ROMA TRE
UNIVERSITÀ DEGLI STUDI

COLLANA DEL
DIPARTIMENTO DI ECONOMIA

PC COMPLEX: PC ALGORITHM FOR COMPLEX SURVEY DATA

Daniela Marella - Paola Vicard

ISSN 2279-6916 Working papers
(Dipartimento di Economia Università degli studi Roma Tre) (online)

Working Paper n° 240, 2018

I Working Papers del Dipartimento di Economia svolgono la funzione di divulgare tempestivamente, in forma definitiva o provvisoria, i risultati di ricerche scientifiche originali. La loro pubblicazione è soggetta all'approvazione del Comitato Scientifico.

Per ciascuna pubblicazione vengono soddisfatti gli obblighi previsti dall'art. l del D.L.L. 31.8.1945, n. 660 e successive modifiche.

Copie della presente pubblicazione possono essere richieste alla Redazione.

esemplare fuori commercio
ai sensi della legge 14 aprile 2004 n.106

REDAZIONE:
Dipartimento di Economia
Università degli Studi Roma Tre
Via Silvio D'Amico, 77 - 00145 Roma
Tel. 0039-06-57335655 fax 0039-06-57335771
E-mail: dip_eco@uniroma3.it
http://dipeco.uniroma3.it

ROMA TRE
UNIVERSITÀ DEGLI STUDI

DIPARTIMENTO DI ECONOMIA

PC COMPLEX: PC ALGORITHM FOR COMPLEX SURVEY DATA

Daniela Marella - Paola Vicard

PC complex: PC algorithm for complex survey data

Daniela Marella*, Paola Vicard** [1]

*Dipartimento di Scienze della Formazione,Università "Roma Tre"

**Dipartimento di Economia, Università "Roma Tre"

Abstract

PC algorithm is one of the most known procedures for Bayesian networks structural learning. The structure is inferred carrying out several independence tests on a database and building a Bayesian network in agreement with the tests results. The PC algorithm is based on the assumption of independent and identically distributed observations. In practice, sample selection in surveys involves more complex sampling designs, then the standard test procedure is not valid even asymptotically. In order to avoid misleading results about the true causal structure the sample selection process must be taken into account in the structural learning process. In this paper, a modified version of the PC algorithm is proposed for inferring casual structure from complex survey data. It is based on resampling techniques for finite population. A simulation experiment showing the robustness with respect to departures from the assumptions and the good performance of the proposed algorithm is carried out.

JEL Classification: C100, C120, C180, C830.

Key words: Bayesian network; complex survey data; pseudo-population; structural learning.

1 Introduction

Bayesian networks (BN) are multivariate statistical models satisfying sets of conditional independence statements contained in a directed acyclic graph (DAG), see [8]. The nodes

[1]email: daniela.marella@uniroma3.it, paola.vicard@uniroma3.it
Corresponding author: Daniela Marella - Dipartimento di Scienze della Formazione, Università "Roma Tre" - daniela.marella@uniroma3.it

of the graph correspond to random variables, while edges represent dependencies. Augmented with parameters tables representing marginal and conditional probabilities, BNs are capable of representing the probabilities over any discrete sample space: the probability of any sample point in that space can be computed from the probabilities in the BN.

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 appear to be very useful in missing item imputation ([9], [23]), contingency table estimation for complex survey sampling ([1]) and measurement errors ([15] and [16]). However, there are still some limitations that may complicate their wide application in official statistics contexts. The main one regards the necessity to take into account the sampling design complexity in the structural learning process.

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 [10]. In case of data driven learning, two broad classes of algorithms can be distinguished: score-plus-search and constraint based algorithms.

The main learning algorithm of the constraint based type is the PC algorithm, see [22]. It has several advantages, among which an intuitive basis. The PC algorithm uses conditional independence tests usually performed using the standard Pearson chi-squared test statistic under *i.i.d.* assumption, which is equivalent to simple random sampling assumption. However, sample selection in surveys involves more complex sampling designs based on stratification, different level of clustering and inclusion probabilities proportional to an appropriate size measure. In such circumstances, the standard test procedure is not valid even asymptotically. The impact of complex designs on *i.i.d.* based methods can be severe, as shown in [21].

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 by [19] and [20].

