Learning Probabilistic Graphical Models

Péter Antal

Department of Measurement and Information Systems

Intelligent Data Analysis, 2017

Topics

- Assumptions:
  - Stability
  - Causal Markov Assumption
  - Priors
  - Interventional data
- Asymptotic learning
  - Hardness of learning: NP
  - IC
  - local Causal Discovery
  - IC*
- Score-based methods
  - An information theoretic approach
  - Score equivalence
  - Learning from interventional data
- Parameter learning
  - From inference to learning
  - Complete data: ML/MAP
  - Incomplete data
- Bayesian structure learning
  - From observational equivalencee to Dirichlet priors
  - Bayesian parameter learning
  - BD scores

Stable distributions

Definition

The distribution P is said to stable (or faithfull), if there exists a DAG called perfect map exactly representing its (in)dependencies (i.e. $(X \perp\!\!\!\perp Y|Z)_G \Leftrightarrow (X \perp\!\!\!\perp Y|Z)_P \; \forall X, Y, Z \subseteq V$). The distribution P is stable w.r.t. a DAG G, if G perfectly represents its (in)dependencies.

However, there can be numerically encoded independencies corresponding to solutions of equation systems and/or to functional dependencies, but they are rare and not stable for numerical perturbations.

1. Consider $p(X, Y, Z)$ with binary X, Z and ternary Y. The conditionals $p(Y|X)$ and $p(Z|Y)$ can be selected such that $p(z|x) = p(z|\neg x)$. That is $(X \not\perp\!\!\!\perp Y)$ and $(Y \not\perp\!\!\!\perp Z)$, but $(X \perp\!\!\!\perp Z)$, demonstrating that the "naturally" expected transitivity of dependency can be destroyed numerically.
2. Consider $P(X, Y, Z)$ with binary variables, where $p(x) = p(y) = 0.5$ and $p(Z|X, Y) = 1(Z = XOR(X, Y))$. That is $(X \perp\!\!\!\perp Z)$ and $(Y \perp\!\!\!\perp Z)$, but $(\{X, Y\} \not\perp\!\!\!\perp Z)$, demonstrating that pairwise independence does not imply total independence.

The Causal Markov Condition I.

Definition

A DAG G is called a causal structure over variables V, if each node represents a variable and edges denote direct influences. A causal model is a causal structure extended with local models $p(X_i|pa(X_i, G))$ for each node describing the dependency of variable X_i on its parents $pa(X_i, G)$. As the conditionals are frequently from a parametric family, they are parameterized by $\boldsymbol{\theta_i}$, and $\boldsymbol{\theta}$ denotes the overall parameterezation, so a causal model is pair $(G, \boldsymbol{\theta})$.

Definition

A causal structure G and distribution P satisfies the Causal Markov Condition, if P obeys the local Markov condition w.r.t. G.

The Causal Markov condition relies on Reichenbach's "common cause principle", i.e.the set of variables V is causally sufficient for P, that is all the common causes for the pairs $X, Y \in V$ are inside V.
(The causal Markov condition implies sufficiency and stability implies necessity of G).

Parameter priors:independence

Definition

For a Bayesian network structure G, the global parameter independence assumption means that

$$P(\boldsymbol{\theta}|G) = \prod_{i=1}^{n} p(\boldsymbol{\theta}_i|G), \tag{1}$$

where $\boldsymbol{\theta}_i$ denotes the parameters corresponding to the conditional $p(X_i|Pa(X_i))$ in G. The local parameter independence assumption means that

$$p(\boldsymbol{\theta}_i|G) = \prod_{j=1}^{q_i} p(\boldsymbol{\theta}_{ij}|G), \tag{2}$$

where q_i denotes the number of parental configurations $(pa(X_i))$ for X_i in G and $\boldsymbol{\theta}_i j$ denotes the parameters corresponding to the conditional $p(X_i|pa(X_i)_j)$ in some fixed ordering of the $pa(X_i)$ configurations. The parameter independence assumption means global and local parameter independence.

