Pp. 231–236 in *Proceedings of Uncertainty 99: The Seventh International Workshop on Artificial Intelligence and Statistics*, edited by D. Heckerman and J. Whittaker. Morgan Kaufmann Publishers, 1999.

Exploring the Robustness of Bayesian and Information-Theoretic Methods for Predictive Inference

Petri Kontkanen, Petri Myllymäki, Tomi Silander, Henry Tirri, Kimmo Valtonen
Complex Systems Computation Group (CoSCo)
P.O.Box 26, Department of Computer Science, FIN-00014 University of Helsinki, Finland
cosco@cs.Helsinki.FI, http://www.cs.Helsinki.FI/research/cosco/

Abstract

Given a set of sample data, we study three alternative methods for determining the predictive distribution of an unseen data vector. In particular, we are interested in the behavior of the predictive accuracy of these three predictive methods as a function of the degree of the domain assumption violations. We explore this question empirically by using artificially generated data sets, where the assumptions can be violated in various ways. Our empirical results suggest that if the model assumptions are only mildly violated, marginalization over the model parameters may not be necessary in practice. This is due to the fact that in this case the computationally much simpler predictive distribution based on a single, maximum posterior probability model shows similar performance as the computationally more demanding marginal likelihood approach. The results also give support to Rissanen's theoretical results about the usefulness of using Jeffreys' prior distribution for the model parameters.

1 Introduction

In this paper we study discrete predictive inference tasks where the goal is to estimate the predictive distribution of a finite number of possible future events. As the sum of the estimated probabilities over all the possible events has to be one, it is clear that no fixed predictive distribution can be consistently "better" than some other probability distribution in the sense that it would give a higher probability for all the possible events than the other probability distribution. Intuitively, a good predictor is such that it gives a high probability to common events, i.e., to events that are likely to occur in the future, and a low probability to rare events.

In the following, we assume that the "commonness" of an event has to be estimated by using a given sample of domain data, with no prior knowledge about the problem domain. The possible events are vectors consisting of values of a set of discrete random variables, and we fix a set of assumptions that determine a parametric model form so that each model (parameter instantiation) produces a joint probability distribution over the discrete random variables. We then consider three alternative approaches for producing the predictive distribution. The first predictive distribution is obtained by using the single *maximum posterior probability (MAP)* model from the parametric family, where the posterior probability over different models is conditioned with respect to the given sample data. The *evidence predictive distribution (EV)* is obtained by computing the marginal likelihood (also known as the evidence) over all the possible models in the model family. The third alternative is motivated by Rissanen's information-theoretic considerations (Rissanen, 1989; Rissanen, 1996). A more detailed description of the three predictive distributions can be found in Section 2.

It can be argued that, theoretically, the predictive distribution based on a single MAP model should generally perform worse than the marginalized EV predictive distribution since the MAP approach is too "eager" in estimating the probabilities of future events: when given a limited amount of data, the resulting MAP model is too sensitive with respect to the sampling bias in the finite set of data, and hence can produce poor estimates of the probabilities of future events. Consequently, the marginalized predictive distribution should be more accurate, as by integrating over all the possible models it automatically filters down the effect of the sampling bias. However, computing the marginalized predictive distribution can be very difficult in practice, and we very often have to accept

the more straightforward MAP approach, although we know that the predictions obtained may not be optimal. An interesting question is whether there are cases where the difference between the MAP approach and the marginalized approach is so small that it is pragmatically sensible to use the computationally much simpler MAP predictive distribution. One of the main goals of this paper is to empirically study the hypothesis that the difference between the MAP and the EV approach is smallest when our domain assumptions are reasonable, and largest when the assumptions do not hold.

To determine the posterior distribution required in the MAP and EV approaches, we need a prior distribution over the different models (parameter instantiations). In this paper we assume that we have no prior knowledge about the problem domain, so it is natural to use the non-informative uniform prior for the parameter values. However, this means that we indeed assume all the models to be equally probable a priori, even those models that produce uniformly distributed, highly irregular data. As our goal is to predict the future based on the regularities found in the sample data, it seems reasonable to assume that such regularities exist in the data in the first place, or otherwise we can give up the prediction problem as a hopeless task. Nevertheless, this means that we can obtain better predictors by assigning a higher prior to those models that produce more regular data. In (Rissanen, 1996), this intuitively appealing line of reasoning was given a firm theoretical foundation by proving certain elegant asymptotic properties of the EV approach when a specific prior, *Jeffreys' prior* (Jeffreys, 1939; Berger, 1985), is used. It is interesting to note that while Jeffreys' prior was originally derived by invariance arguments, Rissanen's approach gives a fundamentally different motivation for its use, based on information-theoretic considerations. One should observe that although in the general case the assumption of the existence of an underlying "true" model can be seriously questioned (see for example (Rissanen, 1989)), in our empirical setting the use of synthetic generated data allows us to identify the actual data generating model.

