (1980), Chao and Lo (1985), Booth et al. (1994), Holmberg (1998), Chauvet (2007), as well as Mashreghi et al. (2016).

Virtually all resampling techniques proposed for finite populations rest on the same justification: in case of linear statistics, the variance of the resampled statistic should be very close to the usual variance estimator, possibly with approximated forms of the second order inclusion probabilities; cfr. Antal and Tillé (2011). This is far from the arguments commonly used to justify the classical bootstrap and its variants, that are based on asymptotic considerations involving the whole sampling distribution of a statistic (cfr., for instance, Bickel and Freedman (1981) and Lahiri (2003)): the asymptotic distribution of a bootstrapped statistic should coincide with that of the "original" statistic. This argument is actually used in Conti and Marella (2015b). In Conti et al. (2019), a class of resampling techniques for finite populations is proposed; cfr. also the paper Jiménez-Gamero et al. (2018), where asymptotic results in Conti et al. (2019) are used to construct various testing procedures based on a design-based estimator of the finite population characteristic function.

However, the class of resampling techniques defined in Conti et al. (2019) does not work in the present case due to the descriptive inference framework. The noticeable exception is the multinomial scheme introduced in Conti et al. (2018) together with the asymptotic results used in Sect. 3.2.1. Such a resampling

procedure has been exploited in Conti et al. (2020) to approximate the asymptotic law of the Lorenz curve estimator when data are collected according to a complex sampling design.

It is based on a two-step procedure consisting in: (i) constructing on the basis of the sampling data a prediction of the population (*i.e.* the multinomial pseudo-population) by sampling N units independently from the original sample where each unit $i \in s$ is selected with probability $\pi_i^{-1} / \sum_{j \in s} \pi_j^{-1}$. Such a prediction is based on the sampling design, and does not essentially involve the superpopulation model; (ii) drawing a sample of the same size of the original one from the pseudo-population according to an appropriate resampling design fulfilling the high entropy requirement. Notice that the idea of using pseudo-populations has been previously used by Mashreghi et al. (2016).

The asymptotic distribution of (14) under the null hypothesis is estimated applying the following procedure.

Step 1 Generate $M = 1000$ bootstrap samples of size n as the original sample size on the basis of the two phase resampling procedure described above. Notice that, the pseudo-population in phase 1 has to be generated under the null hypothesis \mathcal{H}_0 . Specifically, for each unit $i \in s$ such that $Y_i^h = 1$ and $Y_i^k = 1$, the original weight $w_i = 1/\pi_i$ is modified as follows

$$w_i^* = w_i \frac{\hat{p}^h \hat{p}^k}{\hat{p}^{hk}}. \tag{17}$$

Let A^{hk} be the set $\{i \in s : Y_i^h = 1, Y_i^k = 1\}$, the modified weights (17) guarantee that

$$\sum_{i \in A^{hk}} w_i^* = \frac{\sum_{i \in A^h} w_i \sum_{i \in A^k} w_i}{\sum_{i \in s} w_i}. \tag{18}$$

Step 2 For each bootstrap sample, compute the corresponding Hájek estimators $\hat{p}^{hk}, \hat{p}^h, \hat{p}^k$ (6).

Step 3 Compute the M quantities $\chi_H^{2,m}, m = 1, \dots, M$ as in (14).

Step 4 Compute the empirical cumulative distribution function of $\chi_H^{2,m}$ s

$$\hat{T}_{n,M}(t) = \frac{1}{M} \sum_{m=1}^M I_{(\chi_H^{2,m} \leq t)}, \quad t \in \mathbb{R}. \tag{19}$$

Finally, compute the $1 - \alpha$ percentile of $\hat{T}_{n,M}(t)$

$$\widehat{T}_{n,M}^{-1}(1 - \alpha) = \inf\{t : \widehat{T}_{n,M}(t) \geq 1 - \alpha\}, \quad 0 < \alpha < 1. \tag{20}$$

If $\chi_H^2 < \widehat{T}_{n,M}^{-1}(1 - \alpha)$ then \mathcal{H}_0 is not rejected at the $\alpha\%$ significance level.

4 Simulation studies

In this section we proceed to empirically test the PC-cs algorithm performance *via* a simulation study. The study is organized as follows. First of all, in Sect. 4.1 a preliminary analysis is performed. More specifically, the performance of the proposed algorithm when the PC algorithm basic assumptions are violated by the sampling design is investigated. The analysis is based on one sample replicate. Small and larger networks are considered. In Sect. 4.2 a Monte Carlo simulation to evaluate the accuracy of the proposed algorithm is carried out. The aim is to measure how often the PC-cs algorithm is able to eliminate the selection bias recovering the correct dependence structure between the variables. The accuracy is evaluated over a large number of sample replicates. Small and larger networks with fixed or randomly chosen DAG and conditional probabilities are considered.

4.1 Evaluating the accuracy of PC algorithm for complex survey data: a preliminary analysis

In Sect. 4.1.1 the PC-cs algorithm performance when the Markov assumption is violated by the sampling design is investigated. This means that the sample distribution violates the set of conditional independence relations entailed by the DAG to which the population probability distribution P is faithful. The PC-cs algorithm behaviour in case of orientation faithfulness and adjacency faithfulness violations is investigated in sects. 4.1.2 and 4.1.3, respectively. Finally, in Sect. 4.1.4 the robustness of the PC-cs algorithm is evaluated in a ten nodes network where the aforementioned PC algorithm assumptions can be simultaneously violated by the sample distribution.

Fig. 2 a True graph, b Finite population CPDAG

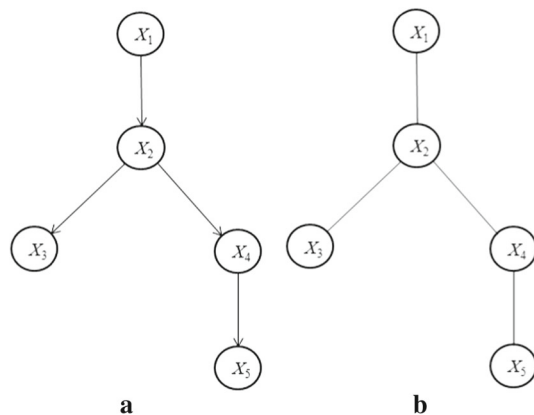


Table 1 Probability distribution of X_1

X_1	$P(X_1 = x_1)$
0	0.25
1	0.35
2	0.40

Table 2 Probability distribution of $X_2|X_1$

X_2	X_1	$P(X_2 = x_2 X_1 = x_1)$
0	0	0.60
1	0	0.40
0	1	0.45
1	1	0.55
0	2	0.25
1	2	0.75

Table 3 Probability distribution of $X_3|X_2$

X_3	X_2	$P(X_3 = x_3 X_2 = x_2)$
0	0	0.30
1	0	0.20
2	0	0.50
0	1	0.55
1	1	0.25
2	1	0.20

Table 4 Probability distribution of $X_4|X_2$

X_4	X_2	$P(X_4 = x_4 X_2 = x_2)$
0	0	0.15
1	0	0.20
2	0	0.35
3	0	0.30
0	1	0.23
1	1	0.50
2	1	0.15
3	1	0.57

4.1.1 Markov assumption violation

A finite population of size $N = 10,000$ has been generated according to the true causal DAG in Fig. 2a. In Tables 1, 2, 3, 4, 5 the conditional probability distributions associated to the nodes are reported.