Conjugate priors, exponential family

Definition

A family $\mathcal{F}$ of prior distributions $p(\theta)$ is said to be conjugate for a class of sampling distributions $p(x|\theta)$, if the posteriors $p(\theta|x)$ also belongs to $\mathcal{F}$.

In general a conjugate prior is updated to posterior using only an appropriate statistics of the observations to update its parametrization. It shows that the parameters frequently has an intuitive interpretation based on observations, that is in the prior specification the parameters corresponds to real or virtual past observations.

The Beta distibution

Assume that x denotes the sum of 1s of n independent and identically distributed (i.i.d.) Bernoulli trials, that is we assume a binomial sampling distribution. If the prior is specified using a Beta distribution, the posterior remains a Beta distribution with updated parameters.

$$p(x|\theta) = Bin(x|n,\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \tag{3}$$

$$p(\theta) = Beta(\alpha,\beta) = c\theta^{\alpha-1}(1-\theta)^{\beta-1} \text{ where } c = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \tag{4}$$

$$p(\theta|x) = \frac{p(\theta)p(x|\theta)}{p(x)} = c'\theta^{\alpha-1+x}(1-\theta)^{\beta-1+n-x} = Beta(\alpha+x,\beta+n-x)$$

Parameter priors: The Dirichlet prior

In case of a fixed structure G (or we shall see for a fixed ordering of the variables), the usage of Dirichlets with parameter independence can be attractive on its own right to specify a parameter distribution $p(\boldsymbol{\theta}|G)$ as follows

$$p(\boldsymbol{\theta}|G) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} Dir(\boldsymbol{\theta_{ij}}|\boldsymbol{N_{ij}}) \propto \prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \theta^{N_{ijk}-1} \tag{5}$$

Structure priors

The global noninformative deviation prior [?] is derived from an a priori "reference" network structure G_0 by modeling each missing or extra edge e_{ij} independently with a uniform probability κ:

$$P(G) \propto \kappa^{\delta}, \text{where } \delta = \sum_{1 \leq i < j \leq n} I_{\{(e_{ij} \in G) \wedge (e_{ij} \notin G_0) \vee (e_{ij} \notin G) \wedge (e_{ij} \in G_0)\}}.$$

The feature priors are defined proportionally by the product of priors for the individual features (as they were totally independent). By denoting the value of feature F_i in G with $F_i(G) = f_i$ $i = 1, \ldots K$

$$P(G) = c \prod_{i=1}^{K} p(F_i(G)), \tag{6}$$

The structure modularity holds, if each feature $F_i(G)$ depends only on the parental set of X_i for $i = 1, \ldots n$, defining the parental prior

$$P(G) \propto \prod_{i=1}^{n} p(pa(X_i, G)). \tag{7}$$

DAG space I.

The cardinality of the space of DAGs is given by the following recursion [**?**]

$$f(n) = \sum_{i=1}^{n} (-1)^{i+1} 2^{i(n-1)} f(n-i) \; with \; f(0) = 1. \tag{8}$$

This is bounded above with the number of the combinations of the edges between different nodes ($2^{n(n-1)}$), because of the exclusions by the DAG-constraint. But it is still super-exponential even with a bound k on the maximum number of parents (consider that the number of parental sets for a given ordering of the variables is in the order of n^{kn}, so $2^{\mathcal{O}(kn \log n)}$.

DAG space II.

The number of orderings, DAGs and order-compatible DAGs with parental constraints. The columns shows respectively the number variables (nodes) (n), DAGs ($|DAG(n)|$), DAGs compatible with a given ordering ($|G_{\prec}|$), DAGs compatible with a given ordering and with maximum parental set size <=4 ($|G_{\prec}^{|\pi|\leq 4}|$) and <=2 ($|G_{\prec}^{|\pi|\leq 2}|$), the number of orderings (permutations) ($|\prec|$) and the total number of parental sets in an order-compatible DAG $|\pi^{\prec}|$ and in an order-compatible DAG with maximum parental set size <=4 ($||\pi^{\prec}| \leq 4|$) and <=2 ($||\pi^{\prec}| \leq 2|$).