Compared to our earlier work on this area (Grünwald et al., 1998a; Kontkanen et al., 1997), the results presented in this paper differ in the following three ways. First of all, the main objective of this paper was to study the robustness of the different predictive distributions in order to see how their performance changes as a function of the degree of the domain assumption violations, while our previous work concentrated on studying the effects of the amount of the training data available. Secondly, for being able to perform this type of experiments, we had to be able to have full control over the data used. For this reason, instead of the real-world datasets used in our earlier work, we generated artificial data for our experiments. Thirdly, in this paper we focus on the joint predictive inference task, while our earlier work was very much concentrated on simple classification problems. Details of the empirical setup can be found in Section 3, and the results are reported in Section 4.

2 Three predictive distributions

Let us model the problem domain by a set $\mathcal{X}$ of m discrete random variables, $\mathcal{X} = \{X_1, \ldots, X_m\}$, where a random variable X_i can take on any of the values in the set $\mathbf{X}_i = \{x_{i1}, \ldots, x_{in_i}\}$. A *data instantiation* $\vec{d} = (x_1, \ldots, x_m)$ is a vector in which all the variables X_i have been assigned a value. A *random sample* $D = (\vec{d}_1, \ldots, \vec{d}_N)$ is a set of N i.i.d. (independent and identically distributed) data instantiations, where each $\vec{d}_j$ is assumed to be sampled from a joint distribution of the variables in $\mathcal{X}$.

Given a random sample D, we are interested in the question of how to define the *predictive distribution* $P(\vec{d}|D)$ for an unseen vector $\vec{d}$. We investigate several candidates for $P(\vec{d}|D)$, relative to a parametric family $\mathcal{M}$ of probabilistic models: each model $\Theta \in \mathcal{M}$ defines a probability $P(\vec{d}|\Theta)$ for each data instantiation $\vec{d}$, and, under the i.i.d. assumption, a probability $P(D|\Theta)$ (the *likelihood*) for each dataset D. Given the likelihood, and a prior distribution $P(\Theta)$ for all $\Theta \in \mathcal{M}$, we can arrive at a posterior distribution for the models:

$$P(\Theta|D) \propto P(D|\Theta)P(\Theta). \tag{1}$$

The *MAP (maximum posterior probability)* predictive distribution is given by

$$\begin{aligned} \mathcal{P}_{\text{map}}(\vec{d} \mid D, \Phi) &= P(\vec{d} \mid D, \hat{\Theta}_\Phi(D)) \\ &= P(\vec{d} \mid \hat{\Theta}_\Phi(D)), \end{aligned} \tag{2}$$

where the last equality follows from the i.i.d. assumption, Φ denotes the (hyper)parameters used for defining the prior distribution $P(\Theta)$, and $\hat{\Theta}(D)$ is the MAP model maximizing the posterior (1).

A more sophisticated approach is to average (integrate) over all the models $\Theta \in \mathcal{M}$, which produces the *evidence* or *marginal likelihood* predictive distribution

$$\begin{aligned} \mathcal{P}_{\text{ev}}(\vec{d}|D, \Phi) &= \int P(\vec{d}|D, \Theta, \Phi) P(\Theta|D, \Phi) d\Theta \\ &= \int P(\vec{d}|\Theta) P(\Theta|D, \Phi) d\Theta. \end{aligned} \tag{3}$$

Both the MAP predictive distribution and the EV predictive distribution are defined by using the posterior

$P(\Theta|D)$, which depends on the prior $P(\Theta)$. In the following we use $\mathcal{P}_{\text{mapu}}$ and $\mathcal{P}_{\text{evu}}$ for denoting the special cases when the prior distribution is uniform. As one can see from (1), in this case the MAP predictive distribution becomes the *Maximum Likelihood (ML) model* of classical statistics, i.e., the model $\tilde{\Theta}$ maximizing the data likelihood $P(D|\Theta)$.

