

Learning Bayesian networks in complex survey sampling

Daniela Marella and Flaminia Musella and Paola Vicard

Abstract Bayesian networks are multivariate statistical models satisfying sets of conditional independence statements. The association structure can be learnt from data by a sequence of independence and conditional independence tests using the PC algorithm. The learning process is based on assumption of independent and identically distributed observations. This assumption is almost never valid for sample surveys data since most of the commonly used survey designs employ stratification and/or cluster sampling and/or unequal selection probabilities. Here, a PC algorithm correction is proposed for taking into account the sampling design complexity.

Key words: complex sample survey, decomposability, PC algorithm.

1 Introduction

In this paper we deal with the problem of learning a Bayesian network (BN) structure when the sample is not drawn according to simple random sampling (*s.r.s.*) but to a complex sampling design. BNs are multivariate statistical models satisfying sets of conditional independence statements contained in a directed acyclic graph (DAG) [3]. A DAG is a pair $G = (V, E)$ where V is the set of nodes (associated with random variables) and E is the set of arrows joining two nodes. Directed cycles are

Daniela Marella

Università Roma Tre, via del Castro Pretorio 20, 00185 Roma, Italy, e-mail: daniela.marella@uniroma3.it

Flaminia Musella

Università Roma Tre, via Silvio D'Amico 77, 00145 Roma, Italy, e-mail: flaminia.musella@uniroma3.it

Paola Vicard

Università Roma Tre, via Silvio D'Amico 77, 00145 Roma, Italy, e-mail: paola.vicard@uniroma3.it

forbidden, *i.e.* it is not possible starting from a node to end up in it following arrow directions. Each node is associated with the conditional probability distribution of the corresponding variable given its parents. The joint probability distribution can be factorised according to the DAG structure.

BNs are a very flexible tool successfully applied in several fields thanks to: their easy readability; the inferential engine enabling to perform *what-if* analysis in a mouse-click time; the modularity property allowing to model increasingly complex problems. Recently BNs have been also applied to official statistics problems, such as missing data imputation [4, 5], and measurement error correction [7]. In all official statistics applications BNs proved to be very promising but, unless the relation structure (the DAG) among the variables of interest is known in advance, BN based imputation and editing can be performed for simple random samples only. Therefore, it is crucial to have structural learning methods available when the sample is not composed of independent and identically distributed (*i.i.d.*) observations from a multivariate random variable. Some initial results developed in a likelihood framework for categorical variables have been proposed in [1]. In this paper we focus on the PC algorithm which is largely used to learn BNs structure for categorical variables, and propose a variation to deal with samples from complex sampling designs. The paper is organized as follows. In Section 2 the PC algorithm for *i.i.d* observations is briefly described. In Section 3 the impact of survey design on conditional independence test for complex survey categorical data and corrections available from existing literature are reviewed. Finally, in Section 4 a variation of PC algorithm for complex survey data is proposed.

2 Learning BNs through PC algorithm

When the background knowledge is not strong enough to build a BN by hand, the association structure (*i.e.* presence/absence of direct edges in the DAG and conditional probability distributions) must be learnt from data. To this aim there are two main approaches [9]: constraint based learning, where the structure is inferred by a sequence of independence and conditional independence tests (for instance the PC algorithm, [13]); score-plus search algorithms that is an optimization based search requiring a scoring function (such as for example the penalized likelihood) and a search strategy [2].

In this paper we focus on the PC algorithm and extend it to deal with data selected according to complex sampling designs. The PC algorithm is a stepwise backward algorithm structured in three main steps [13]:

1. test (conditional) independence and find the skeleton of the graph, *i.e.* the graph without directions. Given a significance level and an ordering of the variables, marginal and conditional independencies given a conditioning set of increasing cardinality, are checked by means of the G^2 statistics that can be expressed in terms of the conditional cross entropy. For example, for three variables X, Y, S

$$G^2 = 2n \sum_s P(S=s) \sum_{x,y} P(X=x, Y=y|S=s) \ln \frac{P(X=x, Y=y|S=s)}{P(X=x|S=s)P(Y=y|S=s)} \quad (1)$$

2. identify colliders, graphically $\rightarrow \circ \leftarrow$, *i.e.* configurations given by two converging arrows outgoing from two unlinked nodes;
3. orient the remaining links without inducing cycles or colliders.

