Noname manuscript No.
(will be inserted by the editor)

Compare and contrast on Bayesian network structure learning metrics

First Author · Second Author · Third Author

Received: date / Accepted: date

Abstract The purpose of this paper is to provide a theoretical review on metrics used for learning Bayesian networks structures. There are at least a handful metrics that can be used if one wants to learn structures using the so called metric-based approach. It is, however, not clear why or why not one metric is prefered over another other than being popular. This paper aims at reviewing these metrics from a theoretical point of view by looking at their assumptions, priors if there are any, and difficulty being applied to general cases.

1 Introduction

Introduce the problem of causal discovery / Bayesian network structure learning. Intrdocue the metric-based approach. Introduce the importance of having a "good" scoring function. And perhaps a short chronological history of the invention of metrics. Assumptions:

- iid samples
- complete data
- independent parameter values
- uniform parameter values
- parameter modularity (heckerman1995)

[Liu *et al.*, 2012] an empirical study on BN structure learning metrics, including AIC, BDeu, MDL, and fNML. look at its reference list for comparison papers.

F. Author
first address
Tel.: +123-45-678910
Fax: +123-45-678910
E-mail: fauthor@example.com

S. Author
second address

Table 1: Notations

G, B_S	a directed acyclic graph or a Bayesian network structure
B_P	the joint distribution with which B_S forms a complete Bayesian network (B_S, B_P)
Θ	a set of conditional probabilities or parameters of a Bayesian network
$\lvert \cdot \rvert$	the cardinality of a set
X	a variable set
X_i	the ith variable/node in X
x_j	the jth state of a variable
Π_i	a parents set of the variable X_i
Π_{ij}	the jth variable in Π_i
$\prod_{j=1}^{\lvert\Pi_i\rvert} \Pi_{ij}$	Cartesian product of parents states
π_{ij}	the jth state in $\prod_{j=1}^{\lvert\Pi_i\rvert} \Pi_{ij}$
r_{Π_i}	$\lvert \prod_{j=1}^{\lvert\Pi_i\rvert} \Pi_{ij} \rvert$, total number of parents' states combination
D	a dataset
$D^{G_i=j}$	the rows of D where the set of variables G_i takes the jth instantiation
$D_i^{G_i=j}$	the ith column of the $D^{G_i=j}$ rows
r_i	the arity (a.k.a., number of states) of a variable X_i
$\mathbf{n_{ij}}$	a vector of counts for all states of X_i given $\P_i$ is in state j
n_{ijk}	$n_{ijk} \in \mathbf{n_{ij}}$ is the count of X_i is in state k given Π_i is in state j, also known as sufficient statistics
n_{ij}	the count of Π_i is in state j i.e., $\sum_{k=1}^{r_i} n_{ijk}$
$\boldsymbol{\alpha}$	a vector of Dirichlet concentration parameters
α_i	the ith parameter in $\boldsymbol{\alpha}$
$B(x, y)$	the beta function

[Allen and Greiner, 2000] an empirical comparison among AIC, MDL and a cross-validation criteria on BN structure learning, suggest MDL is the worst among these three.

2 Bayesian network

Introduce basic concepts in Bayesian networks and define notaions that will be used later.

2.1 Notations

3 Scoring functions

Brief introduction

3.1 Information theoretical approach

[Wallace and Boulton, 1968] MML was first introduced as an inductive inference principle. Then applied on learning causal model by [Wallace *et al.*, 1996], [Neil *et al.*, 1999], [Li *et al.*, 2004], [O'Donnell, 2010].