n	$\|DAG(n)\|$	$\|G_{\prec}\|$	$\|G_{\prec}^{\|\pi\|\leq 4}\|$	$\|G_{\prec}^{\|\pi\|\leq 2}\|$	$\|\prec\|$	$\|\pi^{\prec}\|$	$\|\|\pi^{\prec}\| \leq 4\|$	$\|\|\pi^{\prec}\| \leq 2\|$
5	2.9e+004	1e+003	1e+003	6.2e+002	1.2e+002	30	30	24
6	3.8e+006	3.3e+004	3.2e+004	9.9e+003	7.2e+002	62	61	40
7	1.1e+009	2.1e+006	1.8e+006	2.2e+005	5e+003	1.3e+002	1.2e+002	62
8	7.8e+011	2.7e+008	1.8e+008	6.3e+006	4e+004	2.5e+002	2.2e+002	91
9	1.2e+015	6.9e+010	2.9e+010	2.3e+008	3.6e+005	5.1e+002	3.8e+002	1.3e+002
10	4.2e+018	3.5e+013	7.5e+012	1.1e+010	3.6e+006	1e+003	6.4e+002	1.7e+002
15	2.4e+041	4.1e+031	2.1e+027	3.1e+019	1.3e+012	3.3e+004	4.9e+003	5.7e+002
35	2.1e+213	1.3e+179	1.8e+109	8.5e+068	1e+040	3.4e+010	3.8e+005	7.2e+003

The complexity of BN learning

The NP-hardness of finding a Bayesian network for the observations .

Theorem

Let $\boldsymbol{V}$ be a set of variables with joint distribution $p(\boldsymbol{V})$. Assume that an oracle is available that reveals in $\mathcal{O}(1)$ time whether an independence statement holds in p. Let $0 < k \leq |\boldsymbol{V}|$ and $s = \frac{1}{2}n(n-1) - \frac{1}{2}k(k-1)$. Then, the problem of deciding whether or not there is a (non-minimal) Bayesian network that represents p with less or equal to s edges by consulting the oracle is NP-complete.

The NP-hardness of finding a best scoring Bayesian network (i.e. the NP-hardness of optimization over DAGs).

Theorem

Let $\boldsymbol{V}$ be a set of variables, D_N is a complete data set, $S(G, D_N)$ is a score function and a real value c. Then, the problem of deciding whether or not there exist a Bayesian network structure G_0 defined over the variables $\boldsymbol{V}$, where each node in G_0 has at most $1 < k$ parents, such that $c \leq S(G_0, D_N)$ is NP-complete.

Constraint-based BN learning: IC

The Inductive Causation algorithm (assuming a stable distribution P):

1. *Skeleton:* Construct an undirected graph (skeleton), such that variables $X, Y \in \boldsymbol{V}$ are connected with an edge iff $\forall S (X \perp\!\!\!\perp Y | S)_P$, where $S \subseteq \boldsymbol{V} \setminus \{X, Y\}$.
2. *v-structures:* Orient $X \to Z \leftarrow Y$ iff X, Y are nonadjacent, Z is a common neighbour and $\neg \exists S$ that $(X \perp\!\!\!\perp Y | S)_P$, where $S \subseteq \boldsymbol{V} \setminus \{X, Y\}$ and $Z \in S$.
3. *propagation:* Orient undirected edges without creating new v-structures and directed cycle.

Theorem

The following four rules are necessary and sufficient.

R_1 *if* $(a \not\to c) \wedge (a \to b) \wedge (b - c)$, *then* $b \to c$

R_2 *if* $(a \to c \to b) \wedge (a - b)$, *then* $a \to b$

R_3 *if* $(a - b) \wedge (a - c \to b) \wedge (a - d \to b) \wedge (c \not\to d)$, *then* $a \to b$

R_4 *if* $(a - b) \wedge (a - c \to d) \wedge (c \to d \to b) \wedge (c \not\to b) \wedge (a - d)$, *then* $a \to b$

