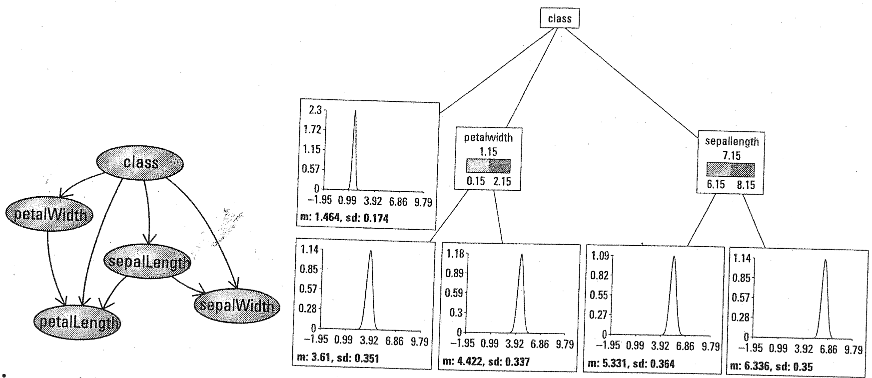

Figure 11.4 On the left is a Bayesian network learned from the iris data set using the MML approach presented in this chapter. On the right is the decision tree used to give a probability density over *petalLength*, given values of the parent attributes — *petalWidth*, *class*, and *sepalLength*.

- **MML-DT:** This is a decision tree tool that infers models from the class of decision trees described in Section 11.4.7. It uses an MML costing metric (see Section 11.4.7) similar to that in [Wallace and Patrick 1993] and a look-ahead-0 greedy search algorithm. This method is equivalent to a supervised network where all nontarget attributes are parents of the target attribute.
- **C5:** C5 [Quinlan] (and its forerunner, C4.5) are popular decision tree tools used for classification. C5 does not use the MML principle and is widely used as a performance benchmark in classification problems.
- **unsup-net:** This is the algorithm presented in this chapter for learning unsupervised asymmetric Bayesian networks.
- **sup-net:** This is the modified algorithm (see Section 11.4.11) that learns supervised asymmetric Bayesian networks.

The results in Figure 11.5 are from a series of ten-fold cross-validation experiments using the *iris* data set, available from [Blake and Merz 1998]. In all, 10 experiments were performed, for a total of 100 learning tasks for each method. In each experiment, each method's performance was averaged over the 10 test sets to yield a score, s. The graph shows the best, worst, and average values of s for each classifier. These results show the two MML asymmetric Bayesian network classifiers performing favorably, on average achieving a lower classification error than the decision trees. This is an example of a situation in which we do better by modeling the target attribute, implicitly using a joint distribution – rather than building an explicit conditional model like the two decision tree classifiers.

11.4.13 Issues for Further Research

The asymmetric Bayesian networks presented in this chapter have already produced encouraging results [Comley and Dowe 2003], and raise several interesting areas

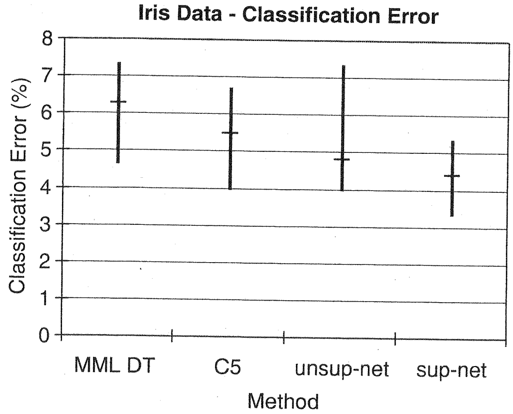

Figure 11.5 Best, average and worst performance of four classification tools on the *iris* data set.

for further research. We feel it would be beneficial to investigate other classes of asymmetric models, for example a multivariate version of the polynomial regression described in Section 11.4.6.

Another issue for future research relates to the estimation of Gaussian density functions in the leaves of decision trees modeling continuous attributes. The probability distribution for a discrete attribute tested by such a decision tree is partly determined by the ratio of these Gaussian distributions. When the estimated variance is small, this ratio can become very large and yield extreme probabilities for certain values of the discrete (target) attribute. This issue is discussed in more detail in [Comley and Dowe 2003]. In [Ng and Jordan 2002] the problem is avoided to some degree by fixing the variance at a value estimated from the entire training set, and allowing only the mean to vary as a function of the discrete target attribute. This seemingly has the effect of avoiding small variance estimates, and producing less dramatic ratios of Gaussian distributions.