In this paper, a novel approach for inferring casual structure from complex survey data

is investigated. A modified version of the PC algorithm (PC complex) is proposed. The sampling design complexity is accounted for *via* a design-based approach by including the sampling weights in the BN parameters estimates. 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 resorting to resampling techniques for finite population. The new test procedure is applied to BN structural learning modifying the skeleton learning step of the PC algorithm.

The paper is organized as follows. In Section 2 the PC algorithm for *i.i.d.* data and the basic assumptions on which it relies are briefly recalled. In Section 3 the PC complex is introduced and described. A simulation study is performed in Section 4. Finally, advantages and disadvantages of the PC complex are discussed in Section 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 missing arrows between nodes imply conditional independence between the corresponding variables. Examples of DAG are shown in Fig. 1a-b. Consider Fig. 1a. In the arrow $X_2 \to X_3$, X_2 is said parent of X_3 and X_3 is 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 v-structure is a triple of nodes (X_i, X_k, X_j) such that the arrows $X_i \to X_k$ and $X_j \to X_k$ are in the DAG, while X_i and X_j are not adjacent. In such configuration, X_k is said collider. For example, in Fig. 1a the triple (X_2, X_3, X_4) constitutes a v-structure where X_3

is a 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.

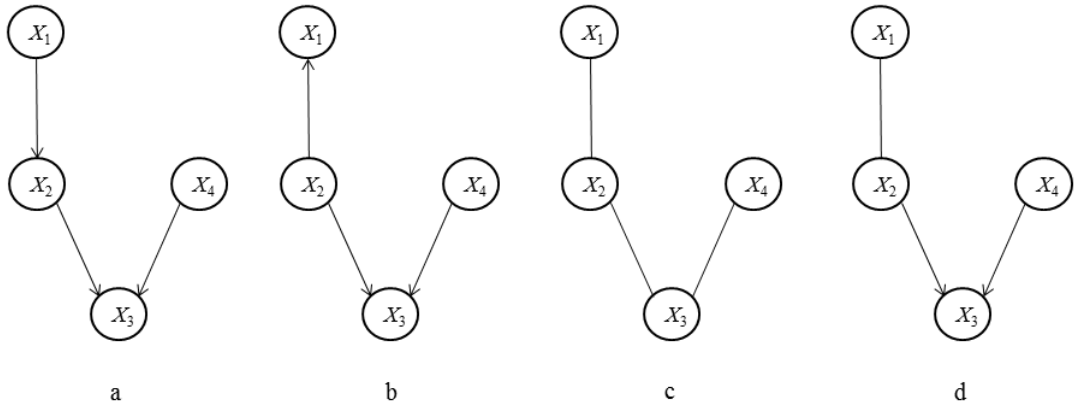

Figure 1: (a) Example of DAG, (b) DAG independence equivalent to (a); (c) skeleton of DAGs in (a) and (b); (d) completed partially DAG, CPDAG for DAGs (a)-(b).

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 [26] and [24].

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 descendents given its parents in G. For example, in Fig.1a X_3 is independent of X_1 given its parent X_2. It 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 [26] a decomposition of the faithfulness 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. However, the leftover part is irrelevant to the correctness of causal discovery procedures such as PC algorithm.

2.3 The original PC algorithm

The PC algorithm starts with a complete undirected graph on the set V and 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.* First, 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 $X_i \to X_k \leftarrow X_j$ 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.

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 is necessary for finding the correct

arrows orientation, then it is related to Phase 2 of the algorithm.

The PC algorithm cannot univocally identify the true DAG, but only the class of independence equivalent DAGs where all members encode the same conditional independence information: two DAGs are independence equivalent if and only if they have the same skeleton and the same v-structures, [25]. For example DAGs a and b in Fig. 1 are equivalent. A common tool for visualizing equivalence classes of DAGs are the completed partially directed acyclic graphs (CPDAG), see [22]. 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 $X_i \to X_j$ and a DAG with $X_i \leftarrow X_j$. An example of CPDAG is shown in Fig. 1d.

3 PC algorithm for complex survey data - PC complex

3.1 The problem

Under the assumptions in section 2.2 and if the sample size is large enough, the original PC algorithm is able to infer the true causal graph from data. This means that, if the input of the PC algorithm is a sample from a population distribution P that is faithful to some DAG, then in large sample limit, the algorithm can identify any probabilistic independence claim with perfect reliability.

In practice 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 casual structure, is far from or close to being unfaithful. 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 could not be faithful to the same DAG because of the sample selection mechanism.