Local Causal inference: inferring about hidden confounders

The Causal Markov Condition (i.e. the assumption of no hidden common causes) guarantees that from the observation of no more than three variables we can infer causal relation as follows. The direct dependencies between X, Y and Y, Z without direct dependence between X, Z and without conditional independence such that $(X \perp\!\!\!\perp Z | \{Y, S\})$ (i.e. with conditional dependence) should be expressed with a unique converging orientation $X \rightarrow Y \leftarrow Z$ according to the global semantics (i.e. DAG-based relation $(X \perp\!\!\!\perp Y | Z)_G$ from Def. **??**) resulting in a v-structure. If potential confounders are not excluded a priori, we have to observe at least one more variable to possibly exclude that direct dependency is caused by a confounder. Continuing the example, assume furthermore that we observe a forth variable W with the direct dependence Y, W and conditional independence $(W \perp\!\!\!\perp \{X, Z\} | Y)$ (because of stability W depends on X and Z). As Y induces independence the global semantics dictates an $Y \rightarrow W$ (note the earlier v-structure) and it cannot be mediated by a confounder $* \; Y \rightarrow * \rightarrow W$ (Y as an effect would not block).

The ML learning: Optimality of relative frequencies

Relative frequency is a ML estimator in multinomial sampling:
Assume $i = 1, \ldots K$ outcomes assuming multinomial sampling with parameters $\theta = \{\theta_i\}$ and observed occurrencies $n = \{n_i\}$ ($N = \sum_i n_i$). Then

$$\log \frac{p(n|\theta^{ML})}{p(n|\theta)} = \log \frac{\prod_i (\theta_i^{ML})^{n_i}}{\prod_i (\theta_i)^{n_i}} = \sum_i n_i \log \frac{\theta_i^{ML}}{\theta_i} = N \sum_i \theta_i^{ML} \log \frac{\theta_i^{ML}}{\theta_i} > 0.$$

where the last quantity is the "KL-distance", which is always positive: if $\hat{p}_i, p_i$ are discrete probability distributions, the Kullback-Leibler (semi)distance KL are as follows (it is always positive)

$$\mathrm{KL}(\underline{\boldsymbol{p}} \| \underline{\hat{\boldsymbol{p}}}) = \sum_i p_i \log(p_i/\hat{p}_i) \tag{9}$$

$$0 < KL(\theta^{ML} || \theta) \tag{10}$$

$$-KL(p||q) = \sum_i p_i \log(q_i/p_i) \leq \sum_i p_i((q_i/p_i) - 1) = 0 \tag{11}$$

using $\log(x) \leq x - 1$.

The ML learning I.

It can be shown that this is maximized by the selection of $\theta^*_{ijk} = N_{ijk}/N_{ij+}$, where N_{ijk} are the occurrences of value x_k and parental configuration q_j for variable X_i and its parental set $Pa(X_i)$ (N_{ij+} is the appropriate sum). By substituting this maximum likelihood parameter selection back, we get

$$ML(G; D_N) = p(D_N|G, \boldsymbol{\theta}^*) = \prod_{l=1}^{N} \prod_{i=1}^{n} p(x_i^{(l)}|pa_i^{(l)}) \tag{12}$$

$$= \prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \frac{N_{ijk}}{N_{ij+}}^{N_{ijk}} \tag{13}$$

by taking logarithm, rearranging and expanding with N

$$\log(ML(G; D_N)) = N \sum_{i=1}^{n} \sum_{j=1}^{q_i} \frac{N_{ij+}}{N} \sum_{k=1}^{r_i} \frac{N_{ijk}}{N_{ij+}} \log(N_{ijk}/N_{ij+}) \tag{14}$$

The ML Learning II

Using conditional entropy $H(Y|X) = \sum_x p(x) \sum_y p(y|x) \log(p(y|x))$, the chain rule $H(X,Y) = H(Y|X) + H(X)$ and the definition of mutual information $I(Y;X) = H(Y) - H(Y|X)$, it can be rewritten as

$$\log(ML(G;D_N)) = -N \sum_{i=1}^{n} H(X_i|Pa(X_i,G)) \quad (15)$$

$$= N \sum_{i=1}^{n} I(X_i;Pa(X_i,G)) - N \sum_{i=1}^{n} H(X_i) \quad (16)$$

$$(17)$$

This shows that the maximization of the ML score is equivalent with finding a BN parameterized with the observed frequencies that has minimum entropy or that we are finding a BN parameterized with the observed frequencies that has maximum mutual information between its children and their parents (16, Note the close connection of this reading to the concept that causal ordering is related to the (maximal) determination of each variable by the earlier variables.