3.1.1 Akaike information criteria

[AKAIKE, 1973] introduced Akaike information criteria (AIC) as a model selection metric

$$AIC = -2\log(L(\hat{\theta} \mid D)) + 2K, \tag{1}$$

where $\hat{\theta}$ is the maximum likelihood estimation of the true model parameters, K is the total number of parameters in a candidate model. The development of AIC was based on Kullback-Leibler divergence (also known as Kullback-Leibler information)

$$\begin{aligned} KLD &= \int_x f(x) \log \frac{f(x)}{g(x)} dx \qquad (2) \\ &= \int_x f(x) \log f(x) dx - \int_x f(x) \log g(x) dx \qquad (3) \end{aligned}$$

where $f(x)$ is the *p.d.f.* of the true distribution (unknown in reality), $g(x)$ is a candidate distribution used to estimate $f(x)$. The first term in equation 3 is a constant when considering candidate models, and hence minimising the Kullback-Leibler divergence is equivalent to maximising the second integration. Since the true distribution is unknown, the KLD can only be estimated from data and hence the motivation is to find the smallest expected KLD over data. AKAIKE proved that the maximised log likelihood value is a biased estimate of the expectation of the second integral with approximately a constant K difference. Hence, the AIC metric was developed. For more details, refer to [Burnham and Anderson, 2004].

AIC is only unbiased estimate of the expected estimated KLD if the true model is in the space of model selection. And its performance can be poor if K is relatively large to the sample size.

AIC tends to overfit, though no theoretical explanation on its behavior of learning BN was given. [Liu *et al.*, 2012] compared AIC, MDL, BDeu and fNML, reported that AIC's behavior is hard to predict.

AIC's penalty term only replies on number of parameters, but not precision of parameters, unlike MML.

3.1.2 Minimum message length

The following is an MML metric for a CPT model written in three different mathematical expression.

$$\begin{aligned} I(\phi_i(\bar{X}), D_{\phi_i}) &= \sum_{k=1}^{K} \ln\left(\frac{(n_k + \alpha_0 - 1)! \prod_{j=1}^{m} (\alpha_j - 1)!}{(\alpha_0 - 1)! \prod_{j=1}^{m} (n_{kj} + \alpha_j - 1)!}\right) \qquad (4) \\ &= \sum_{k=1}^{K} \ln \frac{\Gamma(n_k + \alpha_0) \prod_{j=1}^{m} \Gamma(\alpha_j)}{\Gamma(\alpha_0) \prod_{j=1}^{m} \Gamma(n_{jk} + \alpha_j)} \qquad (5) \\ &= \sum_{k=1}^{K} \ln \frac{B(\boldsymbol{\alpha})}{B(\mathbf{n_{\cdot k}} + \boldsymbol{\alpha})}, \qquad (6) \end{aligned}$$

where $B(\boldsymbol{\alpha}) = \frac{\prod_i \Gamma(\alpha_i)}{\Gamma(\sum_i \alpha_i)}$ is a Gamma function. There is also a constant term for reach parameter that corresponds to the constant difference between the adaptive and MML87 derivation of a CPT model. Ther term is

$$\frac{r_i|\Pi_i|}{2}\ln\frac{\pi e}{6}$$

3.1.3 Minimum description length

Minimum description length (MDL) principle was developed by [Rissanen, 1978] for statistically inferring an optimal model based on a given dataset. It shares some properties with the MML principle but with a fundamental difference of being a non-Bayesian approach (at least in its original derivation). MDL is also a formal version of Occam's razor, so its general form is also a two-part message (or description) length including the cost for encoding a model and the cost for encoding data given this model.

Other than being a non-Bayesian approach, MDL also differs from MML in a few more ways [Baxter and Oliver, 1994]

- the objective of MDL is fine the optimal model class for prediction without having to state a specific model and its parameter, in contrast MML searches for the optimal model and set of parameters that has the shortest total message length.
-

MDL has been applied in a wide range of machine learning problems, including BN structure learning. There are a number of different variations of MDL for structure learning problem. [Bouckaert, 1994] interpreted MDL as

$$MDL = \log p(B_S) - \log p(D \mid (B_S, \hat{B}_P)) - \frac{k}{2}\log N$$

where B_S is a candidate network structure, $\hat{B}_P$ is the maximum likelihood estimation of the parameters in B_S, $k = \sum_{X_i \in X}(r_i - 1)r_{\Pi_i}$ is the number of free parameters, and N is the number of samples in a given dataset D. The author used uniform structure prior for both K2 and MDL and empirically justified that they have similar performance for large samples. But for small and moderate samples, MDL selected networks tend to have less number of parents than K2 selected networks.