Following [8] where a variation of the PC algorithm for ordinal variables has been provided, an extension of the PC algorithm to non *i.i.d.* samples will be developed by studying a suitable statistics to recursively test conditional independence between categorical variables.

3 Conditional independence test for categorical data in complex sample surveys

Statistical methods for the analysis of categorical data have been extensively developed under the assumption of multinomial sampling (equivalent to those of *i.i.d.* observations and of *s.r.s.*), utilizing loglinear models.

Let s be a sample of size n drawn according to a complex sample design $p(s)$. The cells in a multiway contingency table can be numbered lexicographically as $i = 1, \dots, I$. Let π and $\hat{p}$ denote the I -vectors of cell probabilities π_i and their estimates $\hat{p}_i$, respectively. Generally, $\hat{p}_i$ is a ratio estimator depending on survey design weights. A loglinear model may be written

$$\ln \pi = \tilde{u}(\theta)1 + X\theta \quad (2)$$

where $\ln \pi$ is the I -vector of log probabilities $\ln \pi_i$, θ is a r -vector of parameters, X is a known $I \times r$ matrix of full rank $r \leq I - 1$, $X'1 = 0$, 1 is the I -vector of ones and $\tilde{u}$ is the normalizing factor ensuring that $\sum \pi_i = 1$. If $r = I - 1$ we obtain the saturated loglinear model. Let $\theta = (\theta_1, \theta_2)'$ and $X = (X_1, X_2)$ where X_1 is $I \times s$ and X_2 is $I \times u$ and correspondingly θ_1 is $s \times 1$ and θ_2 is $u \times 1$ and $s + u = r$. We are interested in testing the null hypothesis $H_0 : \theta_2 = 0$.

For multinomial sampling, it is well known that both the standard Pearson chi-squared test statistic X^2 and the likelihood ratio test statistic G^2 are asymptotically distributed as χ_u^2 under the null hypothesis H_0 . The same asymptotic result holds under a product multinomial sampling scheme which arises with stratified simple random sampling when the strata correspond to levels of one dimension of the contingency table. However, this will not remain true with more complex survey designs involving clustering or stratification based on variables different from those in the contingency table.

In real cases, most of the used survey designs involve stratification and clustering; hence the multinomial assumption is violated and the standard procedure is not even asymptotically valid. Under nested loglinear models, the impact of survey design on standard multinomial based methods for a multiway contingency table is studied by

[10], [11], where the asymptotic null distribution of test statistic X^2 and G^2 , for any survey design, is obtained as a weighted sum of independent χ_1^2 random variables. Formally, [10] have shown that, under H_0 , X^2 and G^2 are asymptotically distributed as a weighted sum

$$\delta_1 W_1 + \dots + \delta_u W_u \quad (3)$$

of u independent χ_1^2 variables W_i with weights δ_i related to the design effects (deffs). Formally, δ_i are the eigenvalues of the deffs matrix Δ given by

$$\Delta = (\tilde{X}_2' \Omega \tilde{X}_2)^{-1} (\tilde{X}_2' \Sigma \tilde{X}_2) \quad (4)$$

where $\Omega = n^{-1}[\text{diag}(\pi_i) - \pi\pi']$ is the multinomial covariance matrix under H_0 , $\tilde{X}_2 = [I - X_1(X_1' \Omega X_1)^{-1} X_1' \Omega] X_2$ is the projection of X_2 on the orthogonal complement of the space spanned by the columns of X_1 and Σ is the covariance matrix under the actual complex designs. In case of multinomial sampling $\Sigma = \Omega$ then $\delta_i = 1$, $i = 1, \dots, u$; therefore X^2 and G^2 are asymptotically distributed as a χ_u^2 variable. In complex designs the standard test can provide misleading results since the actual type I error rate will usually differ from the nominal level α , assumed when referring X^2 or G^2 to the α th critical point of χ_u^2 . Hence some adjustments to X^2 or G^2 are needed.