Recently, it has been shown (Rissanen, 1996) that there exists a *stochastic complexity* code that is itself not dependent on any prior distribution of parameters, and which yields in a certain interesting sense shorter codelengths than the code based on the marginal likelihood with lengths $-\log \mathcal{P}_{\text{ev}}(D)$. Because of the strong connection between codes and probability distributions (Cover and Thomas, 1991), it is an interesting question whether the stochastic complexity code can be used to produce an accurate predictive distribution. However, although this can be easily done in principle (as demonstrated in, e.g., (Kontkanen et al., 1997)), a straightforward application of this approach for predictive inference is problematic (Grünwald, 1998; Grünwald et al., 1998b). Nevertheless, Rissanen also showed that the stochastic complexity measure can be estimated (asymptotically) by using the marginal predictive distribution (3) with a specific prior, Jeffreys' prior $\pi(\Theta)$ (Jeffreys, 1939),

$$\pi(\Theta) \propto |I(\Theta)|^{1/2}, \tag{4}$$

where $|I(\Theta)|$ denotes the determinant of the *Fisher information matrix* $I(\Theta)$ as defined in (Berger, 1985). The corresponding predictive distribution will in the sequel be denoted by $\mathcal{P}_{\text{evj}}$.

3 Empirical setup

In the experiments reported in this paper, the model family $\mathcal{M}$ was taken to be the family of *Bayesian networks* (see, e.g., (Pearl, 1988)). A Bayesian network is an acyclic directed graph, where the nodes correspond to the domain variables $X_1, \ldots, X_m$. Each network topology defines a set of independence assumptions which allow the joint probability distribution for a data vector $\vec{d}$ to be written as a product of simple conditional probabilities,

$$\begin{aligned} P(\vec{d}) &= P(X_1 = x_1, \ldots, X_m = x_m) \\ &= \prod_{i=1}^{m} P(X_i = x_i | \text{pa}_i = q_i), \end{aligned} \tag{5}$$

where q_i denotes a configuration of (the values of) the parents of variable X_i. Consequently, in the Bayesian network model family, a distribution $P(\vec{d} \mid \Theta)$ is uniquely determined by fixing the values of the parameters $\Theta = (\theta^1, \ldots, \theta^m)$, where $\theta^i = (\theta^i_{11}, \ldots, \theta^i_{1n_i}, \ldots, \theta^i_{c_i 1}, \ldots, \theta^i_{c_i n_i})$, n_i is the number of values of X_i, c_i is the number of configurations of pa_i, and $\theta^i_{q_i x_i} := P(X_i = x_i \mid \text{pa}_i = q_i)$.

In the following all the conditional distributions of the variables, given their parents, are assumed to be multinomial: $X_{i|q_i} \sim \text{Multi}(1; \theta^i_{q_i 1}, \ldots, \theta^i_{q_i n_i})$. It is relatively easy to see that the uniform prior is in this case of the Dirichlet distribution form (see, e.g., (Heckerman et al., 1995)), which is the conjugate distribution of the multinomial. In the following we assume that the structure of the Bayesian network is a simple tree, where one of the variables forms the root, and the other variables are the leaves. In this *Naive Bayes* case Jeffreys' prior is of the conjugate form as well (Kontkanen et al., 1998).

When the prior is of the Dirichlet form, the three predictive distributions described in Section 2 can be computed by using the results presented in (Cooper and Herskovits, 1992; Heckerman et al., 1995), as demonstrated in (Kontkanen et al., 1997). However, these results are based on several assumptions, of which we in the sequel focus on the following two:

Assumption 1 (Parameter independence). The model parameters are independent:

$$P(\Theta) = \prod_{i=1}^{m} P(\theta^i), P(\theta^i) = \prod_{j=1}^{n_m} P(\theta^i_{j1}, \ldots, \theta^i_{jn_i}).$$

Assumption 2 (Variable independence). The leave variables $X_1, \ldots, X_{m-1}$ are independent, given the value of the root variable X_m:

$$P(X_1, \ldots, X_{m-1} | X_m) = \prod_{i=1}^{m-1} P(X_i | X_m).$$

As discussed in the Introduction, our goal was to study the performance of the different predictive distributions as a function of the degree these assumptions were violated. There are of course several different ways to do this. In the experiments used in this paper, Assumption 1 was violated by generating data by using a single naive Bayes model, where some of the conditional distributions $P(X_i|X_m = x_{mj})$ and $P(X_i|X_m = x_{mk})$ were set to be identical. In this setting, it is not reasonable to assume that all the parameters are independent. Assumption 2 was violated by generating data from a mixture of several Naive Bayes models. In addition to this, as discussed in the Introduction, we also wished to compare the performance of the EVU and EVJ approaches as the function of the "skewness" of the data generating model Θ. This was obtained by creating models with different levels of entropy.