An estimate of the finite population underlying causal structure has been obtained using the function $pc()$ in the R-package `pcalg`, with the argument

Table 5 Probability distribution of $X_5|X_4$

X_5	X_4	$P(X_5 = x_5 X_4 = x_4)$
0	0	0.60
1	0	0.40
0	1	0.30
1	1	0.70
0	2	0.75
1	2	0.25
0	3	0.50
1	3	0.50

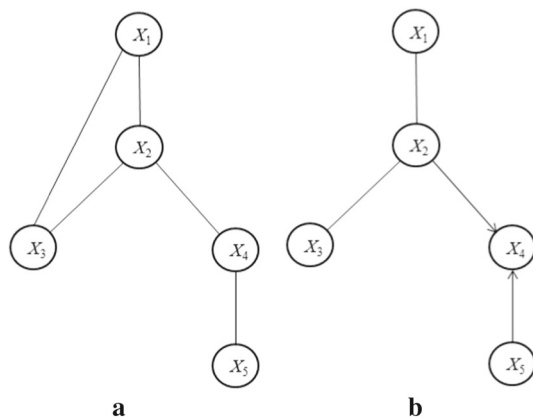
method = “stable”, implementing the PC algorithm, based on conditional independence test for *i.i.d* data, see Kalisch et al. (2012). Notice that, in high dimensional settings the proposal of Lagani et al. (2017) and Tsagris (2019) seems to perform faster and better in terms of returning correct networks. Fig. 2b shows the resulting CPDAG with undirected edges only, representing the Markov equivalence class. We next proceed to generate a sample from our finite population by a complex sampling design. To this aim the variable X_1 has been transformed in a continuous variable Z as follows

$$Z = \begin{cases} N(100, 2) + 15 & \text{if } X_1 = 0 \\ N(10, 2) + 5 & \text{if } X_1 = 1 \text{ or } X_1 = 2 \end{cases} \quad (21)$$

A sample of size $n = 3000$ has been drawn from the finite population according to a conditional Poisson sampling design. Inclusion probabilities are taken proportional to Z -values (21). The effect of the survey design on the causal structure learned using the PC algorithm is shown in the CPDAG in Fig. 3a where an additional edge is placed between the nodes X_1 and X_3 . The conditional independence between X_1 and X_3 given X_2 at the population level is destroyed by the sampling design.

In order to estimate the test statistic distribution under the independence null hypothesis, $M = 1000$ bootstrap replications have been drawn from the selected

Fig. 3 a Markov Condition Violation, b Orientation Faithfulness Violation



sample using the procedure described in sect. 3.3, and the corresponding M bootstrap estimates $\chi_H^{2,m}$, $m = 1, \dots, M$, have been computed. Although the minimum number of bootstrap samples to be considered is 599 as suggested in Wilcox (2010), we use $M=1000$ bootstrap samples to reach a compromise between the computational cost of the procedure and the estimate accuracy.

The significance level for the independence tests has been set equal to 0.05. In this case, the PC-cs algorithm is able to recover the true population equivalence class obtaining the CPDAG shown in Fig. 2b.

The simulation has been repeated considering an alternative sampling design obtained from (21) modifying the parameters of the normal distribution and the intercept value. Specifically, in (21) for $X_1 = 1$ and $X_1 = 2$ the Z variable is generated according to $N(5,1) + 10$ and the effect of sampling design on the association structure is an additional edge between (X_3, X_4) with respect to CPDAG in Fig. 2b. Again the PC-cs algorithm is able to recover the true population equivalence class in Fig. 2b.

4.1.2 Orientation faithfulness violation

In order to investigate the performance of the PC-cs algorithm when the orientation faithfulness is violated, a finite population of size $N = 10,000$ has been generated according to the true causal DAG in Fig. 2a; the conditional probability distributions are in Tables 1, 4, 5 and Tables 6, 7.

A sample of $n = 3000$ has been selected from the finite population according to a conditional Poisson sampling design with inclusion probabilities proportional to Z -values (21). In this case, the survey design produces a failure of the orientation-faithfulness assumption, as shown in Fig. 3b where a v -structure on the triple (X_2, X_4, X_5) is placed. As before, $M = 1000$ bootstrap replications and the corresponding M bootstrap estimates $\chi_H^{2,m}$, $m = 1, \dots, M$, have been computed. Setting the significance level at 0.05, the PC algorithm for complex survey data is able to recover the true population equivalence class obtaining the CPDAG shown in Fig. 2b.

A variation of the PC algorithm, called the conservative PC algorithm, has been proposed in Ramsey et al. (2006) to detect orientation-faithfulness failures. The main difference between the PC algorithm for complex survey data and the conservative PC algorithm can be summarized as follows:

Table 6 Probability distribution of $X_2|X_1$

X_2	X_1	$P(X_2 X_1)$
0	0	0.80
1	0	0.20
0	1	0.45
1	1	0.55
0	2	0.15
1	2	0.85

Table 7 Probability distribution of $X_3|X_2$

X_3	X_2	$P(X_3 X_2)$
0	0	0.7
1	0	0.2
2	0	0.1
0	1	0.2
1	1	0.3
2	1	0.5

1. the PC algorithm for complex survey data takes into account the sampling design *via* a design-based approach, by including the sampling weights in the estimates of the BN parameters. Hence, the PC-cs algorithm adjusts for sample selection bias at the top of the PC algorithm (Phase 1).
2. The conservative PC algorithm assumes the Markov condition and the adjacency faithfulness and tests the orientation faithfulness condition performing additional independence tests. Then the conservative PC algorithm adjusts for sample selection bias at the bottom, *i.e.* in Phase 2, of the PC algorithm. The conservative PC algorithm works in a model-based approach avoiding the use of sampling weights, then it produces bias in the structure learning process if the sampling design is not ignorable, see Pfeffermann (1993).

In our example, the conservative PC algorithm marks the triple (X_2, X_4, X_5) as *ambiguous* since X_4 is in some but not all separating sets. An ambiguous triple is not oriented as a v -structure. Furthermore, no later orientation rule that needs to know whether (X_2, X_4, X_5) is a v -structure or not is applied.