[Cruz-Ramírez *et al.*, 2006]

Look into different version of MDL by Suzuki, Lam and Bacchus.

Cruz-Ramffirez et al. (2006) compared MDL against BIC and claimed the former is simlar as the latter but with a penalty term.

MDL is not a Bayesian approach? This is the fundamental difference.

General form of MDL

$$-\log\hat{p}(x^n) + \frac{k}{2}\log n + O(1)$$

The first two terms is equivalent to BIC.

[Lam and Bacchus, 1994] applied MDL on learning BN structures.

3.1.4 Entropy score

Chow and Liu's work also based on entropy but for learning trees An early work on using information theory to learn BN was developed by [Herskovits and Cooper, 1990] in an algorithm named Kutató. The metric used in it is simply the sum of conditional entropies for each variable given its parents set

$$\begin{aligned} H_{BN} &= \sum_{X_i \in X} H(X_i \mid \Pi_i) \\ &= -\sum_{X_i \in X} \sum_{\pi_{ik}=1}^{r_{\Pi_i}} \sum_{x_i=1}^{r_i} p(X_i = x_i, \Pi_i = \pi_i) \ln p(X_i = x_i \mid \Pi_i = \pi_i). \end{aligned} \tag{7}$$

The Kutató algorithm was computationally inefficient but it has shown an idea of ranking network structures based on the amount of information they carry about a given dataset.

3.1.5 Mutual information test

[de Campos, 2006] introduced a new scoring function called mutual information test (MIT). The motivation was to develop a metric that measures the 'distance' from the true distribution B_P to the estimated distribution $\hat{B}_P$ which is obtained by maximum likelihood estimation of a learned network structure from data. One way of measuring distribution difference is by Kullback-Leibler divergence that can be expressed by entropy and mutual information

$$KL(\hat{B}_P, B_P) = -H_D(X) + \sum_{X_i \in X} H_D(X_i) - \sum_{X_i \in X}^{\Pi_i \neq \emptyset} MI_D(X_i, \Pi_i).$$

Because the first two terms are invariant under structures, minimising the LHS is equivalent to maximising the summation on the RHS. That is, finding an optimal parents set for each variable such that the sum of the multual information is maximised. Since mutual information will not decrease by including additional variables, de Campos used χ^2 values to prevent overfitting and shown it is a legitimate regularization term based on the following theorem.

Theorem 1 ***(Kullback, 1968)***
Given a dataset D with n elements, if the hypothesis that X and Y are conditionally independent given Z is true, then the statistics $2NMI_D(X, Y \mid Z)$ approximates to a distribution $\chi^2(l)$ with $l = (r_X - 1)(r_Y - 1)r_Z$ degrees of freedom, where r_X, r_Y, r_Z represent the number of states of variables X, Y, Z respectively. If $Z = 0$, the statistics $2NMI_D(X, Y)$ approximates to a distribution $\chi^2(l)$ with $l = (r_X - 1)(r_Y - 1)$ degrees of freedom.

The final expression of the MIT metric that measures the fitness of a network structure B_S to a given dataset D with a χ^2 regularization term is

$$g_{MIT}(B_S : D) = \sum_{X_i \in X}^{\Pi_i \neq \emptyset} \left(2NMI_D(X_i, \Pi_i) - \max_{\sigma_i} \sum_{j=1}^{|\Pi_i|} \chi_{\alpha, l_{i\sigma_i(j)}} \right), \tag{8}$$

where the regularizer is a sum of quantiles defined as $p(\chi^2(l_{i\sigma_i(j)}) \leq \chi_{\alpha,l_{i\sigma_i(j)}}) = \alpha$ for a pre-specified significant value α. The degrees of freedom $l_{i\sigma_i(j)} = (r_i - 1)(r_{\sigma_i(j)} - 1) \prod_{k=1}^{j-1} r_{\sigma_i(k)}$, where $\sigma_i(j)$ is the jth element in any permutation of Π_i. The maximisation is over all permutations of Π_i because different ways of decomposing Π_i results in different degrees of freedom hence different χ^2 values.