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. 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.

PC complex represents a modified version of the PC algorithm for complex survey data. In particular the skeleton learning step (Phase 1) of the PC algorithm is modified introducing a procedure for testing association in a two-way table for data coming from complex sample surveys. Such a procedure is introduced in Section 3.2 where the existence of 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

Let $\mathcal{U}_N$ be a finite population of size N, labeled by integers $1, \ldots, N$. Denote by A and B the two characters of interest with H $(A^1, \ldots, A^H)$ and K $(B^1, \ldots, B^K)$ categories, respectively. Furthermore, for each unit i, let $Y_i^{h\cdot}$ $(Y_i^{\cdot k})$ be the indicator variable taking value 1 if the unit i assumes the modality A^h (B^k) and 0 otherwise, for $h = 1, \ldots, H$ $(k = 1, \ldots, K)$. Let $Y_i^{hk} = Y_i^{h\cdot} Y_i^{\cdot k}$ so that for each unit i the following equalities hold

$$\sum_{h=1}^{H} Y_i^{h\cdot} = \sum_{k=1}^{K} Y_i^{\cdot k} = 1, \quad \sum_{h=1}^{H} Y_i^{hk} = Y_i^{\cdot k}, \quad \sum_{k=1}^{K} Y_i^{hk} = Y_i^{h\cdot}. \tag{1}$$

For each unit i of the population $\mathcal{U}_N$, let D_i be a Bernoulli random variable, such that i is in the sample whenever $D_i = 1$, whilst i is not in the sample whenever $D_i = 0$. Let $\boldsymbol{D}_N = (D_1, \ldots, D_N)$. An unordered, without replacement sampling design P is the probability distribution of $\boldsymbol{D}_N$. In particular $\pi_i = E_P[D_i]$ is the first order inclusion probability of unit i. The suffix P denotes the sampling design used to select population units. The effective size is the r.v. $n_s = D_1 + \cdots + D_N$. 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 [6] and [7] (assumptions A1-A6). More specifically, here maximal asymptotic entropy sampling designs are considered. The importance of sampling designs high entropy property is discussed in [4], [12] and references therein. Examples of maximal asymptotic entropy sampling design, as shown in [2] and [3], are simple random sampling, successive sampling, Rao-Sampford design, Chao design, stratified design, two-stage design, etc..

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

$$p_N^{hk} = \frac{1}{N}\sum_{i=1}^{N} Y_i^{hk}, \quad p_N^{h\cdot} = \frac{1}{N}\sum_{i=1}^{N} Y_i^{h\cdot} = \sum_{k=1}^{K} p_N^{hk}, \quad p_N^{\cdot k} = \frac{1}{N}\sum_{i=1}^{N} Y_i^{\cdot k} = \sum_{h=1}^{H} p_N^{hk} \tag{2}$$

where $h = 1, \ldots, H$, $k = 1, \ldots, K$. Parameters (2) can be estimated using the classical Hájek estimators

$$\widehat{p}^{hk} = \frac{\sum_{i=1}^{N} \frac{D_i Y_i^{hk}}{\pi_i}}{\sum_{i=1}^{N} \frac{D_i}{\pi_i}}, \quad \widehat{p}^{h\cdot} = \frac{\sum_{i=1}^{N} \frac{D_i Y_i^{h\cdot}}{\pi_i}}{\sum_{i=1}^{N} \frac{D_i}{\pi_i}} = \sum_{k=1}^{K} \widehat{p}^{hk}, \quad \widehat{p}^{\cdot k} = \frac{\sum_{i=1}^{N} \frac{D_i Y_i^{\cdot k}}{\pi_i}}{\sum_{i=1}^{N} \frac{D_i}{\pi_i}} = \sum_{h=1}^{H} \widehat{p}^{hk} \tag{3}$$

for $h = 1, \ldots, H$, $k = 1, \ldots, K$, where $1/\pi_i$ is the sampling weight, that is the reciprocal of the probability that the unit i is included in the sample. If the sampling design is a simple random sampling, the estimators (3) reduce to the proportion of units in the sample belonging to categories (h, k), h and k, respectively.