The basic idea in the first-order corrections is to note that for X^2 , the true asymptotic null distribution of $X^2/\delta_{\cdot}$, $\delta_{\cdot} = \sum \delta_i/u$, has the same first moment of χ_u^2 , the limiting distribution of X^2 under the multinomial assumption (a similar argument holds for G^2). A first order correction treats

$$X^2(\hat{\delta}_{\cdot}) = \frac{X^2}{\hat{\delta}_{\cdot}}, \quad G^2(\hat{\delta}_{\cdot}) = \frac{G^2}{\hat{\delta}_{\cdot}} \quad (5)$$

as χ_u^2 , where $\hat{\delta}_{\cdot}$ is a consistent estimator of $\delta_{\cdot}$. Formally

$$\hat{\delta}_{\cdot} = \frac{\sum \hat{\delta}_i}{I - r - 1} = \frac{\text{tr}(\hat{\Delta})}{I - r - 1} \quad (6)$$

where $\hat{\delta}_i$, $i = 1, 2, \dots, I - r - 1$ are the eigenvalues of the estimated deffs matrix $\hat{\Delta}$. H_0 is rejected if the corrected test statistic falls into the standard critical region. The test performs well when the coefficient of variation of the $\hat{\delta}_i$ is small. [11] show that $\hat{\delta}_{\cdot}$ requires the knowledge of the cell and the marginal deffs only provided the model admits explicit solution to likelihood equation under multinomial scheme.

The corrections (5) are useful if the researcher must perform secondary analyses from multiway tables reporting the deffs. In such situations only variance (or equivalently deffs) estimates for individual cell estimates (diagonal element of $\hat{\Sigma}$) and for specific marginals (marginal deff estimate) might be reported by which first-order correction can be computed. If the whole covariance matrix estimate of the cell proportions estimators $\hat{\Sigma}$ is available second order corrections, taking into account the variability of $\hat{\delta}_i$, can be implemented, see [10] and [11].

Furthermore, some adjustments to X^2 and G^2 when the models do not admit direct solutions to multinomial likelihood equations are provided in [12].

4 PC algorithm in complex sample surveys

The standard PC algorithm is based on assumption of independent and identically distributed observations. As previously stressed, this assumption is almost never valid for sample surveys data since most of the commonly used survey designs employ stratification and/or cluster sampling and/or unequal selection probabilities.

If the design variables, which may include label information such as cluster or stratum indicators variables, quantitative variables such as measures of size, are known then the sampling design can be ignored and the standard PC algorithm can be used. That is, if all the design variables representing the relevant features of the sampling design are incorporated into the learning process then the design is ignorable. Generally, in secondary analysis the values of the design variables are not included in the data then the design may be not ignorable and it must be taken into account in the learning process. Moreover, in most cases, only the deffs of the cells and of specific marginals are provided. In such situation the first-order Rao Scott corrections can be computed exclusively for those loglinear models admitting an explicit solution to the likelihood equations. This means that attention must be focused on a subset of hierarchical models, the decomposable loglinear models. In particular, here we exploit G^2 test statistics in (1) (and its corrections for survey design) to test conditional independence assumptions.

Conditional independence relations in log-linear models can be alternatively and equivalently described by means of undirected graphs where nodes still represent the variables and undirected edges represent the two-factor terms interactions included in the model. In particular there is a direct correspondence between decomposable loglinear graphical models and decomposable loglinear models. In graphical terms, decomposability can be checked by looking at the graph structure: a graph is decomposable if it is triangulated, *i.e.* if it does not contain cycles of four or more nodes without a chord (an edge between two not consecutive nodes). A decomposable graph can be recursively decomposed in cliques, *i.e.* maximal complete subsets of nodes, and the set of cliques constitutes, in turn, the generating class of the hierarchical loglinear model.

