

Prior beliefs about the structure of a probabilistic network¹

Federico M. Stefanini ²

Dipartimento di Statistica ‘G.Parenti’,
University of Florence, e-mail: stefanini@ds.unifi.it

Abstract: The exploitation of prior information is often advocated to face the curse of dimensionality which affects multivariate statistical models in biomedical research and bioinformatics. In a recent contribution dealing with score-and-search learning of Bayesian Networks, the elicitation of beliefs in large spaces of structures has been considered in a general setup based on elicited network features. In this contribution we build on the above cited work and provide a simpler formalization dealing with incomplete prior information in the form of DAG of network features. A case study illustrates the proposed approach.

Keywords: Bayesian networks, elicitation, structural learning

1. Introduction

The elicitation of prior beliefs about the structure of a probabilistic network is an ingredient of full-Bayesian structural learning but an element-by-element assignment of subjective probability values to all graphs on a fixed set of nodes requires a substantial amount of prior information often unavailable. Moreover, from five nodes on, the enumeration of structures becomes intractable due to a combinatorial explosion in the number of DAGs. In Stefanini (2008), the elicitation of beliefs in large spaces of structures is performed by defining propositions which describe plausible network features. Coherence and anchoring are formally considered in view of the possible need of revising elicited quantities. A fair degree of generality is achieved in Stefanini (2008), for example by avoiding constraints like the a-priori independence among parent sets and the specification of a prior network. A limitation is due to the elicitation of conjunctions of elementary propositions, the canonical features, if conditional independent relationships among reference features are a priori plausible: in this case a conditional approach to elicitation might be preferred. Moreover only elementary propositions which may be true or false for a given structure are considered, so that some notational and computational inefficiency follows when incompatible statements are included in the reference set. Finally the need of dedicated software limits the widespread testing of the approach.

¹This research is partially funded by TUSCANIA s.r.l., Italy.

²Address of correspondence: viale Morgagni 59, 50134 Firenze

2. Methods

A graph $\mathcal{G}$ is a pair (V, E) where $V = \{v_1, v_2, \dots, v_K\}$ is a finite set of nodes and $E \subset V \times V$ is the set of edges. Here we consider Directed Acyclic Graphs (DAGs) and probabilistic networks as defined in Cowell *et al.* (1999), where graph-theoretic details and Markov properties are provided. A one-to-one map $E \longleftrightarrow z$ from the set of DAGs on a fixed V to a subset Ω_Z of natural numbers is defined for ease of notation.

Elements of V denote the components of the random vector X_V , that is $X_{v_1}, \dots, X_{v_K}$ are random variables. DAGs are suited to represent (strong) conditional independence relationships and causal relationships existing among elements of X_V . The probability distribution of X_V is Markov with respect to a DAG z if the following factorization holds: $p(x_{v_1}, x_{v_2}, \dots, x_{v_K}) = \prod_{v_i \in V} p(x_{v_i} \mid x_{pa(v_i)})$, where $x_{pa(v_i)}$ is the subvector made by variables whose labels belong to the parents set of node v_i . More general conditional independence statements may be derived by means of the separation theorem for DAGs or by d-separation. If causal relationships are encoded into a DAG then an arrow $v_i \rightarrow v_j$ represents the immediate causal relation of X_{v_i} on X_{v_j} . Stronger assumptions are needed for causal encoding, for example that the graph is stable under external manipulation (intervention) on a subset of nodes. Full-Bayesian structural learning points towards the posterior distribution of Z given the database $\mathcal{D} = \{d_1, d_2, \dots, d_{n_d}\}$ of n_d conditionally independent realizations of the random vector X_V . The posterior distribution on Ω_Z is typically intractable, therefore a search in the space Ω_Z is performed by optimizing the marginal distribution $p(\mathcal{D}, z \mid \xi) = \int p(\mathcal{D} \mid \theta, z, \xi) \cdot p(\theta \mid z, \xi) \cdot p(z \mid \xi) \cdot d\theta$, where θ is the vector of parameters embed into the conditional probability distributions of the Markovian factorization.