Next, the existence of the limiting distribution of the Hájek estimators (3), as the sample size and the population size increase is proved. The obtained results are asymptotic and come from [7], where the behaviour of some widely used estimators of the population distribution function for the class of high entropy sampling designs is analyzed. Similarly to [7], we analyze the behaviour of the stochastic processes

$$W_N^{HK} = \{\sqrt{n}(\widehat{p}^{hk} - p_N^{hk}), h = 1, \ldots, H, k = 1, \ldots, K\}, \quad (4)$$

$$W_N^{H} = \{\sqrt{n}(\widehat{p}^{h\cdot} - p_N^{h\cdot}), h = 1, \ldots, H\}, \quad (5)$$

$$W_N^{K} = \{\sqrt{n}(\widehat{p}^{\cdot k} - p_N^{\cdot k}), k = 1, \ldots, K\} \quad (6)$$

as n and N increase. Proposition 1 establishes the convergence of processes $(4)-(6)$ to Gaussian distributions.

Proposition 1. *Under the assumptions A1-A6 of Proposition 1 in [7], as n and N increase, the sequence:*

1. *W_N^{HK} converges in distribution to a degenerate multivariate normal distribution with mean zero and singular covariance matrix Σ_{HK} of order HK;*

2. *W_N^{H} converges in distribution to a degenerate multivariate normal distribution with mean zero and singular covariance matrix Σ_{H} of order H;*

3. *W_N^{K} converges in distribution to a degenerate multivariate normal distribution with mean zero and singular covariance matrix Σ_{K} of order K.*

The proof rests on the same ideas as the proof of Proposition 1 in [7] and it can be seen as a consequence.

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

$$\mathcal{H}_0 : p_N^{hk} = p_N^{h\cdot} p_N^{\cdot k} \quad \text{against} \quad \mathcal{H}_1 : p_N^{hk} \neq p_N^{h\cdot} p_N^{\cdot k}. \quad (7)$$

The used test statistic is

$$\chi^2_H = n \sum_{h=1}^{H} \sum_{k=1}^{K} \frac{(\widehat{p}^{hk} - \widehat{p}^{h\cdot}\widehat{p}^{\cdot k})^2}{\widehat{p}^{h\cdot}\widehat{p}^{\cdot k}} \tag{8}$$

where the sampling weights in the Hájek estimators $\widehat{p}^{hk}$, $\widehat{p}^{h\cdot}$ and $\widehat{p}^{\cdot k}$ (3) compensate for different selection probabilities. In Proposition 3, we show that for complex survey data: (i) the statistic (8) 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*

$$\widehat{\boldsymbol{p}} = (\widehat{p}^{11}, \ldots, \widehat{p}^{HK}, \widehat{p}^{1\cdot}, \ldots, \widehat{p}^{H\cdot}, \widehat{p}^{\cdot 1}, \ldots, \widehat{p}^{\cdot K})$$

and

$$\boldsymbol{p}^{\mathcal{H}_0} = (p_N^{1\cdot} p_N^{\cdot 1}, \ldots, p_N^{H\cdot} p_N^{\cdot K}, p_N^{1\cdot}, \ldots, p_N^{H\cdot}, p_N^{\cdot 1}, \ldots, p_N^{\cdot K})$$

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

$$\boldsymbol{Z}_N = \sqrt{n}(\widehat{\boldsymbol{p}} - \boldsymbol{p}^{\mathcal{H}_0}) \tag{9}$$

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 .

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

$$f : \boldsymbol{Z}_N \to \chi^2_H = n \sum_{h=1}^{H} \sum_{k=1}^{K} \frac{(\widehat{p}^{hk} - \widehat{p}^{h\cdot}\widehat{p}^{\cdot k})^2}{\widehat{p}^{h\cdot}\widehat{p}^{\cdot k}}. \tag{10}$$

From Proposition 2, it follows that χ^2_H tends in distribution to a quadratic form of a degenerate multinormal distribution.

Remark 1. *In general, the χ^2_H limiting distribution does not necessarily approach a chi-square*

distribution as happens for independent and identically distributed observations. However, as stated in [5], if the relationship between the variables of interest and the design variables is weak, then χ^2_H *converges in distribution to* $f(A-1)\chi^2_{(H-1)(K-1)}$ *where*

$$A = lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\pi_i} \tag{11}$$

and $f = n/N$ *. The proportionality factor* $f(A-1)$ *essentially is a finite population correction term, representing the effect of the sampling design on the limiting distribution. The results obtained so far can be easily particularized to the case of simple random sampling design of size* n, i.e. χ^2_H *converges in distribution to* $(1-f)\chi^2_{(H-1)(K-1)}$.