4.1.3 Adjacency faithfulness violation

In order to investigate the performance of PC-cs algorithm when the adjacency faithfulness is violated, a finite population of size $N = 10,000$ has been generated

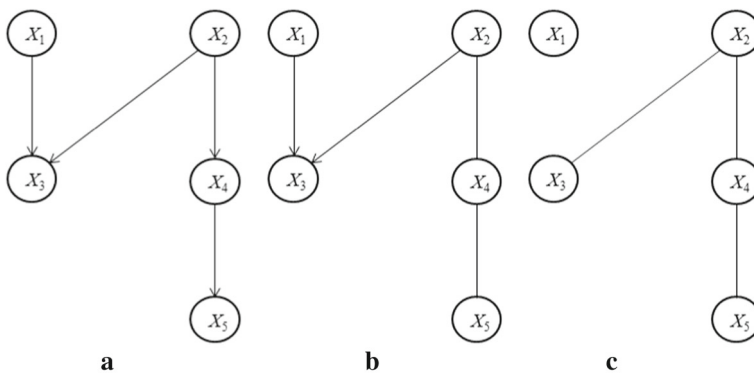


Fig. 4 a Superpopulation graph, b finite population CPDAG, c adjacency faithfulness violation

according to the true causal DAG in Fig. 4a, where a v -structure is introduced. The conditional probability distributions associated to the nodes are in Tables 8, 9, 10, 11, 12.

An estimate of the underlying causal structure in the finite population has been obtained using the function $pc()$ in the R-package `pcalg`. The result is shown in Fig. 4b.

A sample of size $n = 3000$ has been drawn from the finite population according to a conditional Poisson sampling design. Inclusion probabilities are taken proportional to Z -values, defined in (21). The sampling design effect is reported in Fig. 4c where an edge between X_1 and X_3 is missing. On the basis of $M = 1000$ bootstrap replications and a significance level equal to 0.05, the PC algorithm for complex survey data is able to recover the true population equivalence class in Fig. 4b.

4.1.4 Robustness of PC-cs algorithm: a ten nodes network

In this section the performance of the PC-cs algorithm is evaluated in a ten nodes network. Specifically, a finite population of size $N = 10000$ has been generated according to the network in Fig. 5a, where X_3 and X_9 are discrete variables assuming the values (0, 1, 2) while the remaining nodes are dichotomous variables (0, 1).

As before, an estimate of the finite population causal structure has been obtained using the function $pc()$. The result is shown in Fig. 5b. A sample of size $n = 3000$ has been drawn from the finite population according to a conditional Poisson sampling design. Inclusion probabilities are taken proportional to Z -values, defined as follows

$$Z = \begin{cases} N(10, 2) + 5 & \text{if } X_3 + X_9 = 0 \\ N(20, 4) + 10 & \text{if } X_3 + X_9 = 1 \\ N(30, 6) + 15 & \text{if } X_3 + X_9 = 2 \end{cases} \quad (22)$$

and the original PC-algorithm is applied. The sampling design effect on the association structure is reported in Fig. 6a.

Clearly, as the number of nodes increases the Markov condition, the orientation faithfulness, the adjacency faithfulness discussed in sects. 4.1.1–4.1.3 can be simultaneously violated by the sample distribution. Specifically, as Fig. 6a shows,

1. the edge (X_2, X_7) is missing;
2. the edge (X_3, X_9) is added;
3. the directions of the edges (X_2, X_4) and (X_2, X_3) are wrong.

Table 8 Probability distribution of X_1

X_1	$P(X_1)$
0	0.15
1	0.45
2	0.40

Table 9 Probability distribution of X_2

X_2	$P(X_2)$
0	0.5
1	0.5

Table 10 Probability distribution of $X_3|(X_1, X_2)$

X_3	X_1	X_2	$P(X_3 (X_1, X_2))$
0	0	0	0.10
1	0	0	0.50
2	0	0	0.40
0	1	0	0.40
1	1	0	0.20
2	1	0	0.40
0	2	0	0.40
1	2	0	0.30
2	2	0	0.30
0	0	1	0.70
1	0	1	0.20
2	0	1	0.10
0	1	1	0.30
1	1	1	0.50
2	1	1	0.20
0	2	1	0.35
1	2	1	0.25
2	2	1	0.40

Table 11 Probability distribution of $X_4|X_2$

X_4	X_2	$P(X_4 X_2)$
0	0	0.25
1	0	0.25
2	0	0.20
3	0	0.30
0	1	0.23
1	1	0.50
2	1	0.15
3	1	0.12

On the basis of $M = 1000$ bootstrap replications and a significance level equal to 0.05, the PC algorithm for complex survey data is able to recover the population equivalence class in Fig. 6b.

Finally, the structural Hamming distance (SHD, Tsamardinos et al. (2006)) has been computed. Recall that the structural Hamming distance between two CPDAGs is defined as the number of operations (addition, deletion, flips) required to make the

Table 12 Probability distribution of $X_5|X_4$

X_5	$P(X_4)$	$P(X_5 X_4)$
0	0	0.60
1	0	0.40
0	1	0.40
1	1	0.60
0	2	0.55
1	2	0.45
0	3	0.50
1	3	0.50

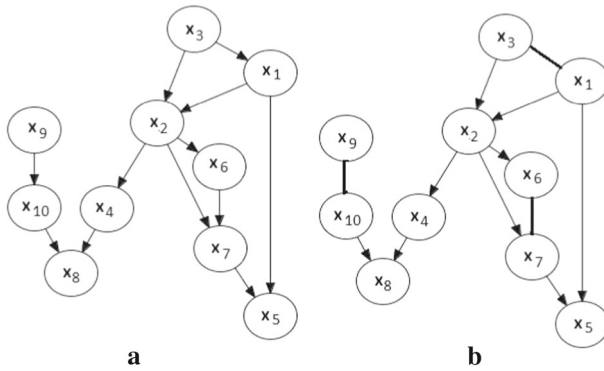


Fig. 5 **a** Superpopulation graph, **b** finite population CPDAG

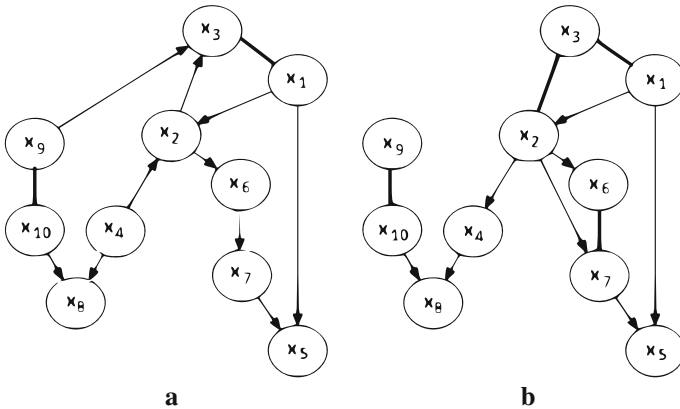


Fig. 6 **a** CPDAG from PC-algorithm, **b** CPDAG from the PC-cs algorithm

CPDAGs match. The *SHDs* between the population CPDAG (Fig. 5b) and the CPDAG obtained by the PC algorithm (Fig. 6a) and the PC-cs algorithm (Fig. 6b) are 7 and 1, respectively, showing the good performance of our proposed methodology in taking into account the sampling design complexity.