Complexity regularization

Because of the monotonicity of mutual information — if $Pa(X_i) \subset Pa(X_i)'$, then $I(X_i; Pa(X_i)) \leq I(X_i; Pa'(X_i))$ — so the complete network maximizes the maximum likelihood score. However score functions such as the MDL-score derived from the minimum description length (MDL) principle or the Bayesian information criterion (BIC)-score derived with a non-informative Bayesian approach contains various complexity penalty terms. We shall use only the BIC-score defined as follows (for overviews of other score functions and for the derivation of the BIC-score)

$$BIC(G; D_N) = \log(ML(G; D_N)) - 1/2 dim(G) \log(N) \tag{18}$$

where $dim(G)$ denotes the number of free parameters.

Score equivalence: BIC

Definition

A score function $S(G; D_N)$ is called score equivalent, if for each pair of observationally equivalent Bayesian network structure G_1, G_2 the scores are equal $S(G_1; D_N) = S(G_2; D_N)$ for all D_N.

Theorem

The $BIC(G; D_N)$ scoring metric is score equivalent .

The score equivalence of the *BIC* score is the direct consequence of the result that the number of free parameters (that is the term $dim(G)$) are equal in observationally equivalent Bayesian networks.

Asymptotic consistency

Theorem

Let $\boldsymbol{V}$ be a set of variables. Let the prior distribution $p(G)$ over Bayesian network structures be positive. Let $p(\boldsymbol{V})$ be a positive and stable distribution and G_0 is a corresponding perfect map (i.e. a Bayesian network representing exactly all the independencies in $p(\boldsymbol{V})$, see Def. **??***). Now, let D_N is an i.i.d. data set generated from $p(\boldsymbol{V})$. Then, for any network structure G over $\boldsymbol{V}$ that is not a perfect map of $p(\boldsymbol{V})$ we have that*

$$\lim_{N\to\infty} BD_e(G_0; D_N) - BD_e(G; D_N) = -\infty \text{ and also} \quad (19)$$

$$\lim_{N\to\infty} BIC_e(G_0; D_N) - BD_e(G; D_N) = -\infty \quad (20)$$

For further results about the asymptotic optimality of the scores for not stable distributions.

Rate of convergence

Furthermore, a rate of convergence result is also derived and a corresponding sample complexity $N(\epsilon, \delta)$ to select an appropriate sample size for a given accuracy between the target distribution p_0 and the distribution p_{BN} represented by the learned Bayesian network with a given confidence

$$p(D_N : KL(p_0|\hat{p}_{BN}(D_N)) > \epsilon) < \delta \tag{21}$$

The Dirichlet distribution

Assume that the observed sequence $D_n = \{X_i; i = 1, 2 \ldots, n\}$ contains i.i.d. multinomial samples with L discrete values. The prior is a Dirichlet prior with hyperparameters $\boldsymbol{\alpha} = \alpha_1, \ldots, \alpha_L$ and $\alpha_. = \sum_i \alpha_i$.

$$p(\theta) = Di(\boldsymbol{\alpha}) = c \prod_i \theta^{\alpha_i - 1} \text{ where } c = \frac{\Gamma(\alpha_.)}{\prod_i \Gamma(\alpha_i)} \tag{22}$$

Dirichlet distribution II.