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

The simplest way to estimate the limiting sampling distribution of the test statistic (8) under the independence null hypothesis consists in resorting to resampling methods. In sampling from finite populations, the original bootstrap method proposed by [11] 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 are required.

In [7] a class of resampling techniques for finite populations under complex sampling designs is introduced. It is based on a two-step procedure consisting in: (i) constructing a design-based predictor of the population on the basis of sample data; (ii) drawing a sample from the predicted population according to an appropriate resampling design.

The asymptotic distribution of (8) under the null hypothesis is estimated applying the following procedure.

Step 1 Generate a pseudo-population under the null hypothesis $\mathcal{H}_0$ from the selected sample. Specifically, for each unit $i \in s$ such that $Y_i^{h\cdot} = 1$ and $Y_i^{\cdot k} = 1$, the original weight w_i is

modified as follows

$$w_i^* = w_i \frac{\widehat{p}^{h\cdot}\widehat{p}^{\cdot k}}{\widehat{p}^{hk}}. \tag{12}$$

Let A^{hk} be the set $\{i \in s : Y_i^{h\cdot} = 1, Y_i^{\cdot k} = 1\}$, the modified weights (12) guarantee that

$$\sum_{i \in A^{hk}} w_i^* = \frac{\sum_{i \in A^{h\cdot}} w_i \sum_{i \in A^{\cdot k}} w_i}{\sum_{i \in s} w_i}. \tag{13}$$

A randomization step is introduced to deal with non integer weights w_i^*. Formally, for each unit $i \in \boldsymbol{s}$, let $r_i = w_i^* - \lfloor w_i^* \rfloor$, and consider independent Bernoulli r.v.s ϵ_is with $P(\epsilon_i = 1) = r_i$, then the integer weights are $\lfloor w_i^* \rfloor + \epsilon_i$ as in [13].

Step 2 Generate $M = 1000$ bootstrap samples of size n as the original sample size from the pseudo-population using the original sampling design.

Step 3 For each bootstrap sample, compute the corresponding Hájek estimators $\widehat{p}^{hk}$, $\widehat{p}^{h\cdot}$,$\widehat{p}^{\cdot k}$ (3).

Step 4 Compute the M quantities $\chi_H^{2,m}$, $m = 1, \ldots, M$ as in (8).

Step 5 Compute the empirical cumulative distribution function of $\chi_H^{2,m}$s

$$\widehat{T}_{n,M}(t) = \frac{1}{M} \sum_{m=1}^{M} I_{(\chi_H^{2,m} \leqslant t)}, \quad t \in \mathbb{R} \tag{14}$$

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

$$\widehat{T}_{n,M}^{-1}(1-\alpha) = \inf\{t : \ \widehat{T}_{n,M}(t) \geqslant 1 - \alpha\}, \quad 0 < \alpha < 1. \tag{15}$$

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 Study

4.1 Simulation Plan

In this section we proceed to empirically test the PC complex performance *via* a simulation study. The study is organized as follows. First of all, a preliminary analysis is performed. More specifically, in subsection 4.2 the PC complex 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 complex behaviour in case of orientation faithfulness and adjacency faithfulness violations is investigated in sections 4.3 and 4.4, respectively. Finally, a Monte Carlo simulation to evaluate the accuracy of the proposed algorithm is carried out in section 4.5.

4.2 Markov assumption violation

A finite population of size $N = 10000$ has been generated according to the true causal DAG in Fig. 2a. In Tables 1-5 the nodes conditional probability distributions are reported.

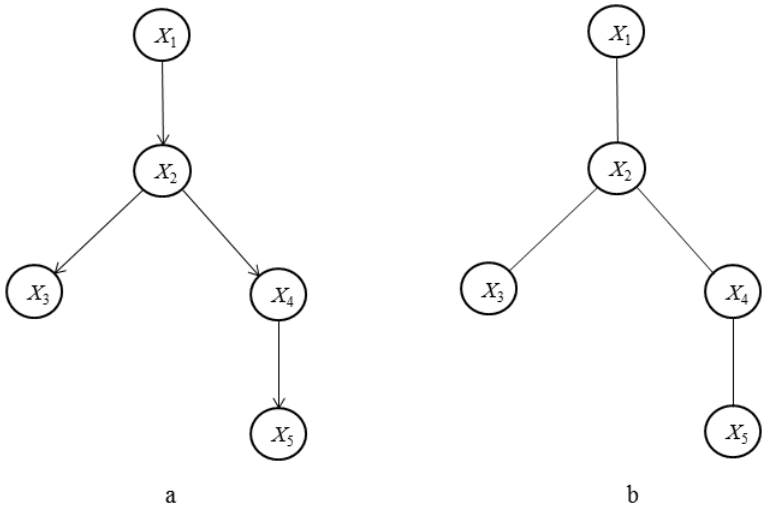

Figure 2: (a) True graph, (b) Finite population CPDAG.