4.2 Evaluating the accuracy of PC algorithm for complex survey data: a Monte Carlo study

In this section the performance of the PC-cs algorithm is evaluated over a large number of sample replicates. Small and larger networks with fixed or randomly chosen DAG and conditional probabilities are considered. More specifically, in Sect. 4.2.1 a network of three nodes with fixed DAG and conditional probabilities is considered and the effects of sampling design on the association structure are graphically showed over 500 sample replicates. In Sect. 4.2.2 the ten nodes network of Fig. 5a is analyzed over 350 sample replicates. Finally, in Sect. 4.2.3 the accuracy of the PC-cs algorithm is evaluated when the DAG and the conditional probabilities are randomly chosen. The number of sample replicates is 350 and sample sizes equal to 3000 and 6000 are considered.

4.2.1 A three nodes network with fixed DAG and conditional probabilities

In this section a Monte Carlo simulation is performed to assess the PC-cs algorithm accuracy. A finite population of size $N = 10000$ has been generated according to the network in Fig. 7a. The probability distributions of the nodes X_1 , X_2 and $X_3|X_1, X_2$ are reported in Tables 8, 13 and 14, respectively. An estimate of the finite population causal structure has been obtained using the function $pc()$ in the R-package pcalg. The finite population CPDAG in Fig. 7a has been obtained. In order to investigate the effect of the sampling design on the structural learning process, 500 samples of size $n = 3000$ have been selected from the finite population according to (i) a simple random sampling design; (ii) a conditional Poisson sampling design with inclusion probabilities proportional to the Z -values, defined as follows

$$Z = \begin{cases} N(200, 2) + 10 & \text{if } X_2 = 0 \\ N(10, 2) + 5 & \text{if } X_2 = 1 \end{cases} \quad (23)$$

The significance level is fixed to 0.05. When the sample is selected according to a simple random sampling, the PC algorithm is not able to recover the true association structure in 3% of the selected samples.

The percentage of wrong graphs rises to 10.7% when the sample is selected according to a conditional Poisson sampling. In Fig. 7 the survey design effects on the association structure are shown. The edge between the nodes X_2 and X_3 is

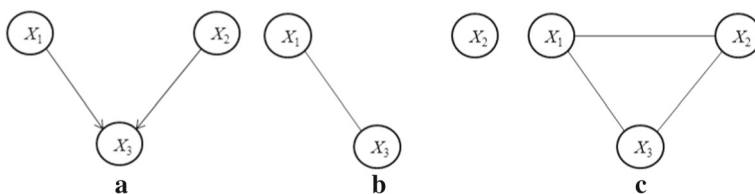


Fig. 7 a True graph and finite population CPDAG, b, c Sampling design effects

Table 13 Probability distribution of X_2

X_2	$P(X_2)$
0	0.7
1	0.3

Table 14 Probability distribution of $X_3|(X_1, X_2)$

X_3	X_1	X_2	$P(X_3 (X_1, X_2))$
0	0	0	0.10
1	0	0	0.50
2	0	0	0.40
0	1	0	0.40
1	1	0	0.20
2	1	0	0.40
0	2	0	0.20
1	2	0	0.20
2	2	0	0.60
0	0	1	0.70
1	0	1	0.20
2	0	1	0.10
0	1	1	0.30
1	1	1	0.50
2	1	1	0.20
0	2	1	0.35
1	2	1	0.25
2	2	1	0.40

missing in 34% of the wrong graphs (Fig. 7b). An additional edge is placed between the nodes X_1 and X_2 in the remaining 66% (Fig. 7c). For each sample, a pseudo population has been constructed and $M = 1000$ bootstrap samples have been drawn. The percentage of wrong graphs decreases to 3.2% when the PC algorithm for complex survey data is applied.

4.2.2 A ten nodes network with fixed DAG and conditional probabilities

With regard to the ten nodes network of Fig. 5a (sect. 4.1.4), in order to assess the PC-cs algorithm accuracy, 350 samples of size $n = 3000$ have been selected from the finite population according to a conditional Poisson sampling design with inclusion probabilities proportional to the Z-values defined in (22). The mean, the standard deviation and the quartiles of SHDs for the PC algorithm and the PC-cs algorithm over the 350 have been computed and denoted by SHD_{iid} and SHD_{cs} , sd_{iid} and sd_{cs} , $Q_{iid}^1, Q_{iid}^2, Q_{iid}^3$ and $Q_{cs}^1, Q_{cs}^2, Q_{cs}^3$ respectively. They are shown in Table 15, where it can be noticed that $SHD_{iid} = 9.3$ while $SHD_{cs} = 3.4$, confirming the good performance of PC-cs algorithm in taking into account the sample selection effects.

Table 15 Mean, standard deviations and quartiles of *SHDs* for PC-algorithm and PC-cs algorithm over the 350 samples

n	SHD_{iid}	SHD_{cs}	sd_{iid}	sd_{cs}	Q_{iid}^1	Q_{iid}^2	Q_{iid}^3	Q_{cs}^1	Q_{cs}^2	Q_{cs}^3
3000	9.3	3.4	1.4	0.3	8.3	9	10.1	3.2	3.3	3.6

4.2.3 A four nodes network with randomly chosen DAG and conditional probabilities

In this section, the accuracy of the PC-cs algorithm is evaluated by a simulation where the DAG and the conditional probabilities are not fixed but randomly chosen. To this aim, the following procedure has been applied:

- Step 1 Generate a random DAG with 4 nodes and sparsity parameter 0.6 by the function *randomDAG* in R package *pcalg*.
- Step 2 Generate a multivariate dataset from the DAG defined in Step 1 by the function *rmvDAG* with nodes corresponding to normal random variables.
- Step 3 Discretize the variables (X_1, X_2, X_3, X_4) according to the DAG dependence structure. First of all, the range of the variables without parents has been divided into 3 intervals according to the h th percentiles of data, for $h = 0.5, 0.75$. Next, if $pa(X_j) \neq \emptyset$, then the range of $X_j|pa(X_j)$ has been divided into 2 intervals according to the 80th percentile of data.
- Step 4 Generate a finite population of size $N = 10000$ from the discretized variables in Step 3 and learn its structure by function *pc()*. This represents the population CPDAG.
- Step 5 Draw a sample of size n from the finite population according to a conditional Poisson sampling design. Inclusion probabilities are taken proportional to Z -values defined as

$$Z = \begin{cases} N(10, 2) + 5 & \text{if } X_1 = 0 \\ N(20, 4) + 10 & \text{if } X_1 = 1 \\ N(30, 6) + 15 & \text{if } X_1 = 2 \end{cases} \quad (24)$$

- Step 6 Perform structural learning using the original PC algorithm and the PC-cs algorithm with a 0.05 significance level.

Steps 1-6 have been repeated 350 times and for $n = 3000, 6000$. Algorithms performance have been compared in terms of (i) the percentage of wrong graphs denoted by W_{iid} and W_{cs} for the PC algorithm and the PC-cs algorithm, respectively; (ii) the mean of *SHDs* denoted by SHD_{iid} and SHD_{cs} , respectively; (iii) the standard deviation of *SHDs* denoted by sd_{iid} and sd_{cs} , respectively; (iv) the first, the second and the third quartile of *SHDs* denoted by $Q_{iid}^1, Q_{iid}^2, Q_{iid}^3$ and $Q_{cs}^1, Q_{cs}^2, Q_{cs}^3$, respectively. Results are reported in Table 16.