Finally, we believe that the network structure coding scheme and search strategy presented in this chapter could be further refined, and have begun work on a promising variation based on incrementally adding directed links to an initially unordered, empty network.

11.5 Summary

This chapter has described minimum message length (MML) — a statistically invariant information-theoretic approach to Bayesian statistical inference dating back to Wallace and Boulton Wallace and Boulton [1968] — and highlighted some of the differences between MML and the subsequent minimum description length (MDL) principle. Furthermore, in Section 11.3, we have addressed several common misconceptions regarding MML, and (in Section 11.3.1) we mentioned Dowe's question as to whether Bayesianism is inherently necessary to guarantee statistical invariance and consistency.

This chapter has also presented an application of MML to a general class of Bayesian network that uses decision trees as conditional probability distributions. It can efficiently express context-specific independence, and is capable of modeling a combination of discrete and continuous attributes. We have suggested that when we know which attribute is to be predicted, it may be better to use a 'supervised' network rather than an 'unsupervised' one. We have proposed a modification to our algorithm to allow for this.

The main contribution here, other than extending Bayesian networks to handle continuous *and* discrete data, is the identification of 'asymmetric' networks, and the proposal of an efficient scheme to search for node order and connectivity.

11.6 Acknowledgments

We thank Peter Grünwald and the other editors both for the opportunity to write this chapter and for editorial feedback. David L. Dowe humbly thanks his dear mother, family and friends for their support for this work and his every endeavor.

References

Allison, L., C. S. Wallace, and C. Yee (1990). When is a string like a string? In *Proceedings of the International Symposium on Artificial Intelligence and Mathematics.*

Barron, A.R., and T.M. Cover (1991). Minimum complexity density estimation. *IEEE Transactions on Information Theory, 37*, 1034–1054.

Baxter, R.A., and D.L. Dowe (1996). Model selection in linear regression using the MML criterion. Technical report 96/276, Department of Computer Science, Monash University, Clayton, Victoria, Australia.

Bernardo, J., and A. Smith (1994). *Bayesian Theory.* New York: Wiley.

Blake, C., and C. Merz (1998). UCI repository of machine learning databases. Department of Information and Computer Sciences, University of California, Irvine. See also `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

Boulton, D. (1975). *The Information Criterion for Intrinsic Classification.* Ph.D. thesis, Department of Computer Science, Monash University, Clayton, Victoria, Australia.

Boulton, D.M., and C.S. Wallace (1969). The information content of a multistate distribution. *Journal of Theoretical Biology, 23*, 269–278.

Boulton, D.M., and C.S. Wallace (1970). A program for numerical classification.

Computer Journal, *13*(1), 63–69.

Boulton, D.M., and C.S. Wallace (1973). An information measure for hierarchic classification. *Computer Journal*, *16*(3), 254–261.

Boulton, D.M. and C.S. Wallace (1975). An information measure for single-link classification. *Computer Journal*, *18*(3), 236–238.

Boutilier, C., N. Friedman, M. Goldszmidt, and D. Koller (1996). Context-specific independence in Bayesian networks. In *Uncertainty in Artificial Intelligence: Proceedings of the Twelfth Conference (UAI-1996)*, pp. 115–123. San Francisco, CA :Morgan Kaufmann.

Chaitin, G.J. (1966). On the length of programs for computing finite sequences. *Journal of the Association for Computing Machinery*, *13*, 547–569.

Clarke, B. (1999). Discussion of the papers by Rissanen, and by Wallace and Dowe. *Computer Journal*, *42*(4), 338–339.

Comley, J.W. and D.L. Dowe (2003). General Bayesian networks and asymmetric languages. In *Proceedings of the Second Hawaiian International Conference on Statistics and Related Fields.*

Dawid, A.P. (1999). Discussion of the papers by Rissanen and by Wallace and Dowe. *Computer Journal*, *42*(4), 323–326.

Deakin, M.A.B. (2001). The characterisation of scoring functions. *Journal of the Australian Mathematical Society*, *71*, 135–147.

Dom, B. E. (1996). MDL estimation for small sample sizes and its application to linear regression. Technical report RJ 10030 (90526), IBM Almaden Research Center, San Jose, CA.