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

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

An estimate of the finite population underlying causal structure has been obtained using the function *pc()* in Package pcalg based on conditional independence test for *i.i.d* data, see [14]. Fig. 2b shows the resulting CPDAG, representing the independence equivalence class, where each edge is undirected.

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

$$Z = \begin{cases} N(100, 2) + 15 & X_1 = 0 \\ N(10, 2) + 5 & X_1 = 1, 2 \end{cases} \tag{16}$$

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 (16). The effect of the survey design on the casual structure is shown in the CPDAG learned using the PC algorithm 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 sample using the procedure described in section 3.3, and the corresponding M bootstrap estimates $\chi^{2,m}_H$, $m = 1, \ldots, M$, have been computed. The significance level for the independence tests has been set equal to 0.05. In this case, the PC complex is able to recover the true population equivalence class obtaining the CPDAG shown in Fig. 2b.

4.3 Orientation faithfulness violation

In order to investigate the PC complex performance when the orientation faithfulness is violated, a finite population of size $N = 10000$ has been generated according to the true causal DAG in

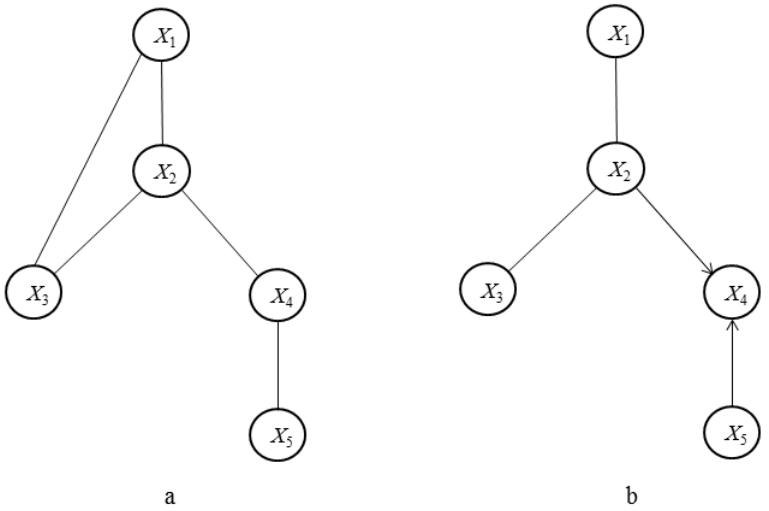

Figure 3: (a) Markov Condition Violation, (b) Orientation Faithfulness Violation.

Fig. 2a; the conditional probabilities distributions are in Tables 1-3 and Tables 6-7.

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

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 (16). 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 $\chi_H^{2,m}$, $m = 1, \ldots, M$, have been computed. Setting the significance level 0.05, the PC complex is able to recover the true population equivalence class obtaining the CPDAG shown in Fig. 2b.

[18] proposed a variation of the PC algorithm, called the conservative PC algorithm, to detect orientation-faithfulness failures. The main difference between the PC complex and the conservative PC algorithms can be summarized as follows:

1. the PC complex 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 complex adjusts for sample selection bias at the top of the PC algorithm.

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 [17].

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.

4.4 Adjacency faithfulness violation

In order to investigate the performance of PC complex when the adjacency faithfulness is violated, a finite population of size $N = 10000$ has been generated according to the true causal DAG in Fig. 4a, where a v-structure is introduced. The nodes conditional probability distributions are in Tables 8-12.

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

An estimate of the underlying causal structure in the finite population has been obtained using the function *pc()* in 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 (16). The sampling design effect is reported in Fig. 4c where and edge between X_1

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

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

and X_3 is missing. On the basis of $M = 1000$ bootstrap replications and a significance level equal to 0.05, the PC complex is able to recover the true population equivalence class in Fig. 4b.