The way equation 8 is expressed makes it a decomposable metric, but not score-equivalent due to χ^2's degrees of freedom unless all variables have the same number of states. MIT, however, was proved to be score-equivalent in the *restricted partial DAG* (RPDAG) space which was used by [Acid and de Campos, 2003] for Bayesian network structure learning. Assumptions:

- complete data,
- no latent variable,

3.2 Posterior approach

A Bayesian measure of the goodness of fit of a model to a given dataset is the posterior probability of the model given such dataset.

3.2.1 Bayesian information criteria

[Schwarz, 1978]

$$BIC = -2 \ln L + K \log n \tag{9}$$

When apply BIC, the model selection space does not have to contain the true model, the BIC selected model does not have to be the data generating model.

3.2.2 Buntine's posterior metric

[Buntine, 1991a] draw the connection between learning classification trees and Bayesian networks (as well as any other probabilistic models), and applied the same strategy for learning trees to BNs. [Buntine, 1991b] addressed the learning of an optimal Bayesian network starting from partial domain knowledge is an example of theory refinement. Such a refinement process can be completed by a heuristic search and an objective function that is expressed as the product of local conditional probability posteriors (i.e., a parents set and the conditional probabities) over all nodes. The full conditional probabilities can be expressed by a CPT which relies on the parents set. Hence, the first problem to be solved is the a BN structure. The structural posterior is obtained by integrating over all parameters, which is propotional to prior times likelihood

$$\begin{aligned}
p(G \mid D) &= \int_{\Theta} p(G, \Theta \mid D) d\Theta \\
&\propto \int_{\Theta} p(G)p(\Theta \mid G)p(D \mid G, \Theta) d\Theta \\
&= \int_{\Theta} \prod_{X_i \in X} p(\Pi_i) \frac{\prod_{j=1}^{|\Pi_i|} \prod_{k=1}^{r_i} \theta_{ijk}^{\alpha_k - 1}}{B(\boldsymbol{\alpha})} \prod_{j=1}^{|\Pi_i|} \prod_{k=1}^{r_i} \theta_{ijk}^{n_{ijk}} d\Theta \\
&= \prod_{X_i \in X} p(\Pi_i) \prod_{j=1}^{|\Pi_i|} \frac{1}{B(\boldsymbol{\alpha})} \int_{\Theta} \prod_{k=1}^{r_i} \theta_{ijk}^{n_{ijk} + \alpha_k - 1} d\Theta \\
&= \prod_{X_i \in X} p(\Pi_i) \prod_{j=1}^{|\Pi_i|} \frac{B(\mathbf{n_{ij}} + \boldsymbol{\alpha})}{B(\boldsymbol{\alpha})}.
\end{aligned} \tag{10}$$

Given a parents set of each node, a network structure is defined. Assuming the choice of parents set for each node is independent, the structure prior is a product of parents set prior for each node. The structure prior is $p(\Pi)$ is often assumed to be uniform hence omitted for being a constant. Since each local structure of a node given its parents set forms a conditional probability table, which in turn can be patitioned into $|\Pi_x|$ many multinomial distributions, the parameter prior $p(\Theta \mid \Pi)$ is assumed to follow a symmetric Dirichlet distribution. An computational advantage of using a Dirichlet prior its conjugate property that ensures the posterior is in the same distribution family as the prior.