Table 16 Percentage of wrong graphs and mean, standard deviations and quartiles of $SHDs$ for PC-algorithm and PC-cs algorithm over the 350 samples, for $n=3000,6000$

n	W_{iid}	W_{cs}	SHD_{iid}	SHD_{cs}	sd_{iid}	sd_{cs}	Q_{iid}^1	Q_{iid}^2	Q_{iid}^3	Q_{cs}^1	Q_{cs}^2	Q_{cs}^3
3000	8.2	3.7	3.4	1.8	0.6	0.3	2.8	3.2	3.7	1.5	1.7	2
6000	6.5	2.8	2.6	1.2	0.4	0.1	2.1	2.4	2.8	1.1	1.2	1.4

As results in Table 16 show, the percentage of wrong graphs W_{cs} , the Structural Hamming Distance SHD_{cs} , the standard deviation and the interquartile difference decrease when the PC-cs algorithm is applied and the complexity of sampling design is taken into account in the learning process. Finally, the accuracy of the results improves as the sample size increases.

5 Conclusions

In this paper a modified version of the PC algorithm is proposed for inferring causal structure from complex survey data. The complexity of sampling design is accounted for via a design-based approach. In the PC algorithm for complex survey data the chi-square statistic does not necessarily approach to a chi-square distribution, and the limiting distribution under the null hypothesis is estimated by a resampling method allowing to make inference on the superpopulation parameters in a finite population setting. Corrections based on design effects have been proposed by Rao & Scott in Rao and Scott (1981) and Rao and Scott (1984). However, while the PC algorithm for complex survey data adjusts for the sample selection bias including the sampling weights in the BN parameters estimates, Rao & Scott corrections use the classical chi-square test statistic adjusted on the basis of the design effects. Furthermore, the second order Rao & Scott correction requires availability of the full covariance matrix estimate of the cell proportions estimators. In secondary analysis this estimate is not necessarily provided, but cell design-effect estimate, possibly with marginal design effect estimate, might be reported. Approximate first-order corrections can then be obtained by using the design effect estimates. These results require that published two-way tables report at least the cells design effects and their marginal along with the cell estimates, otherwise variance estimates must be computed from microdata files using resampling methods for finite population. Hence, both the approaches require to resort to resampling methods: in the PC algorithm for complex survey data the distribution under the independence null hypothesis is estimated; in Rao & Scott approach the variances need to be estimated. Extension of the proposed approach to complex multi-stage designs is under investigation. Finally, future research will be devoted to the development of a structural learning procedure for complex survey sampling in the score-plus-search framework.

Appendix

Proof of Proposition 1 Here the main lines showing how Proposition 1 descends from Proposition 1 in Conti et al. (2018) are provided.

Define the cumulative distribution functions (c.d.f.s),

$$F^{hk} = \sum_{u=1}^h \sum_{v=1}^k p^{uv}, \quad h = 1, \dots, H, \quad k = 1, \dots, K$$

the empirical c.d.f.s

$$\widehat{F}^{hk} = \sum_{u=1}^h \sum_{v=1}^k \widehat{p}^{uv}, \quad h = 1, \dots, H, \quad k = 1, \dots, K$$

where \widehat{p}^{uv} are estimated using the classical Hájek estimators as in (6), and the corresponding random vectors (with elements in lexicographic order)

$$\mathbf{F}^{HK} = \begin{bmatrix} F^{11} \\ F^{12} \\ \dots \\ F^{HK} \end{bmatrix} \quad \widehat{\mathbf{F}}^{HK} = \begin{bmatrix} \widehat{F}^{11} \\ \widehat{F}^{12} \\ \dots \\ \widehat{F}^{HK} \end{bmatrix}$$

and

$$\mathbf{T}^{HK} = \sqrt{n} \left(\widehat{\mathbf{F}}^{HK} - \mathbf{F}^{HK} \right).$$

Note that the random vector \mathbf{T}^{HK} lies on a hyperplane of dimension $HK - 1$, due to the relationships $\widehat{F}^{HK} = F^{HK} = 1$ (then the last component of \mathbf{T}^{HK} is 0).

From Conti et al. (2018) it follows that \mathbf{T}^{HK} tends in distribution, as $n, N \rightarrow \infty$, to a degenerate multivariate Normal r.v. with mean vector $\mathbf{0}^{HK}$ (with HK components) and covariance matrix $\mathbf{\Omega}^{HK}$. Since the limiting distribution is degenerate (it lies in a sub-space of dimension $HK - 1$), the matrix $\mathbf{\Omega}^{HK}$ is degenerate. However, this does not affect neither its definition, nor its basic properties (cfr. Rao (1973), pp. 184-185). In addition, again from Conti et al. (2018), the relationship

$$\mathbf{\Omega}^{HK} = \mathbf{\Omega}_1^{HK} + f\mathbf{\Omega}_2^{HK} \tag{25}$$

holds, where $\mathbf{\Omega}_1^{HK}$ is the part of the total variability due to sampling design, $\mathbf{\Omega}_2^{HK}$ is the part of variability due to superpopulation model, and f is the limiting sampling fraction.

Define now

$$W^{HK} = \sqrt{n} \begin{bmatrix} \hat{p}^{11} - p^{11} \\ \hat{p}^{12} - p^{12} \\ \dots \\ \hat{p}^{HK} - p^{HK} \end{bmatrix}$$

From

$$p^{hk} = F^{hk} - F^{h k-1} - F^{h-1 k} + F^{h-1 k-1}$$

$$\hat{p}^{hk} = \hat{F}^{hk} - \hat{F}^{h k-1} - \hat{F}^{h-1 k} + \hat{F}^{h-1 k-1}$$

where $h = 1, \dots, H, k = 1, \dots, K$, it is immediate to verify that the map

$$T^{HK} \mapsto W^{HK} \tag{26}$$

is linear, and hence continuous. From the continuous mapping theorem, W^{HK} tends in distribution to a degenerate multivariate Normal distribution with mean θ^{HK} and (singular) covariance matrix Σ^{HK} . In view of (25), the matrix Σ^{HK} can be decomposed as

$$\Sigma^{HK} = \Sigma_1^{HK} + f \Sigma_2^{HK}. \tag{27}$$

Next, define

$$W^H = \sqrt{n} \begin{bmatrix} \hat{p}^{1\cdot} - p^{1\cdot} \\ \hat{p}^{2\cdot} - p^{2\cdot} \\ \dots \\ \hat{p}^{H\cdot} - p^{H\cdot} \end{bmatrix} = \sqrt{n} \begin{bmatrix} \sum_{k=1}^K (\hat{p}^{1k} - p^{1k}) \\ \sum_{k=1}^K (\hat{p}^{2k} - p^{2k}) \\ \dots \\ \sum_{k=1}^K (\hat{p}^{Hk} - p^{Hk}) \end{bmatrix}.$$