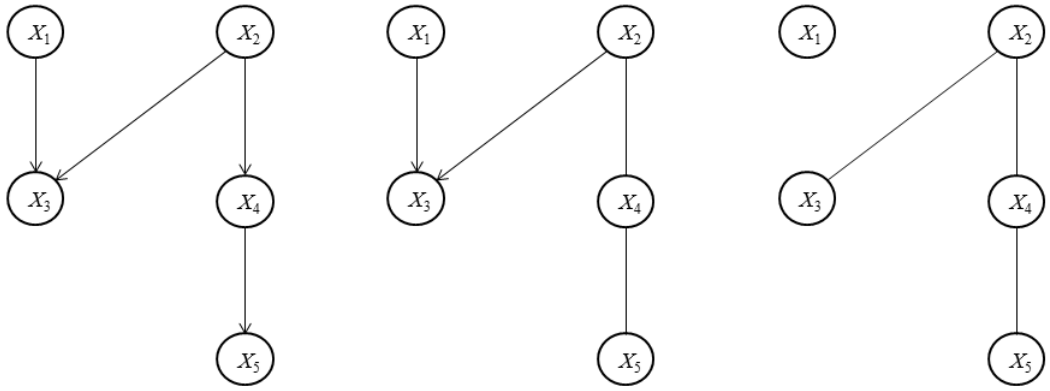

Figure 4: (a) Superpopulation graph, (b) finite population CPDAG, (c) adjacency faithfulness violation.

4.5 Evaluating the accuracy of PC complex

In this section a Monte Carlo simulation is performed to assess the PC complex accuracy. A finite population of size $N = 10000$ has bee generated according to the network in Fig. 5a. The probability distributions of the nodes X_1, X_2 and $X_3|(X_1, X_2)$ are reported in Table 8, Table 13 and Table 14, respectively. An estimate of the finite population causal structure has been obtained using the function *pc()* in Package pcalg. The finite population CPDAG in Fig. 5a 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

$$Z = \begin{cases} N(200, 2) + 10 & X_2 = 0 \\ N(10, 2) + 5 & X_2 = 1 \end{cases} \tag{17}$$

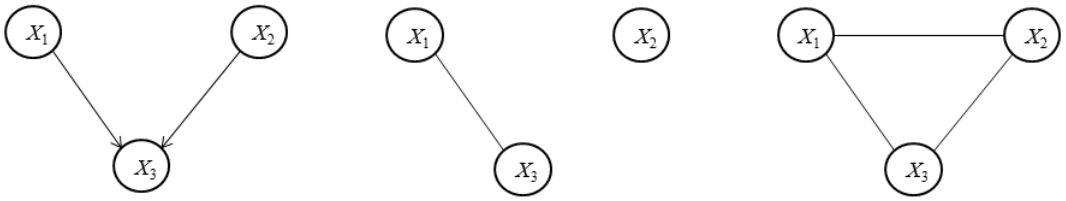

Figure 5: (a) True graph and finite population CPDAG, (b)-(c) Sampling design effects.

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

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. 5 the survey design effects on the association structure are shown. The edge between the nodes X_2 and X_3 is missing in 34% of the wrong graphs (Fig. 5b). An additional edge is placed between the nodes X_1 and X_2 in the remaining 66% (Fig. 5c).

For each sample, a pseudo population has been constructed and $M = 1000$ bootstrap samples have been drawn. The percentage of wrong graphs decreases to 5.2% when the PC complex is applied.

5 Conclusions

In this paper the complexity of sampling design in PC algorithm is accounted for *via* a design-based approach. Corrections based on design effects have been proposed in the literature by [19] and [20]. While PC complex 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 design effects.

In the PC complex the chi-square statistic does not necessarily approach a chi-square distribution and the limiting distribution under the null hypothesis is estimated by resampling methods for finite population. 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 for finite population, although the quantities to be estimated are different. In our approach the distribution under the independence null hypothesis is estimated; in Rao & Scott approach the

variances are estimated.

Acknowledgement

We thank Claudia Tarantola for useful comments and remarks.

References

[1] Ballin, M., Scanu, M. & Vicard, P. (2010). Estimation of contingency tables in complex survey sampling using probabilistic expert systems. *J. Statist. Plann. Infer.*, **140**, 6, 1501–1512.

[2] Berger, Y.G. (1998). Rate of convergence to normal distribution for the Horvitz-Thompson estimator. *Journal of Statistical Planning and Inference*, **67**, 2, 209–226.