Dowe, D.L., L. Allison, T. Dix, L. Hunter, C. Wallace, and T. Edgoose (1996). Circular clustering of protein dihedral angles by minimum message length. In *Proceedings of the First Pacific Symposium on Biocomputing (PSB-1)*, Mauna Lani, HI, U.S.A., pp. 242–255. Singapore: World Scientific.

Dowe, D.L., R.A. Baxter, J.J. Oliver, and C.S. Wallace (1998). Point estimation using the Kullback-Leibler loss function and MML. In *Proceedings of the Second Pacific Asian Conference on Knowledge Discovery and Data Mining (PAKDD'98)*, Melbourne, Australia, pp. 87–95. Berlin: Springer Verlag.

Dowe, D.L., G.E. Farr, A. Hurst, and K.L. Lentin (1996). Information-theoretic football tipping. In N. de Mestre (Ed.), *Third Australian Conference on Mathematics and Computers in Sport*, Bond University, Queensland, Australia, pp. 233–241. See also http://www.csse.monash.edu.au/~footy .

Dowe, D.L., and A.R. Hajek (1998). A non-behavioural, computational extension to the Turing Test. In *Proceedings of the International Conference on Computational Intelligence and Multimedia Applications (ICCIMA'98)*, Gippsland, Australia, pp. 101–106.

Dowe, D.L., and K.B. Korb (1996). Conceptual difficulties with the efficient market hypothesis: Towards a naturalized economics. In D. Dowe, K. Korb, and

J. Oliver (Eds.), *Proceedigns of the Conference on Information, Statistics and Induction in Science (ISIS'96)*, Melbourne, Australia, pp. 212–223. Singapore: World Scientific.

Dowe, D.L. and N. Krusel (1993). A decision tree model of bushfire activity. Technical report 93/190, Department of Computer Science, Monash University, Clayton, Victoria 3800, Australia.

Dowe, D.L., J.J. Oliver, R.A. Baxter, and C.S. Wallace (1995). Bayesian Estimation of the von Mises concentration parameter. In *Proceedings of the Fifteenth International Workshop on Maximum Entropy and Bayesian Methods (Max-Ent '95)*, Santa Fe, NM. Boston: Kluwer.

Dowe, D.L., J.J. Oliver, T.I. Dix, L. Allison, and C.S. Wallace (1993). A decision graph explanation of protein secondary structure prediction. In *Proceedings of the 26th Hawaii International Conference on System Sciences (HICSS-26)*, Volume 1, Maui, HI, pp. 669–678. Los Alamitos, CA: IEEE Computer Society Press.

Dowe, D.L., J.J. Oliver, and C.S. Wallace (1996). MML estimation of the parameters of the spherical Fisher distribution. In *Proceedings of the Seventh International Workshop on Algorithmic Learning Theory (ALT'96)*, Sydney, Australia, pp. 213–227. Volume 1160 of *Lecture Notes in Artificial Intelligence (LNAI)*. Berlin: Springer Verlag.

Dowe, D.L., and G.R. Oppy (2001). Universal Bayesian inference? *Behavioral and Brain Sciences*, *24*(4), 662–663.

Dowe, D.L. and C.S. Wallace (1997). Resolving the Neyman-Scott problem by Minimum Message Length. In *Computing Science and Statistics — Proceedings of the 28th Symposium on the Interface*, Sydney, Australia, pp. 614–618.

Dowe, D.L., and C.S. Wallace (1998). Kolmogorov complexity, minimum message length and inverse learning. In *Proceedings of the Fourteenth Australian Statistical Conference (ASC-14)*, Gold Coast, Queensland, Australia, p. 144.

Edgoose, T., and L. Allison (1999). MML Markov classification of sequential data. *Statistics and Computing*, *9*(4), 269–278.

Edgoose, T., L. Allison, and D.L. Dowe (1996). An MML classification of protein structure that knows about angles and sequence. In *Proceedings of Third Pacific Symposium on Biocomputing (PSB-98)*, Mauna Lani, HI, pp. 585–596. Singapore: World Scientific.

Edwards, R., and D. Dowe (1998). Single factor analysis in MML mixture modeling. In *Proceedings of the Second Pacific Asian Conference on Knowledge Discovery and Data Mining (PAKDD'98)*, Melbourne, Australia, pp. 96–109. Berlin: Springer Verlag.

Farr, G.E., and C.S. Wallace (2002). The complexity of strict minimum message length inference. *Computer Journal*, *45*, 285–292.