The map $W^{HK} \mapsto W^H$ is linear, and hence continuous. From the continuous mapping theorem, it follows that W^H tends in distribution to a (degenerate) multivariate Normal distribution, with mean vector θ^H and covariance matrix Σ^H . From (27), it also follows that the following decomposition holds:

$$\Sigma^H = \Sigma_1^H + f \Sigma_2^H.$$

Finally, using exactly the same arguments as above, it is not difficult to see that the degenerate r.v.

$$W^K = \sqrt{n} \begin{bmatrix} \hat{p}^{\cdot 1} - p^{\cdot 1} \\ \hat{p}^{\cdot 2} - p^{\cdot 2} \\ \dots \\ \hat{p}^{\cdot K} - p^{\cdot K} \end{bmatrix}$$

tends in distribution to a (degenerate) multivariate Normal distribution, with mean vector θ^K and covariance matrix Σ^K . Again, the decomposition

$$\Sigma^H = \Sigma_1^H + f\Sigma_2^H$$

holds. □

In order to prove Propositions 2, 3, define the vectors \hat{p}^{HK} and \bar{p}^{HK} of length HK

$$\tilde{p}^{HK} = \begin{bmatrix} \hat{p}^1 \cdot \hat{p}^1 \\ \hat{p}^1 \cdot \hat{p}^2 \\ \dots \\ \hat{p}^1 \cdot \hat{p}^K \\ \hat{p}^2 \cdot \hat{p}^1 \\ \hat{p}^2 \cdot \hat{p}^2 \\ \dots \\ \hat{p}^2 \cdot \hat{p}^K \\ \dots \\ \hat{p}^H \cdot \hat{p}^1 \\ \hat{p}^H \cdot \hat{p}^2 \\ \dots \\ \hat{p}^H \cdot \hat{p}^K \end{bmatrix} \quad \bar{p}^{HK} = \begin{bmatrix} p^1 \cdot p^1 \\ p^1 \cdot p^2 \\ \dots \\ p^1 \cdot p^K \\ p^2 \cdot p^1 \\ p^2 \cdot p^2 \\ \dots \\ p^2 \cdot p^K \\ \dots \\ p^H \cdot p^1 \\ p^H \cdot p^2 \\ \dots \\ p^H \cdot p^K \end{bmatrix}$$

and the matrices ($H \times HK$ and $K \times HK$, respectively)

$$\begin{aligned} \mathbf{A} &= [\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_H] \\ \mathbf{B} &= [\mathbf{B}_1, \mathbf{B}_2, \dots, \mathbf{B}_H] \end{aligned}$$

where

- i) \mathbf{A}_h is a matrix of size $H \times K$ with all entries equal to 0 but the entries of the h th row which are equal to 1, for $h = 1, \dots, H$.
- ii) \mathbf{B}_h is an identity matrix of order K , for $h = 1, \dots, H$.

If we set

$$\hat{p}^{H.} = \begin{bmatrix} \hat{p}^1. \\ \hat{p}^2. \\ \dots \\ \hat{p}^H. \end{bmatrix} \quad \hat{p}^{.K} = \begin{bmatrix} \hat{p}^1. \\ \hat{p}^2. \\ \dots \\ \hat{p}^K. \end{bmatrix}$$

then the relationships

$$\begin{aligned} \hat{p}^{H.} &= \mathbf{A} \hat{p}^{HK} \\ \hat{p}^{.K} &= \mathbf{B} \hat{p}^{HK} \end{aligned}$$

hold. Next, define the matrices ($HK \times H$, $HK \times H$ and $HK \times K$, respectively)

$$\mathbf{\Pi} = \begin{bmatrix} \mathbf{\Pi}_1 \\ \mathbf{\Pi}_2 \\ \dots \\ \mathbf{\Pi}_H \end{bmatrix} \quad \widehat{\mathbf{\Pi}} = \begin{bmatrix} \widehat{\mathbf{\Pi}}_1 \\ \widehat{\mathbf{\Pi}}_2 \\ \dots \\ \widehat{\mathbf{\Pi}}_H \end{bmatrix} \quad \mathbf{\Psi} = \begin{bmatrix} \mathbf{\Psi}_1 \\ \mathbf{\Psi}_2 \\ \dots \\ \mathbf{\Psi}_H \end{bmatrix}$$

where

1. $\mathbf{\Pi}_h$ is a matrix of size $K \times H$ having all entries equal to zero but the entries in the h th column that are equal to p^1, p^2, \dots, p^K , for $h = 1, \dots, H$.
2. $\widehat{\mathbf{\Pi}}_h$ is a matrix of order $K \times H$ having all entries equal to zero but the entries in the h th column that are equal to $\widehat{p}^1, \widehat{p}^2, \dots, \widehat{p}^K$, for $h = 1, \dots, H$.
3. $\mathbf{\Psi}_h$ is a diagonal matrix of order $K \times K$, with all entries in the main diagonal equal to p^h , for $h = 1, \dots, H$.

With this symbols, we may write

$$\sqrt{n} \begin{bmatrix} \widehat{\mathbf{p}}^{HK} - \mathbf{p}^{HK} \\ \widehat{\mathbf{p}}^H - \mathbf{p}^H \\ \widehat{\mathbf{p}}^K - \mathbf{p}^K \end{bmatrix} = \begin{bmatrix} \mathbf{I}^{HK} \\ \mathbf{A} \\ \mathbf{B} \end{bmatrix} \sqrt{n}(\widehat{\mathbf{p}}^{HK} - \mathbf{p}^{HK}) \tag{28}$$

where \mathbf{I}^{HK} is the identity matrix of size $HK \times HK$.

Lemma 1 $\widehat{p}^{hk} - p^{hk}$ converges in probability to 0 as, n, N go to infinity, for each h, k .

Proof Immediate consequence of Proposition 1. □

Note that Proposition 1 actually implies that $\widehat{p}^{hk} - p^{hk} = O_p(n^{-1/2})$, for each h, k .

Lemma 2 $\widehat{p}^h - p^h, \widehat{p}^k - p^k$ converge in probability to 0 as, n, N go to infinity, for each h, k .