It is conjugate for multinomial sampling, so the posterior predictive distributions are the updated Dirichlet with hyperparameters $\boldsymbol{\alpha}_j$ at step j and the posterior prediction for x_j (i.e. the marginal posterior probability $E[\theta_{x_j}]$) is

$$p(x_j|x_1,\ldots,x_{j-1}) = \int p(x_j|\theta)p(\theta|x_1,\ldots,x_{j-1})d\theta \quad (23)$$

$$= \int p(x_j|\theta)Dir(\boldsymbol{\theta}|\boldsymbol{\alpha_j})d\theta \quad (24)$$

$$= c\int \prod_{i=1}^{L}\theta_i^{1(x_j=r_i)}\prod_i \theta^{\alpha_{ji}-1}d\theta \text{ where } c=\frac{\Gamma(\alpha_{j,.})}{\prod_i\Gamma(\alpha_{j,i})} \quad (25)$$

$$= c\int \prod_i \theta^{\alpha_{j+1}-1}d\theta \quad (26)$$

$$= \frac{\Gamma(\alpha_{j.})}{\Gamma(\alpha_{j+1,.})}\frac{\prod_i\Gamma(\alpha_{j+1,i})}{\prod_i\Gamma(\alpha_{ji})} \quad (27)$$

$$= \frac{\alpha_{j,x_j}}{\alpha_{j,.}} \quad (28)$$

(29)

Dirichlet distribution III.

The marginal probability of the data set D_n with n_i occurrences of value r_i

$$p(x_1,\ldots,x_n|Dir(\boldsymbol{\alpha}_1)) = \prod_{i=1}^{n} p_i(x_i|x_1,\ldots,x_{i-1}) \quad (30)$$

$$= \frac{\prod_{i=1}^{L} \alpha_{1,i}..(\alpha_{1,i}+n_i)}{\alpha_{1,.}\ldots(\alpha_{1,.}+n)} \quad (31)$$

$$= \frac{\Gamma(\alpha_{1,.})}{\Gamma(\alpha_{1,.}+n)} \prod_{i=1}^{L} \frac{\Gamma(\alpha_{1,i}+n_i)}{\prod_{i=1}^{L}\Gamma(\alpha_{1,i})} \quad (32)$$

Parameter priors:likelihood equivalence

The concept of likelihood equivalence extends observational equivalence of the structure coherently to the parameters .

Definition

The likelihood equivalence assumption means that for two observationally equivalent Bayesian network structures G_1, G_2,

$$p(\boldsymbol{\theta}_V|G_1) = p(\boldsymbol{\theta}_V|G_2), \tag{33}$$

where $\boldsymbol{\theta}_V$ denotes a non-redundant set of the multinomial parameters for the joint distribution over V.

Parameter priors: Dirichlet priors

Theorem

The assumption of positive densities, likelihood equivalence and parameter independence for complete structures G_c implies that $p(\boldsymbol{\theta}_U|\xi)$ is a Dirichlet distribution with hyperparameters $N_{x_1,\ldots,x_n}$.

The $p(\boldsymbol{\theta_i}|G_i) = J_{G_i} p(\boldsymbol{\theta_V}|\xi)$, where J_{G_i} is the Jacobian of the transformation from $\boldsymbol{\theta_V}$ to $\boldsymbol{\theta_{G_i}}$. To state the following theorem it is convenient to rewrite the hyperparameters as $N' = \sum_{x_1,\ldots,x_n} N_{x_1,\ldots,x_n}$ called prior/virtual sample size and $p^{prior} x_1,\ldots,x_n = E[\theta_{x_1,\ldots,x_n}] = N_{x_1,\ldots,x_n}/N'$. Furthermore, we need the following concept.

Definition

The parameter modularity assumption means that if $pa(X_i)$ are identical in two Bayesian network structures G_1, G_2, then

$$p(\boldsymbol{\theta_{ij}}|G_1) = p(\boldsymbol{\theta_{ij}}|G_2), \tag{34}$$

where $\boldsymbol{\theta}ij$ denotes the parameters corresponding to the conditional $p(X_i|pa(X_i)_j)$ in some fixed ordering of the $pa(X_i)$ configurations.

Parameter priors: Dirichlet priors II.

The assumption of parameter modularity allows to induce parameter distributions for incomplete models from complete model.