Following Stefanini (2008), we assume that relevant network features may be expressed by propositions $\mathcal{P}_i$ which belong to a reference set $\mathcal{R} = \{\mathcal{P}_i : i = 1, \dots, n_f\}$, with V fixed. Note that an atomic proposition $\mathcal{P}_i$ applied to a candidate structure z may either be true or false. Here we introduce parameterized propositions that depend on a parameter defined on suitable set, so that a collection of atomic and incompatible propositions are simultaneously defined, for example $\mathcal{P}_i(a) = \text{'the cardinality of } pa(v_i) \text{ is } a\text{'}$ where a is value in $\mathcal{A}$, a subset of natural numbers. Parameterized propositions are useful to state plausible values of counts regarding a candidate structure. The notation is simplified by using indicator variables $R_i : \{\neg\mathcal{P}_i, \mathcal{P}_i\} \rightarrow \{0, 1\}$, with $\mathcal{P}_i \mapsto 1$ and $\neg\mathcal{P}_i \mapsto 0$ if atomic propositions are considered. Otherwise, a variable R_i is introduced for each parameterized proposition $\mathcal{P}_i(a)$, with sample space $\mathcal{A}_i$. Vector $R = (R_1, R_2, \dots, R_{n_f})$ takes values $r_j = (r_{1,j}, r_{2,j}, \dots, r_{n_f,j}) \in \Omega_R$ that correspond to conjunctions of propositions (canonical features) so that the degree of belief about a joint statement of all reference features may be elicited. Such belief about r_j corresponds to a statement specifying that the unknown structure has structural features as coded by the configuration $(r_{1,j}, r_{2,j}, \dots, r_{n_f,j})$. The reference set $\mathcal{R}$ is a basis for the elicitation because it induces a partition $\mathcal{Z} = \{\mathcal{Z}_j : j \in \mathcal{J}\}$ on the set of structures for a fixed set V , where $\mathcal{Z}_j = \{z : z \mapsto r_j\}$ is the subset of structures satisfying the same joint assertion obtained by conjunction of reference features, thus sharing the same value of prior degree of belief. The prior distribution $p(z \mid \xi)$ is elicited by extension on vector R , that is

$$p(z \mid \xi) = \sum_{r_j \in \Omega_R} P[Z = z \mid R = r_j, \xi] \cdot P[R = r_j \mid \xi].$$

The probability $P[Z = z \mid R = r_j, \xi]$ is null for all but one configuration or R , which is the value $r_{j(z)}$ representing the structural features of candidate z . Moreover, the working

definition of prior distribution is derived by indifference of beliefs among members within an equivalence class (Stefanini, 2008, modified): $p(z \mid \xi) = \frac{1}{n_{j(z)}} \cdot P[R = r_{j(z)} \mid \xi]$. This equation lies upon the ability of counting the size of equivalence classes. In Stefanini (2008), a Monte Carlo algorithm is proposed to estimate cardinalities $n_{j(z)}$. Moreover the second factor $P[R = r_{j(z)} \mid \xi]$ entails a prior distribution $p(r_1, r_2, \dots, r_{n_f} \mid \xi)$ on R . Here we propose to represent prior beliefs on network features belonging to $\mathcal{R}$ by means of a Bayesian Network for R . We follow the rules below:

1. Select a preferred order of elicitation $(R_{i_1}, R_{i_2}, R_{i_3}, \dots, R_{i_{n_f}})$.
2. Check out that at step $k = 1, 2, \dots, n_f$ the elicitation of R_{i_k} is feasible given the configurations of subvector $(R_{i_1}, R_{i_2}, \dots, R_{i_{k-1}})$, or just the elements in the parent set of node i_k if the DAG of features is not complete.
3. Elicit the conditional distribution of R_{i_k} given $R_{pa(i_k)}$ for each configuration of the conditioning subvector. A directed method based on roulette wheel or a standard urn containing two kinds of balls is typically employed. The question at each step is: ‘Knowing that a structure shows features as coded by the configuration of $r_{i_1}, r_{i_2}, \dots, r_{i_{k-1}}$, how much plausible is feature r_{i_k} ?’
4. Check out marginal probabilities, full margins of vector R first: if the expert does not confirm the implied beliefs then the elicited values must be revised.
5. If several candidate orders exist in step (1.) of the elicitation then the ‘overelicitation’ may be exploited to check for the presence of elicitation bias due to anchoring on a selected order. The comparison of elicited distributions given different orders should not reveal differences in probability values otherwise reconciliation of elicited distributions is needed.