Proof Consequence of Lemma 1. □

Proof of Proposition 2 It is enough to use the relationship (28). Proposition 2 follows from (28), Proposition 1, and the continuous mapping theorem. □

Lemma 3 Under the independence hypothesis \mathcal{H}_0 , the limiting distribution of $\sqrt{n}(\widehat{\mathbf{p}}^{HK} - \mathbf{p}^{HK})$ coincides with the limiting distribution of

$$(\mathbf{I}^{HK} - \mathbf{\Pi A} - \mathbf{\Psi B}) \{ \sqrt{n}(\widehat{\mathbf{p}}^{HK} - \mathbf{p}^{HK}) \}$$

that turns out to be (degenerate) multivariate Normal with null mean vector and covariance matrix,

$$\mathbf{\Gamma}^{HK} = (\mathbf{I}^{HK} - \mathbf{\Pi A} - \mathbf{\Psi B}) \mathbf{\Sigma}^{HK} (\mathbf{I}^{HK} - \mathbf{\Pi A} - \mathbf{\Psi B})^T$$

Proof From the relationship

$$\widehat{p}^{hk} - \widehat{p}^h \widehat{p}^k = (\widehat{p}^{hk} - p^h p^k) - \widehat{p}^h (\widehat{p}^k - p^k) - \widehat{p}^k (\widehat{p}^h - p^h)$$

it follows that, in matrix terms,

$$\sqrt{n}(\widehat{\mathbf{p}}^{HK} - \widetilde{\mathbf{p}}^{HK}) = \sqrt{n}(\mathbf{I}^{HK} - \widehat{\mathbf{\Pi}}\mathbf{A} - \mathbf{\Psi}\mathbf{B})(\widehat{\mathbf{p}}^{HK} - \widetilde{\mathbf{p}}^{HK}).$$

Next, from Lemma 2, the matrix $\widehat{\mathbf{\Pi}}$ tends in probability to $\mathbf{\Pi}$, as n, N go to infinity. Using the Slutsky Theorem (Serfling 1980), this implies, in its turns, that the limiting distribution of,

$$\begin{aligned} \sqrt{n}(\widehat{\mathbf{p}}^{HK} - \widetilde{\mathbf{p}}^{HK}) &= (\mathbf{I}^{HK} - \mathbf{\Pi}\mathbf{A} - \mathbf{\Psi}\mathbf{B})\{\sqrt{n}(\widehat{\mathbf{p}}^{HK} - \widetilde{\mathbf{p}}^{HK})\} \\ &\quad - (\widehat{\mathbf{\Pi}} - \mathbf{\Pi})\mathbf{A}\{\sqrt{n}(\widehat{\mathbf{p}}^{HK} - \widetilde{\mathbf{p}}^{HK})\} \end{aligned}$$

coincides with the limiting distribution of

$$(\mathbf{I}^{HK} - \mathbf{\Pi}\mathbf{A} - \mathbf{\Psi}\mathbf{B})\{\sqrt{n}(\widehat{\mathbf{p}}^{HK} - \widetilde{\mathbf{p}}^{HK})\} \tag{29}$$

The linearity of (29) and the continuous mapping theorem complete the proof. \square

For the sake of simplicity, from now the notation

$$\mathbf{C} = \mathbf{I}^{HK} - \mathbf{\Pi}\mathbf{A} - \mathbf{\Psi}\mathbf{B}$$

will be used.

Lemma 4 Define

$$\chi_{2H}^2 = n \sum_{h=1}^H \sum_{k=1}^K (\widehat{p}^{hk} - \widehat{p}^h \widehat{p}^k)^2 \left(\frac{1}{\widehat{p}^h \widehat{p}^k} - \frac{1}{p^h p^k} \right) \tag{30}$$

Under the null hypothesis of independence \mathcal{H}_0 , χ_{2H}^2 converges in probability to 0 as n, N go to infinity.

Proof First of all, we have

$$|\chi_{2H}^2| \leq \max_{h,k} \left| \frac{1}{\widehat{p}^h \widehat{p}^k} - \frac{1}{p^h p^k} \right| \left\{ n \sum_{h=1}^H \sum_{k=1}^K (\widehat{p}^{hk} - \widehat{p}^h \widehat{p}^k)^2 \right\}.$$

Since convergence in probability is preserved under continuous transformations, the term

$$\max_{h,k} \left| \frac{1}{\widehat{p}^h \widehat{p}^k} - \frac{1}{p^h p^k} \right| \xrightarrow{P} 0 \quad \text{as } n, N \rightarrow \infty \tag{31}$$

In addition, from Lemma 3 it follows that

$$n \sum_{h=1}^H \sum_{k=1}^K (\hat{p}^{hk} - \hat{p}^h \hat{p}^k)^2 = \{ \sqrt{n}(\hat{\mathbf{p}}^{HK} - \tilde{\mathbf{p}}^{HK}) \}^T \{ \sqrt{n}(\hat{\mathbf{p}}^{HK} - \tilde{\mathbf{p}}^{HK}) \} \tag{32}$$

$$\xrightarrow{d} \mathbf{X}^T \mathbf{X}$$

where \mathbf{X} is a singular multivariate (HK) Normal r.v. with null mean vector and covariance matrix $\mathbf{\Gamma}^{HK} = \mathbf{C}\mathbf{\Sigma}^{HK}\mathbf{C}^T$. The lemma follows from (31) and (32) and the continuous mapping theorem. \square

Lemma 5 *Define*

$$\chi_{1H}^2 = n \sum_{h=1}^H \sum_{k=1}^k \frac{(\hat{p}^{hk} - \hat{p}^h \hat{p}^k)^2}{\hat{p}^h \hat{p}^k}. \tag{33}$$

Under the null hypothesis of independence \mathcal{H}_0 , χ_{1H}^2 tends in distribution to $\mathbf{X}^T \mathbf{p}^{HK} (\mathbf{p}^{HK})^T \mathbf{X}$ where \mathbf{X} is a (singular) multivariate HK Normal r.v. with null mean vector and covariance matrix $\mathbf{\Gamma}^{HK} = \mathbf{C}\mathbf{\Sigma}^{HK}\mathbf{C}^T$.

Proof It is enough to observe that

$$\chi_{1H}^2 = \{ \sqrt{n}(\hat{\mathbf{p}}^{HK} - \tilde{\mathbf{p}}^{HK}) \}^T \mathbf{p}^{HK} (\mathbf{p}^{HK})^T \{ \sqrt{n}(\hat{\mathbf{p}}^{HK} - \tilde{\mathbf{p}}^{HK}) \}$$

and apply Lemma 3 and the continuous mapping theorem. \square

Proof of Proposition 3 The statistic χ_H^2 can be written as $\chi_{1H}^2 + \chi_{2H}^2$, where χ_{1H}^2 and χ_{2H}^2 are defined in (33) and (30) respectively. The proof is a simple application of Lemma 4, 5. \square

Acknowledgements We want to thank the anonymous referees whose comments considerably improved an earlier version of the paper.