Originally, our intention was to explore independently how the different aspects mentioned above affect the predictive performance. However, in the experiments it was soon discovered that the effects of these aspects are so intertwined that performing this type of a study is extremely difficult in practise. For this reason, we measured the domain assumption violation level indirectly by measuring the general level of predictive accuracy instead. The idea is that the more the domain assumptions are being violated, the worse results our predictive distributions should give, and vice versa. The results can be found in the next Section.

4 Results

In our experiments we used the following three predictive distributions described in Section 2:

- The maximum likelihood predictive distribution $\mathcal{P}_{\text{mapu}}$ computed using formula (2) with the uniform prior.
- The marginal likelihood predictive distribution $\mathcal{P}_{\text{evu}}$ computed using formula (3) with the uniform prior.
- The (approximative) stochastic complexity predictive distribution $\mathcal{P}_{\text{evj}}$ computed using formula (3) with Jeffreys' prior (4).

The accuracy of each predictive method was measured by

$$\frac{1}{T}\sum_{t=1}^{T} -\log P(\vec{d}_t \mid D),$$

the average of minus the predictive log-likelihood of a previously unseen test vector $\vec{d}$, given a sample D of training data. Consequently, in this set of experiments, the different test vectors were processed independently of each other.

Both the test vector $\vec{d}$ and the training vectors D were i.i.d. samples from a probability distribution violating Assumptions 1 and 2 as described in Section 3. Figure 1 compares the results obtained using the MAPU and EVU approaches. A similar comparison between the results obtained by the EVU and EVJ approaches is given in Figure 2.

In Figures 1 and 2, each data point corresponds to an average of $T = 100$ independent tests. As it is well known that all three predictive distributions converge asymptotically as the size of the training set D increases, we concentrated in our tests on the small sample size behavior of the different methods. In the experiments shown in Figures 1 and 2, the size of the training set D was 30. In these tests, the 210 data points shown (corresponding to 210 training set–test set pairs) were generated by using 10 different data generating models, each violating the model assumptions in a different way.

The results plotted in Figure 1 show that the difference between the MAPU and EVU approaches is highest on the right when the domain assumptions are being violated severely (the minus predictive log-likelihood is on the average high), and decreases when we move to the left and the assumptions hold better and better. This suggests that the marginalization of the model parameters helps when the domain assumptions are not reasonable, but may not be beneficial at all (with respect to the MAP approach) if the domain assumptions hold well.

In Figure 2, we see empirical evidence for the theoretical arguments concerning the use of Jeffreys' prior: when the data is highly irregular (and hence the results are generally bad on the average), EVU produces better results than EVJ, whereas the situation reverses moving towards left in the plot, when the data becomes more and more regular.

5 Conclusion and Future Work

We have studied empirically the robustness of different predictive distributions by using artificial data sets generated by models violating our domain assumptions with a varying degree. Theoretically, it is well known that marginalization over model parameters filters down the bias caused by sampling error, and hence should lead to better predictive accuracy than predictive distributions based on a single parameter instantiation. In the light of this observation, it seems natural to assume that marginalization also helps in decreasing the error caused by adopting an unreasonable set of domain assumptions. In this paper, we study this hypothesis, and show empirically that marginalization over the model parameters seems to be most beneficial in cases where the domain assumptions do not hold, and may not be very useful if the assumptions are not very seriously violated.