Theorem

If $p(\boldsymbol{\theta}_V|\xi)$ is a Dirichlet distribution with hyperparameters $N_{x_1,\ldots,x_n} = N'px_1,\ldots,x_n$ and the parameter modularity assumption holds and for all complete network G_c $p(G_c) > 0$, then for any network structure G the parameter independence and the likelihood equivalence holds and the decomposed distribution of the parameters is the product of Dirichlet distributions

$$p(\boldsymbol{\theta}|G) = \prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i}\theta^{N'p^{prior}(X_i=k,pa(X_i,G)=pa_{ij})-1} \tag{35}$$

where r_i denotes the number of values of X_i, q_i denotes the number of parental configurations $(pa(X_i, G))$ for X_i in G and pa_{ij} denotes the values of the parents for the jth parental configuration in some fixed ordering of the $pa(X_i)$ configurations.

BD score

By assuming N complete observations, i.i.d. multinomial sampling, Bayesian network model with parameter independence and Dirichlet parameter priors, the observation of a complete case results in a local standard Bayesian updating of the hyperparameters of the appropriate Dirichlets and retains the parameter independence. The maintained parameter independence allows a standard parental decomposition w.r.t. the Bayesian network G for each observation, which allows the following rearrangement

$$p(\boldsymbol{C_1}, \ldots, \boldsymbol{C_N}|G) = \prod_{l=1}^{N}\prod_{i=1}^{n} p_l(x_i^{(l)}|pa_i^{(l)}) \tag{36}$$

$$= \prod_{i=1}^{n}\prod_{l=1}^{N} p_l(x_i^{(l)}|pa_i^{(l)}) \tag{37}$$

$$= \prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{l=1}^{N} p_l(x_i^{(l)}|pa_{ij})^{1(pa_{ij}=pa_i^{(l)})} \tag{38}$$

where $pa_i^{(l)}$ denotes the value(s) of parental set of X_i in case l.

BD score II

This can be combined with the earlier result of the marginal probability of the data for a single Dirichlet prior and multinomial sampling.That is for each variable X_{i_0} and parental configurations j_0 independently

$$\prod_{l=1}^{N} p_l(x_{i_0}^{(l)}|pa_{i_0j_0}, G)^{1(pa_{i_0j_0}=pa_{i_0}^{(l)})} = \frac{\prod_{k=1}^{r_{i_0}} \alpha_{i_0j_0k}..(\alpha_{i_0j_0k}+n_k)}{\alpha_{i_0j_0+}\ldots(\alpha_{i_0j_0+}+n)} \quad (39)$$

$$= \frac{\Gamma(\alpha_{i_0j_0+})}{\Gamma(\alpha_{i_0j_0+}+n_{i_0j_0+})} \prod_{k=1}^{r_{i_0}} \frac{\Gamma(\alpha_{i_0j_0k+n_{i_0j_0k}})}{\Gamma(\alpha_{i_0j_0k})}$$

where r_i denotes the cardinality of the discrete values of variable X_i, α_{ijk} the initial Dirichlet hyperparameters and n_{ijk} the number of occurrences for the variable X_i, its parental configuration pa_{ij} and its value r_k. The sign $+$ denotes the appropriate marginals.

BD score III.

Putting everything together, if the prior satisfies the structure modularity, then the posterior of the Bayesian network (structure) has the following product form

$$p(G|D_N) \propto \prod_{i=1}^{n} p(Pa(X_i, G))S(X_i, Pa(X_i, G), D_N) \text{ where} \tag{40}$$

$$S(X_i, Pa(X_i, G), D_N) = \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij+})}{\Gamma(\alpha_{ij+} + n_{ij+})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk+n_{ijk}})}{\Gamma(\alpha_{ijk})}. \tag{41}$$

Score equivalence: BD

Theorem

The $BD_e(G; D_N)$ scoring metric is likelihood equivalent, that is if G_1, G_2 are observational equivalent, then $p(D_N|G_1) = p(D_N|G_2)$. Furthermore, if the hypotheses are the equivalence classes or the prior is equal for such G_1, G_2, then the BD_e scoring metric is score equivalent.

Thank you for your attention!

Questions?