As illustrated in Section 2, the PC algorithm first builds the undirected graph and then assigns suitable directions to the existing edges. By the definition of decomposable graphs as triangulated ones, it follows that when an undirected decomposable graph is given, collider configurations cannot occur in the direction assignment phase.

Therefore, by all the considerations above, the PC algorithm correction is still a stepwise backward algorithm but it is now structured in two steps only as follows.

1. Test (conditional) independence by comparing the loglinear model representing such an independence statement with the saturated model in a recursive way until

the skeleton of the graph is found. At each step the candidate link to be removed is selected in way that the resulting graph is still decomposable, *i.e.* the candidate link belongs to one clique only [6]. Notice that, due to tested model decomposability, local computations can be performed, *i.e.* it is possible to work on one single clique at a time (specifically, on the clique containing the edge proposed for the deletion). Consequently, there is a gain in terms of computational efficiency.

2. Orient the remaining links without inducing cycles and in the respect of directional constraints, eventually fixed by the user.

The goodness of the adjusted procedure is investigated through simulation studies. The methodology here proposed helps accounting for survey designs different from *s.r.s.* when learning a BN. Nevertheless model search is limited to decomposable graphs. This is quite a substantial restriction, although it is often introduced in much of graphical model research. In future work it will be necessary to work on an extension of this algorithm to non decomposable graphs.

References

1. Ballin, M., Scanu, M., Vicard, P.: Estimation of contingency tables in complex survey sampling using probabilistic expert systems. *Journal of Statistical Planning and Inference*, **140**, 1501-1512 (2010).
2. Cooper, G. F. and Herskovits, E.: A bayesian method for the induction of probabilistic networks from data. *Mach. Learn.*, **9**, 4, 309-347 (1992)
3. Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J.: *Probabilistic Networks and Expert Systems: Exact Computational Methods for Bayesian Networks*. Springer Publishing Company, Incorporated, 1st edition. (2007)
4. Di Zio, M., Scanu, M., Coppola, L., Luzi, O., Ponti, A.: Bayesian Networks for Imputation. *J. Roy. Statist. Soc. A* **167**, 2, 309-322 (2004)
5. Di Zio, M., Sacco, G., Scanu, M., Vicard, P.: Multivariate techniques for imputation based on Bayesian networks. *Neural Network World* **4**, 303-309 (2005)
6. Frydenberg, M., Lauritzen, S. L.: Decomposition of maximum likelihood in mixed interaction model. *Biometrika* **76**, 539-555 (1989)
7. Marella, D., Vicard P.: Object-Oriented Bayesian Network for modeling the respondent measurement error. *Communications in Statistics-Theory and Methods*, **42**:19, 3463-3477 (2013)
8. Musella F.: A PC algorithm variation for ordinal variables, *Computational Statistics*, **28**, 6, 2749-2759 (2013)
9. Neapolitan, R.E.: *Learning Bayesian Networks*. Prentice Hall, Upper Saddle River (2004)
10. Rao J.N.K. and Scott A.J.: The analysis of categorical data from complex sample surveys: Chi-squared Tests, *JASA*, **76**, 221-230 (1981)
11. Rao J.N.K. and Scott A.J.: On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data, *The Annals of Statistics*, **12**, 1, 46-60 (1984)
12. Rao J.N.K. and Scott A.J.: On simple adjustments to chi-square tests with sample survey data, *The Annals of Statistics*, **15**, 1, 385-397 (1987)
13. Spirtes, P., Glymour, C., and Scheines, R.: *Causation, Prediction, and Search*. The MIT Press, Cambridge, Massachusetts, 2 edition. (2000)