3. Results and discussion

Here the case study presented in Stefanini (2008) is reworked. Classical biomarkers in breast cancer studies include progesterone receptors (PR), oestrogen receptors (ER), age (AG), menopausal status (MS), number of positive lymph nodes (PL). Variables of interest for patients are tumor grade (TG) and tumor size (TS). The expert defines the reference set $\mathcal{R}$ as listed below: $\mathcal{P}_1 = \text{‘Nodes AG, MS are ancestors of all other nodes’}$; $\mathcal{P}_2 = \text{‘TS, TG, NL are descent of all other nodes’}$; $\mathcal{P}_3 = \text{‘} |pa(v_i)| \leq 3 \text{ for each node } v_i \in V\text{’}$; $\mathcal{P}_4 = \text{‘ER is independent on AG given MS’}$. The total number of DAGs for a set V of seven nodes is 1’138’779’265. The elicited order relations is (R_1, R_2, R_3, R_4) and a complete DAG of features is employed. Key entries of the first three features are: $P[R_1 = 1] = \frac{7}{12}$, $P[R_2 = 1 \mid R_1 = 0] = \frac{3}{5}$, $P[R_2 = 1 \mid R_1 = 1] = \frac{4}{7}$, $P[R_3 = 1 \mid R_1 = 0, R_2 = 0] = \frac{5}{8}$, $P[R_3 = 1 \mid R_1 = 0, R_2 = 1] = \frac{7}{12}$, $P[R_3 = 1 \mid R_1 = 1, R_2 = 0] = \frac{7}{12}$, $P[R_3 = 1 \mid R_1 = 1, R_2 = 1] = \frac{9}{16}$. The DAG of features is useful to compute the joint probability value for each $r_j \in \Omega_R$, and to calculate the marginal probability distributions representing implied beliefs for margins of R (Table 1). The expert did confirm implied beliefs therefore the revision of the elicited distribution did not take place. In Table (1), results of replicated simulations (100 replicates) performed to estimate the cardinality values of elements in $\mathcal{Z}$ are shown together with the estimated standard deviation of the estimates.

The DAG of features described above and the elicited conditional probability tables are available for download at the author’s website as text and in gRain format (a package for the R environment <http://www.r-project.org>).

Table 1: *Estimated fraction of DAGs in each equivalence class (second left). Mean and standard deviation of 100 replicated simulations are shown on the right. Prior probability values for each configuration of R are on the first column right. Note that $f(7)$ is the number of DAGs on a set V of seven nodes.*

$R = r_j$	Estimate $n_j/f(7)$	Mean	Standard deviation	Prior belief
(0, 0, 0, 0)	0.5289	0.528400	0.004797	0.0208
(0, 0, 0, 1)	0.0171	0.020090	0.001437	0.0417
(0, 0, 1, 0)	0.3181	0.317800	0.004741	0.0417
(0, 0, 1, 1)	0.0184	0.019260	0.001533	0.0625
(0, 1, 0, 0)	0.0191	0.017510	0.001206	0.0417
(0, 1, 0, 1)	0.0032	0.003116	0.000517	0.0625
(0, 1, 1, 0)	0.0131	0.012900	0.001067	0.0625
(0, 1, 1, 1)	0.0013	0.002176	0.000475	0.0833
(1, 0, 0, 1)	0.0345	0.032870	0.001839	0.0417
(1, 0, 1, 0)	0.0075	0.008087	0.000893	0.0625
(1, 0, 1, 1)	0.0182	0.019200	0.001428	0.0625
(1, 1, 0, 0)	0.0066	0.005697	0.000710	0.0833
(1, 1, 0, 1)	0.0056	0.004484	0.000709	0.0625
(1, 1, 1, 0)	0.0034	0.003010	0.000594	0.0833
(1, 1, 1, 1)	0.0029	0.003126	0.000572	0.0833
(1, 1, 1, 1)	0.0021	0.002298	0.000505	0.1040

The elicitation based on a DAG of features is suited to problem domains in which an order relation among features may be elicited and the relations among features are sparse, so that marginal independence and complete dependence of features are just two extreme instances. The elicitation scheme proposed in this work complements the approach in Stefanini (2008) which is focused on canonical features and is particularly useful if the order relation on features exploited in this work is not easy to elicit. An approach accounting for a partial order on the reference set of features is feasible and it will be considered elsewhere.

References

- Cowell R.G., Dawid A.P., Lauritzen S.L. and Spiegelhalter D.J. (1999) *Probabilistic Networks and expert systems*, Springer-Verlag.
- Stefanini F.M. (2008) Using weak prior information on structures to learn bayesian networks, in: *PGM2008, Congress Acta of the Fourth European Workshop on Probabilistic Graphical Models*, Hirtshals, Denmark, http://pgm08.cs.aau.dk/Papers/32_Paper.pdf.