References

Antal E, Tillé Y (2011) A direct bootstrap method for complex sampling designs from a finite population. *J Amer Statist Assoc* 106:534–543

Ballin M, Scanu M (2010) Vicard P (2010) Estimation of contingency tables in complex survey sampling using probabilistic expert systems. *J Stat Plan Inference* 140:1501–1512

Beaumont J-F, Patak Z (2012) On the generalized bootstrap for sample surveys with special attention to poisson sampling. *Int Stat Rev* 80:127–148

Berger YG (2011) Asymptotic consistency under large entropy sampling designs with unequal probabilities. *Pak J Stat* 27:407–426

Bickel PJ, Freedman DA (1981) Some asymptotic theory for the bootstrap. *Ann Statist* 9:1196–1217

Boistard H, Lophuhaä HP, Ruiz-Gazen A (2017) Functional central limit theorems for single-stage sampling design. *Ann Stat* 45:1728–1758

Booth JG, Butler RW, Hall P (1994) Bootstrap methods for finite populations. *J Amer Statist Assoc* 89:1282–1289

Chao MT, Lo S-H (1985) A bootstrap method for finite population. *Sankhya Ser A* 47:399–405

Chauvet G (2007) Méthodes de bootstrap en population finie. Ph.D. Dissertation, Laboratoire de statistique d'enquêtes, CREST-ENSAI, Universioté de Rennes 2,

Chatterjee A (2011) Asymptotic properties of sample quantiles from a finite population. *Ann Inst Statist Math* 63:157–179

- Conti PL (2014) On the estimation of the distribution function of a finite population under high entropy sampling designs, with applications. *Sankhya B* 76:234–259
- Conti PL, Marella D (2015) Inference for quantiles of a finite population: asymptotic vs. resampling results. *Scand J Stat* 42:545–561
- Conti PL, Marella D (2015) Inference for quantiles of a finite population: Asymptotic versus resampling results. *Scand J Stat* 42:545–561
- Conti PL, Marella D, Mecatti F, Andreis F (2019) A unified principled framework for resampling based on pseudo-populations: asymptotic theory. *Bernoulli* 26:1044–1069
- Conti PL, Di Iorio A (2018) Analytic inference in finite populations via resampling, with applications to confidence intervals and testing for independence, [arXiv:1809.08035](https://arxiv.org/abs/1809.08035). Submitted under second review
- Conti PL, Di Iorio A, Guandalini A, Marella D, Vicard P, Vitale V (2020) On the estimation of the Lorenz curve under complex sampling designs. *Stat Meth Appl* 29:1–24
- Cowell RG, Dawid P, Lauritzen SL, Spiegelhalter DJ (2007) Probabilistic networks and expert systems: exact computational methods for bayesian networks, Springer Publishing Company
- Di Zio M, Scanu M, Coppola L, Luzi O, Ponti A (2004) Bayesian networks for imputation. *J Royal Stat Soc A* 167:309–322
- Drton M, Maathuis MH (2017) Structure learning in graphical modeling. *Annu Rev Stat Appl* 4:365–393
- Efron B (1979) Bootstrap methods: another look at the jackknife. *Ann Stat* 7:1–26
- Friedman N, Goldszmidt M, Wyner A (1999) Data analysis with bayesian networks: a bootstrap approach. Proceedings of the 15th annual conference on uncertainty in artificial intelligence, 196–201,
- Gräfström A (2010) Entropy of unequal probability sampling designs. *Stat Methodol* 7:84–97
- Hájek J (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Ann Math Stat* 35:1491–1523
- Holmberg A (1998) A bootstrap approach to probability proportional-to-size sampling. Proceedings of the ASA Section on Survey research Methods, 378–383
- Jiménez-Gamero MD, Moreno-Rebollo JL, Mayor-Gallego JA (2018) On the estimation of the characteristic function in finite populations with applications. *Test* 27:95–121
- Gross ST (1980) Median estimation in sample surveys. In Proceedings of the section on survey research methods. American Statistical Association 181–184
- Kalisch M, Mächler M, Colombo D, Maathuis MH, Bühlmann P (2012) Causal inference using graphical models with the R package pcalg. *J Stat Softw* 47:1–26
- Lagani V, Athineou G, Farcomeni A, Tsagris M, Tsamardinos I (2017) Feature selection with the R package MXM: discovery statistically-equivalent feature subsets. *J Stat Softw* 80:7
- Lahiri SN (2003) Resampling methods for dependent data. Springer series in statistics. Springer, New York
- Mashreghi Z, Haziza D, Leger C (2016) A survey of bootstrap methods in finite population sampling. *Stat Surv* 10:1–52
- Marella D, Vicard P (2013) Object-oriented bayesian networks for modeling the respondent measurement error. *Commun Stat* 42:3463–3477
- Marella D, Vicard P (2015) Object-oriented bayesian network to deal with measurement error in household surveys. *Advances in Statistical Models for Data Analysis*, Springer
- Marella D, Pfeffermann D (2019) Matching information from two independent informative samples. *J Stat Plan Inference* 203:70–81
- McCarthy PJ, Snowden CB (1985) The bootstrap and finite population sampling. In *Vital and health statistics* 95(2): 1–23. Washington, DC: Public Health Service Publication, U.S. Government Printing,
- Pfeffermann D (1993) The role of sampling weights when modeling survey data. *Int Stat Rev* 61:317–337
- Pfeffermann D (2001) Modelling of complex survey data: why model? Why is it a problem? How can we approach it? *Surv Methodol* 37:115–136
- Ramsey J, Spirites P, Zhang J (2006) Adjacency-faithfulness and conservative causal inference, Proceedings of 22nd conference on uncertainty in artificial intelligence, 401–408. Oregon: AUAI Press,
- Ranalli MG, Mecatti F (2012) Comparing recent approaches for bootstrapping sample survey data: a first step towards a unified approach. In Proceedings of the ASA section on survey research methods, 4088–4099,
- Rao CR (1973) *Linear statistical inference and its applications*, 2nd edn. Wiley, New York

- Rao JNK, Scott AJ (1981) The analysis of categorical data from complex sample surveys: chi-squared tests for goodness-of-fit and independence in two-way tables. *J Am Stat Assoc* 76:221–230
- Rao JNK, Scott AJ (1984) On chi-squared tests for multi-way tables with cell proportions estimated from survey data. *Ann Stat* 12:46–60
- Rao JNK, Wu C-FJ (1988) Resampling inference with complex survey data. *J Amer Statist Assoc* 83:231–241
- Serfling RJ (1980) *Approximation theory of mathematical statistics*. Wiley, New York
- Sitter RR (1992) A resampling procedure for complex survey data. *J Amer Statist Assoc* 87:755–765
- Skinner CJ, Holt D, Smith MF (1989) *Analysis of complex surveys*. Wiley
- Spirtes P, Glymour G, Scheines R (2000) *Causation, Prediction, and Search*, MIT Press, Cambridge, MA, 2nd ed. with additional material by D. Heckerman, C. Meek, G. F. Cooper and T. Richardson
- Thibaudeau Y, Winkler WE (2002) Bayesian networks representations, generalized imputation, and synthetic micro-data satisfying analytic constraints, Research Report RRS2002/92002. U.S. Bureau of the Census
- Tsagris M (2019) Bayesian network learning with PC algorithm: an improved and correct variation. *Appl Artif Intell* 33(2):101–123
- Tsamardinos IL, Brown E, Aliferis CF (2006) The max-min climbing Bayesian network structure learning algorithm. *Mach Learn* 65(1):31–78
- Uhler C, Raskutti G, Bühlmann P, Yu B (2013) Geometry of the faithfulness assumption in causal inference. *Ann Stat* 41:436–463
- Verma T, Pearl J (1990) On equivalence of causal models. Technical Report R-150, Department of Computer Science, University of California at Los Angeles
- Wilcox RR (2010) *Fundamentals of modern statistical methods, Substantially improving power and accuracy*. Springer
- Zhang J, Spirtes P (2008) Detection of unfaithfulness and robust causal inference. *Minds Mach* 18:239–271

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.