The assumptions Buntine made were:

- a node (possibly causal) ordering being provided by domain experts instead of the usual faithfulness (perfect-map) assumption, because faithfulness does not always hold in practice. The metric stays unchanged with or without a node ordering.
- complete data with no missing values, which later briefly mentioned can be dealt with EM algorithm.
- conditional probabilities of a node given its parents set is expressed by a full CPT. This assumption can also be relaxed by approximating the condition probaibilities by lower dimension distributions (or restricted/local probability models such as noisy-OR-gate [Pearl, 1988], trees, lower order logit models [Neil *et al.*, 1999] that have linear number of parameters.
- parameter priors are indpendent under parents instantiations, so that total parameter prior is a product of individual parameter priors.
- independence betweem local structure (i.e., a node and its parents sets) so that the posterior of a global structure can be a product of the local structure posteriors,
- samples are iid

To ensure equivalent structures have the same score, [Buntine, 1991b] used a hyper-parameter α which is known as equivalent sample size (ess) to derive symmetric Dirichlet's concentration paremters for each variable

$$\boldsymbol{\alpha} =< \frac{\alpha}{r_i|\Pi_i|} > . \tag{11}$$

The factorization $p(X) = \prod_{X_i \in X} p(X_i \mid \Pi_i)$ and the assumption of an ess (11) makes the instantiations of the joint ditribution $p(X)$ uniformly distributed. Hence, this score is often refered as the BDeu (Bayesian Dirichlet equivalent uniform).

3.2.3 K2

Around the same time, [Cooper and Herskovits, 1992] derived the same metric as [Buntine, 1991b] using similar assumptions, except the specific uniform distribution assumption on model parameters. The motivation was to compare two network structure's posterior $p(G \mid D$, which is equivalent to comparing their joint density $p(G, D)$ because the normaling term is invariant under structures. The metric has the form

$$\begin{aligned} p(G, D) &= p(G) \prod_{X_i \in X} \prod_{j=1}^{|\Pi_i|} \frac{(r_i - 1)! \prod_{k=1}^{r_i} n_{ijk}!}{(n_{ij} + r_i - 1)!} \\ &= \prod_{X_i \in X} p(\Pi_i) \prod_{j=1}^{|\Pi_i|} \frac{B(\mathbf{n_{ij}} + < 1 >)}{B(< 1 >)} \qquad (12) \\ & \qquad (13) \end{aligned}$$

It is the same as equation 10 under the uniform parameter prior circumstance when $\boldsymbol{\alpha} =< 1 >$. Cooper also extented the score to general case of Dirichlet prior.

Assumptions:

- X is a set of discrete variables so that the likelihood function is a probability mass function and integration is over finite sets. (but why? what's the problem with continuous?)
- no hidden/latent variables (relaxed later)
- D is complete. That is, no missing values. (relaxed later)
- samples in D are *i.i.d.* so that the likelihood $p(D \mid G, \Theta)$ is a product of each sample.
- parameter values are uniformly distributed.

Nothing about equivalent networks is mentioned. Heckerman1995 said K2 does not give same score to equivalent structures under the uniform Dirichlet assumption.

K2 uses an uninformative prior $\alpha_{ijk} = 1$ for parameters.

K2 is the same as the original mml metric (i.e., start counting from 1 in adaptive approach).

3.2.4 BDe

parameter modularity and score equivalent (metrics for bn should be equivalent, but not for causal net)

With the assumptions stated by [Cooper and Herskovits, 1992] and [Buntine, 1991a] and an additional parameter modularity assumption (that states local model parameters only depend on parents set), [Heckerman *et al.*, 1995] developed a more general metric named Bayesian Dirichlet equivalent (BDe). Although the BDe metric uses Dirichlet parameter priors, but [Heckerman *et al.*, 1995] proved the Dirichlet assumption can be relaxed as long as parameters takes positive real values. The metric relies on the same hyper-parameter ess as [Buntine, 1991b]'s metric, but assign non-uniform priors to parameters based on user's prior knowlege of the joint distribution $p(X)$. [Heckerman *et al.*, 1995] proved the BDe metric is score equivalent and worked through a toy example to demonstrate this property. The sensitivity of BDe about ess was also emphasized by the authors and concluded that small values of ess will result the metric quickly in favor of different learned BN from the prior BN. The conclusion is not surprising since small α results in small Dirichlet prior for each parameter, hence its impact on the outcome quicly diminishes as more data come in.