Fitzgibbon, L., D. Dowe, and L. Allison (2002a). Change-point estimation using new minimum message length approximations. In *Proceedings of the Seventh Pacific Rim International Conference on Artificial Intelligence (PRICAI-2002)*, pp. 244–254. Volume 2417 of *Lecture Notes in Artificial Intelligence (LNAI)*. Berlin: Springer-Verlag.

Fitzgibbon, L., D. Dowe, and L. Allison (2002b). Univariate polynomial inference by Monte Carlo message length approximation. In *Proceedings of the 19th International Conference on Machine Learning (ICML-2002)*, pp. 147–154. San Francisco: Morgan Kaufmann.

Good, I.J. (1952). Rational decisions. *Journal of the Royal Statistical Society, Series B*, *14*, 107–114.

Good, I. J. (1968). Corroboration, explanation, evolving probability, simplicity, and a sharpened razor. *British Journal of Philosophy of Science*, *19*, 123–143.

Grünwald, P., P. Kontkanen, P. Myllymaki, T. Silander, and H. Tirri (1998). Minimum encoding approaches for predictive modeling. In *Proceedings of the Fourteenth International Conference on Uncertainty in Artificial Intelligence (UAI98)*, pp. 183–192.

Hernandez-Orallo, J., and N. Minaya-Collado (1998). A formal definition of intelligence based on an intensional variant of algorithmic complexity. In *Proceedings of the International Symposium on Engineering of Intelligent Systems (EIS'98)*, pp. 244–254.

Hodges, A. (1983). *Alan Turing : The Enigma*. New York: Simon & Schuster.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London A*, *186*, 453–454.

Kearns, M., Y. Mansour, A.Y. Ng, and D. Ron (1997). An experimental and theoretical comparison of model selection methods. *Machine Learning Journal*, *27*, 7–50.

Kissane, D., S. Bloch, D. Dowe, D.M.R.D. Snyder, P. Onghena, and C. Wallace (1996). The Melbourne family grief study, I: Perceptions of family functioning in bereavement. *American Journal of Psychiatry*, *153*, 650–658.

Kolmogorov, A.N. (1965). Three approaches to the quantitative definition of information. *Problems of Information Transmission*, *1*, 4–7.

Kornienko, L., D.L. Dowe, and D.W. Albrecht (2002). Message length formulation of support vector machines for binary classification - a preliminary scheme. In *Proceedings of the 15th Australian Joint Conference on Artificial Intelligence*, Canberra, Australia, 2-6 December 2002, pp. 119–130. Volume 2557 of *Lecture Notes in Artificial Intelligence (LNAI)*. Berlin: Springer Verlag, 2002.

Lanterman, A.D. (2005). Hypothesis testing for Poisson versus geometric distributions using stochastic complexity. In P.D. Grünwald, I.J. Myung, and M.A. Pitt (Eds.), *Advances in Minimum Description Length: Theory and Applications*. Cambridge MA: MIT Press, 2005.

Li, M. and P. Vitanyi (1997). *An Introduction to Kolmogorov Complexity and its*

Applications (2nd ed.). Springer-Verlag.

Liang, F. and Barron, A. (2005). Exact minimax predictive density estimation and MDL. In P. D. Grünwald, I. J. Myung, and M. A. Pitt (Eds.), *Advances in Minimum Description Length: Theory and Applications.* MIT Press, 2004.

Lindley, D. (1972). Bayesian statistics, a review. *SIAM*, 71.

McLachlan, G., and D. Peel (2000). Finite mixture models. *Wiley Series in Probability and Statistics.* New York: Wiley.

Murphy, P., and M. Pazzani (1994). Exploring the decision forest: An empirical investigation of Occam's razor in decision tree induction. *Journal of Artificial Intelligence, 1*, 257–275.

Needham, S.L. and D.L. Dowe (2001). Message length as an effective Ockham's razor in decision tree induction. In *Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics (AISTATS 2001)*, Key West, FL, pp. 253–260.

Ng, A.Y. and M.I. Jordan (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In T.G. Dietterich, S. Becker, and Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems 14.* Cambridge, MA: MIT Press.

Oliver, J.J. (1993). Decision graphs - an extension of decision trees. In *Proceedings of the Fourth International Workshop on Artificial Intelligence and Statistics*, pp. 343–350. Extended version available as technical report 173, Department of Computer Science, Monash University, Clayton, Victoria, Australia.