[3] Berger, Y.G. (2011). Asymptotic consistency under large entropy sampling designs with unequal probabilities. *Pakistan Journal of Statistics*, **27**, 4, 407–426.

[4] Brewer, K.R.W., & Donadio, M.E. (2003). The high entropy variance of the Horvitz-Thompson estimator. *Survey Methodology*, **29**, 2, 189–196.

[5] Conti, P.L. (2014). On the estimation of the distribution function of a finite population under high entropy sampling designs, with applications. *Sankhya B*, **76**, 2, 234–259.

[6] Conti, P.L. & Marella, D. (2015). Inference for quantiles of a fnite population: asymptotic vs. resampling results. *Scandinavian Journal of Statistics*, **42**, 2, 545–561.

[7] Conti, P.L., Marella, D., Mecatti, F. & Andreis, F. (2017) A unified principled framework for resampling based on pseudo-populations: asymptotic theory. *arXiv:1705.03827*

[8] Cowell, R. G., Dawid, P., Lauritzen, S. L. & Spiegelhalter, D. J. (2007). *Probabilistic Networks and Expert Systems: Exact Computational Methods for Bayesian Networks*. Summit: Springer Publishing Company.

[9] Di Zio, M., Scanu, M., Coppola, L., Luzi, O. & Ponti, A. (2004). Bayesian networks for imputation. *J. Roy. Statist. Soc.* A, **167**, 2, 309–322.

[10] Drton, M. & Maathuis, M. H. (2017). Structure learning in graphical modeling. *Annual Review of Statistics and its Applications*, **4**, 1, 365–393.

[11] Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, **7**, 1, 1–26.

[12] Grafström, A. (2010). Entropy of unequal probability sampling designs. *Statistical Methodology*, **7**, 2, 84–97.

[13] Holmberg, A. (1998). A bootstrap approach to probability proportional-to-size sampling. *Proceedings of the ASA Section on Survey research Methods*, 378–383.

[14] Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H. & Bühlmann, P. (2012). Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, **47**, 11, 1–26.

[15] Marella, D. & Vicard, P. (2013). Object-Oriented Bayesian Networks for Modeling the Respondent Measurement Error. *Communications in Statistics - Theory and Methods*, **42**, 19, 3463–3477.

[16] Marella, D. & Vicard, P. (2015). Object-Oriented Bayesian Network to deal with measurement error in household surveys. In *Advances in Statistical Models for Data Analysis*, 157–164. Springer International Publishing Switzerland.

[17] Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, **61**, 2, 317–337.

[18] Ramsey, J., Spirtes, P. & Zhang J. (2006). Adjacency-faithfulness and conservative causal inference. In *Proceedings of 22nd Conference on Uncertainty in Artificial Intelligence*, 401-408. Oregon: AUAI Press.

[19] Rao, J.N.K. & Scott, A.J. (1981). The analysis of categorical data from complex sample surveys: chi-squared tests for goodness-of-fit and independence in two-way tables. *Journal of the American Statistical Association*, **76**, 374, 221-230.

[20] Rao, J.N.K. & Scott, A.J. (1984). On chi-squared tests for multi-way tables with cell proportions estimated from survey data. *The Annals of Statistics*, **12**, 1, 46-60.

[21] Skinner, C.J., Holt, D. & Smith, M.F. (1989). *Analysis of complex surveys.* Summit: Wiley.

[22] Spirtes, P., Glymour, G. & Scheines, R. (2000). *Causation, Prediction, and Search.* Summit: MIT Press, Cambridge, MA, 2nd ed. with additional material by D. Heckerman, C. Meek, G. F. Cooper and T. Richardson.

[23] Thibaudeau, Y. & Winkler, W.E. (2002). Bayesian networks representations, generalized imputation, and synthetic micro-data satisfying analytic constraints. In: *Research Report RRS2002/92002. U.S. Bureau of the Census.*

[24] Uhler, C., Raskutti, G., Bühlmann, P. & Yu, B. (2013). Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, **41**, 2, 436–463.

[25] Verma, T., Pearl, J. (1990). On Equivalence of Causal Models. *Technical Report, R-150*, Department of Computer Science, University of California at Los Angeles.

[26] Zhang, J., Spirtes, P. (2008). Detection of Unfaithfulness and Robust Causal Inference. *Minds and Machines*, **18**, 2, 239-271.