Assumptions:

- X is a set of discrete variables.
- D's rows are exachangable, meaning iid.
- D is complete.
- parameter independence.
- defined an event B_S^e for the structure B_S being B_S is an I-map (perhaps perfect-map) of the underlying probability distribution.
- hence introduced the event and score equivalence propositions.
- parameter modularity, meaning CPT parameters only depends on parents set, not the global network structure.
- parameters follow a dirichlet distribution.

3.2.5 Others

[Spiegelhalter *et al.*, 1993] a different score, perhaps similar as BDe. (hard to understand)

[de Campos, 2006] MIT score.

[Silander *et al.*, 2008] a new score using factorized nomalized maximum likelihood (fNML), no tunable hyper-parameter as BDe, hence avoid having sensitivity issue for small samples. AIC and BIC are decomposable scores. The penalty terms are not a function of the data, but functions of structure and r_i. BDe with $\alpha_{ijk} = 1$ is K2, and $\alpha_{ijk} = \frac{\alpha}{r_i|\Pi_i|}$ is BDeu [Buntine, 1991b], where the single hyper-parameter α is called the equivalent sample size (ess), which is equivalent to the initial counting value in MML.

[Riggelsen, 2008] proposed a new score (but similar as BDe) for learning Bayesian networks. The only difference is that this new score finds the optimal pair of DAG and parameters for a given dataset, rather than just finding the optimal DAG as BDe does (according to the author). The parameters are estimated using MAP estimation, and hence this new score is called MAP BN.

4 Assumptions

Common assumptions from the previous mentioned metrics. Can they be relaxed or not? Which one has a stronger assumption? Does it make any difference in real world?

5 Similarities and differences

Compare and contrast each metrics. Draw conclusion on their similarities, such as under what circumstances or what sort of data, several metrics produce the same score. And when will they be different? The difference is caused by what? Assumtipion on priors?

[Steck and Jaakkola, 2003] emphasised the sensitivity of ess.

[Silander *et al.*, 2007] emphasized the sensitivity of the BDeu score on the number of arcs identified for variety values of the equivalent sufficient statisics α in it. The conclusion of large α results in more arcs was made mainly on experimental results, but with reasonable explanations.

[Steck, 2008] on optimizing ess.

6 Summary

References

[Acid and de Campos, 2003] S. Acid and L. M. de Campos. Searching for Bayesian network structures in the space of restricted acyclic partially directed graphs. *Journal of Artificial Intelligence Research*, 18:445–490, 2003.

[AKAIKE, 1973] H. AKAIKE. Information theory and an extension of the maximum likelihood principle. In *Proceedings of the 2nd International Symposium on Information Theory*, pages 267–281. Akademiai Kiado, 1973.

[Allen and Greiner, 2000] T. V. Allen and R. Greiner. Model Selection Criteria for Learning Belief Nets: An Empirical Comparison. In *Proceedings of the 17th International Conference on Machine Learning*, pages 1047–1054. Morgan Kaufmann Publishers Inc., 2000.

[Baxter and Oliver, 1994] R. A. Baxter and J. J. Oliver. MDL and MML: similarities and differences. Technical report, Monash University, Department of Computer Science, 1994.

[Bouckaert, 1994] R. R. Bouckaert. Properties of Bayesian belief network learning algorithms. In *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence*, pages 102–109. Morgan Kaufmann Publishers Inc., 1994.

[Buntine, 1991a] W. L. Buntine. Classifiers: a theoretical and empirical study. In *Proceedings of the 12th International Joint Conference on Artificial Intelligence*, pages 638–644. Morgan Kaufmann Publishers Inc., 1991.

[Buntine, 1991b] W. L. Buntine. Theory refinement on Bayesian networks. In *Proceedings of the 7th Conference on Uncertainty in Artificial Intelligence*, pages 52–60. Morgan Kaufmann Publishers Inc., 1991.

[Burnham and Anderson, 2004] K. P. Burnham and D. R. Anderson. Multimodel inference: understanding AIC and BIC in model selection. *Sociological methods & research*, 33(2):261–304, 2004.