Our second goal in this paper was to compare the predictive accuracy obtained by using the marginalized predictive distribution with uniform prior and with Jeffreys' prior. The use of Jeffreys' prior can be motivated by Rissanen's information-theoretic arguments, which state that theoretically, Jeffreys' prior should lead to better performance in cases where the data is highly skewed (i.e., regular), while it may not be the optimal approach when the data is non-skewed (i.e., irregular). To be able to study this question empirically, we generated artificial data so that we gained full control over the skewness of the data. The results

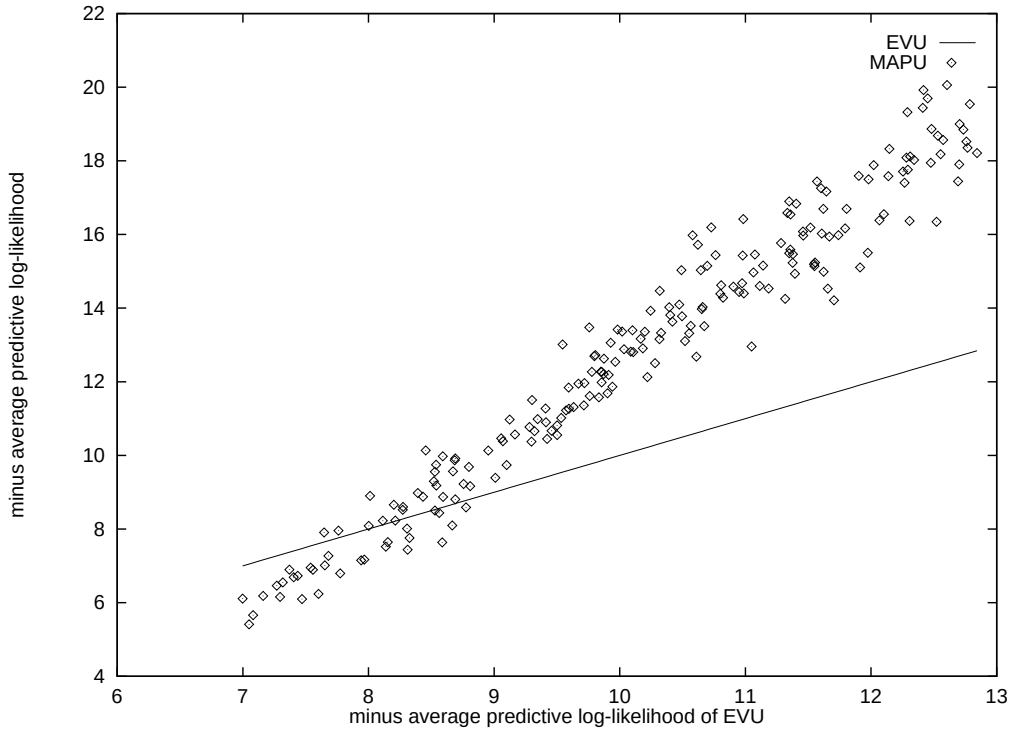

Figure 1: A comparison between the results obtained by the MAPU and EVU predictive distributions.