[Cooper and Herskovits, 1992] G. F. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309–347, 1992.

[Cruz-Ramírez *et al.*, 2006] N. Cruz-Ramírez, H. G. Acosta-Mesa, R. E. Barrientos-Martínez, and L. A. Nava-Fernández. How good are the Bayesian information criterion and the minimum description length principle for model selection? A Bayesian network analysis. In *Proceedings of the 5th Mexican International Conference on Artificial Intelligence*, pages 494–504. Springer, 2006.

[de Campos, 2006] L. M. de Campos. A scoring function for learning Bayesian networks based on mutual information and conditional independence tests. *Journal of Machine Learning Research*, 7(Oct):2149–2187, 2006.

[Heckerman *et al.*, 1995] D. Heckerman, D. Geiger, and D. M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3):197–243, 1995.

[Herskovits and Cooper, 1990] E. Herskovits and G. F. Cooper. Kutató: an entropy-driven system for construction of probabilistic expert systems from databases. In *Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence*, pages 117–128. Elsevier Science Inc., 1990.

[Lam and Bacchus, 1994] W. Lam and F. Bacchus. Learning Bayesian belief networks: an approach based on the MDL principle. *Computational intelligence*, 10(3):269–293, 1994.

[Li *et al.*, 2004] G. Li, H. Dai, and Y. Tu. Identifying Markov blankets using lasso estimation. In *Advances in Knowledge Discovery and Data Mining*, number 2004, pages 308–318. Springer, 2004.

[Liu *et al.*, 2012] Z. Liu, B. Malone, and C. Yuan. Empirical evaluation of scoring functions for Bayesian network model selection. In *BMC bioinformatics*, volume 13, page S14. BioMed Central, 2012.

[Neil *et al.*, 1999] J. R. Neil, C. S. Wallace, and K. B. Korb. Learning Bayesian networks with restricted causal interactions. In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, pages 486–493. Morgan Kaufmann Publishers Inc., 1999.

[O'Donnell, 2010] R. T. O'Donnell. *Flexible Causal Discovery with MML*. Monash University, 2010.

[Pearl, 1988] J. Pearl. *Probabilistic reasoning in intelligent systems*. Morgan Kaufmann San Mateo, CA, 1988.

[Riggelsen, 2008] C. Riggelsen. Learning Bayesian networks: a MAP criterion for joint selection of model structure and parameter. In *Preceedings of the 8th IEEE International Conference on Data Mining*, pages 522–529. IEEE, 2008.

[Rissanen, 1978] J. Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

[Schwarz, 1978] G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[Silander *et al.*, 2007] T. Silander, P. Kontkanen, and P. Myllymäki. On sensitivity of the MAP Bayesian network structure to the equivalent sample size parameter. In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence*, pages 360–367. AUAI Press, 2007.

[Silander *et al.*, 2008] T. Silander, T. Roos, P. Kontkanen, and P. Myllymäki. Factorized normalized maximum likelihood criterion for learning Bayesian network structures. In *Proceedings of the 4th European Workshop on Probabilistic Graphical Models*, 2008.

[Spiegelhalter *et al.*, 1993] D. J. Spiegelhalter, A. P. Dawid, S. L. Lauritzen, and R. G. Cowell. Bayesian analysis in expert systems. *Statistical science*, pages 219–247, 1993.

[Steck and Jaakkola, 2003] H. Steck and T. S. Jaakkola. On the Dirichlet prior and Bayesian regularization. In *Advances in Neural Information Processing Systems*, pages 713–720, 2003.

[Steck, 2008] H. Steck. Learning the Bayesian network structure: dirichlet prior versus data. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, pages 511–518. AUAI Press, 2008.

[Wallace and Boulton, 1968] C. S. Wallace and D. M. Boulton. An information measure for classification. *The Computer Journal*, 11(2):185–194, 1968.

[Wallace *et al.*, 1996] C. Wallace, K. B Korb, and H. Dai. Causal discovery via MML. In *ICML*, volume 96, pages 516–524, 1996.