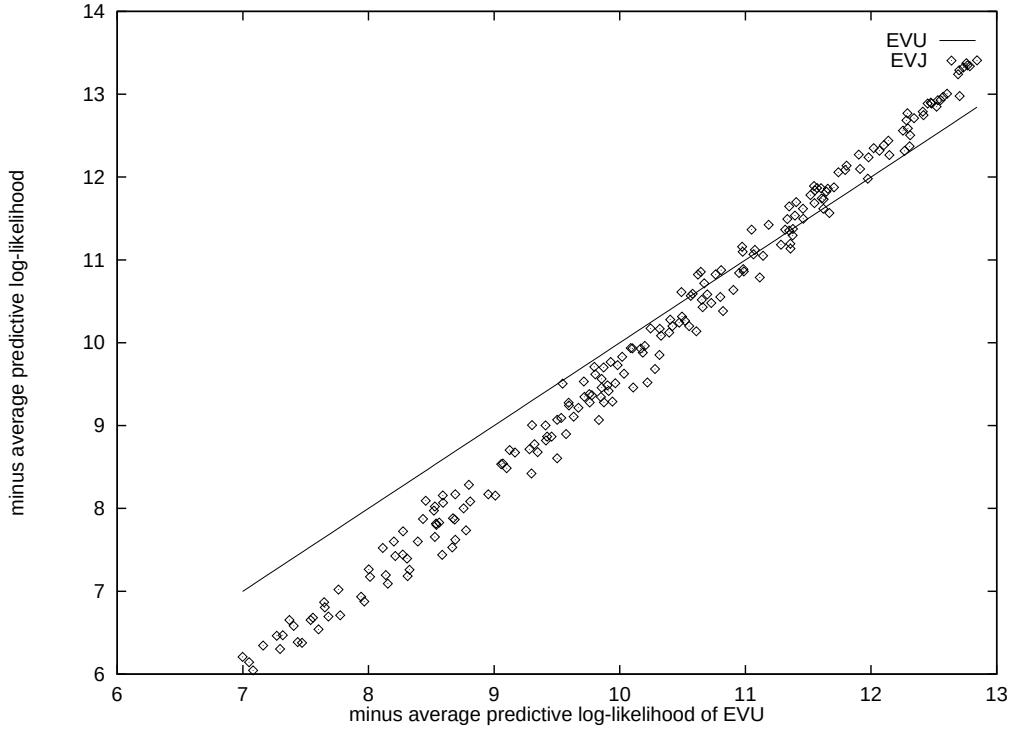

Figure 2: A comparison between the results obtained by the EVU and EVJ predictive distributions.

support Rissanen's hypothesis. Intuitively, this means that in cases where we obtain good predictive accuracy, we can still improve our results by a considerable amount by using Jeffreys' prior instead of the uniform prior. On the other hand, when the situation is such that with our current model structure it is not possible to gain a good predictive accuracy, then Jeffreys' prior may lead to slightly worse performance than the uniform prior. However, it can now be argued that in this case our domain assumptions obviously do not hold, and we should change our model structure accordingly.

It should be kept in mind that the results presented in this paper deal only with the non-informative case where no prior knowledge of the problem domain is available. If such information were available, the Bayesian way to incorporate such knowledge in making good predictors would be to use an appropriate prior distribution for the model parameters. In contrast to this, the discussion above suggests that the prior knowledge should be used for selecting the model structure only, while the parameters should be determined by using sample data and Jeffreys' prior. This naturally leads to the problem of model structure selection, which is a highly controversial issue in statistical inference, and not discussed here further.

As the work reported in this paper was concentrated on the computationally simple Naive Bayes model, it would be interesting to see whether the results extend to more complex domains, such as multi-connected Bayesian networks. However, according to our experience, this may be more difficult than one might expect: the different aspects studied in this report (different types of assumption violations, skewness of the data generating model) are obviously very severely intertwined with each other, and separating the effects of different aspects is very difficult. Development of a good experimental setup for studying these different aspects independently in more complex domains is a goal for our future work in this area.

Acknowledgments

This research has been supported by the Technology Development Center (TEKES) and the Academy of Finland.

References

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York.

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

Cover, T. and Thomas, J. (1991). *Elements of Information Theory*. John Wiley & Sons, New York, NY.

Grünwald, P. (1998). *The Minimum Description Length Principle and Reasoning under Uncertainty*. PhD thesis, CWI, ILLC Dissertation Series 1998-03.

Grünwald, P., Kontkanen, P., Myllymäki, P., Silander, T., and Tirri, H. (1998a). Minimum encoding approaches for predictive modeling. In Cooper, G. and Moral, S., editors, *Proceedings of the 14th International Conference on Uncertainty in Artificial Intelligence (UAI'98)*, Madison, WI. Morgan Kaufmann Publishers, San Francisco, CA.

Grünwald, P., Kontkanen, P., Myllymäki, P., Silander, T., and Tirri, H. (1998b). On predictive distributions and Bayesian networks. *Statistics and Computing (to appear)*.

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

Jeffreys, H. (1939). *Theory of Probability*. Clarendon Press, Oxford.

Kontkanen, P., Myllymäki, P., Silander, T., Tirri, H., and Grünwald, P. (1997). Comparing predictive inference methods for discrete domains. In *Proceedings of the Sixth International Workshop on Artificial Intelligence and Statistics*, pages 311–318, Ft. Lauderdale, Florida.

Kontkanen, P., Myllymäki, P., Silander, T., Tirri, H., and Grünwald, P. (1998). Bayesian and information-theoretic priors for Bayesian network parameters. In Nédellec, C. and Rouveirol, C., editors, *Machine Learning: ECML-98, Proceedings of the 10th European Conference*, Lecture Notes in Artificial Intelligence, Vol. 1398, pages 89–94. Springer-Verlag.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, San Mateo, CA.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry*. World Scientific Publishing Company, New Jersey.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, 42(1):40–47.