Oliver, J.J. and C.S. Wallace (1991). Inferring decision graphs. In *Proceedings of Workshop 8 — Evaluating and Changing Representation in Machine Learning IJCAI-91.*

Quinlan, J.R. C5.0. Available at `http://www.rulequest.com`.

Quinlan, J.R. and R.L. Rivest (1989). Inferring decision trees using the Minimum Description Length Principle. *Information and Computation, 80*(3), 227–248.

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

Rissanen, J.J. (1987). Stochastic complexity. *Journal of the Royal Statistical Society (Series B), 49*, 260–269.

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

Rissanen, J. J. (1996b). A universal regression model. In D. Dowe, K. Korb, and J. Oliver (Eds.), *Proceedings of the Conference on Information, Statistics and Induction in Science (ISIS'96)*, Melbourne, Australia, p. 4. Singapore: World Scientific.

Rissanen, J.J. (1999a). Discussion of paper "Minimum message length and Kolmogorov complexity" by C. S. Wallace and D. L. Dowe. *Computer Journal*, *42*, 327–329.

Rissanen, J.J. (1999b). Hypothesis selection and testing by the MDL principle. *Computer Journal 42*, 223–239.

Rissanen, J.J. (1999c). Rejoinder. *Computer Journal*, *42*, 343–344.

Russell, S., and P. Norvig (1995). *Artificial Intelligence: a Modern Approach.* Prentice Hall.

Sanghi, P. and D.L. Dowe (2003). A computer program capable of passing I.Q. tests. In *Proceedings of the Joint International Conference on Cognitive Science*, UNSW, Sydney, Australia.

Scheines, R., P. Spirtes, C. Glymour, and C. Meek (1994). *Tetrad II: User's Manual.* Hillsdale, NJ: Lawrence Erlbaum.

Shen, A. (1999). Discussion on Kolmogorov complexity and statistical analysis. *Computer Journal*, *42*(4), 340–342.

Solomonoff, R.J. (1964). A formal theory of inductive inference. *Information and Control 7*, 1–22,224–254.

Solomonoff, R.J. (1996). Does algorithmic probability solve the problem of induction? In D. Dowe, K. Korb, and J. Oliver (Eds.), *Proceedings of the Conference on Information, Statistics and Induction in Science (ISIS'96)*, Melbourne, Australia, pp. 7–8. Singapore: World Scientific.

Solomonoff, R.J. (1999). Two kinds of probabilistic induction. *Computer Journal*, *42*(4), 256–259.

Tan, P.J. and D.L. Dowe (2002). MML inference of decision graphs with multi-way joins. In *Proceedings of the Fifteenth Australian Joint Conference on Artificial Intelligence*, Canberra, Australia, pp. 131–142. Volume 2557 of *Lecture Notes in Artificial Intelligence (LNAI).* Berlin: Springer-Verlag.

Tan, P.J., and D.L. Dowe (2003, December). MML inference of decision graphs with multi-way joins and dynamic attributes In *Proceedings of the Sixteenth Australian Joint Conference on Artificial Intelligence (AI'03)*, Perth, Australia.

Vahid, F. (1999). Partial pooling: A possible answer to "To pool or not to pool". In R. Engle and H. White (Eds.), *Festschrift in Honor of Clive Granger*, pp. 410–428. Chapter 17. Oxford, UK: Oxford University Press.

Viswanathan, M., and C.S. Wallace (1999). A note on the comparison of polynomial selection methods. In *Proceedings of Uncertainty 99: the Seventh International Workshop on Artificial Intelligence and Statistics*, Fort Lauderdale, FL, pp. 169–177. San Francisco: Morgan Kaufmann.

Viswanathan, M., C.S. Wallace, D.L. Dowe, and K.B. Korb (1999). Finding cutpoints in noisy binary sequences - a revised empirical evaluation. In *Proceedings of the Twelfth Australian Joint Conference on Artificial Intelligence.* Volume 1747 of *Lecture Notes in Artificial Intelligence (LNAI)*, Sydney, Australia, pp. 405–416.

Vitányi, P., and M. Li (1996). Ideal MDL and its relation to Bayesianism. In D. Dowe, K. Korb, and J. Oliver (Eds.), *Proceedings of the Conference on Information, Statistics and Induction in Science (ISIS'96)*, Melbourne, Australia, pp. 405–416. Singapore: World Scientific.

Vitányi, P., and M. Li (2000). Minimum description length induction, Bayesianism, and Kolmogorov complexity. *IEEE Transactions on Information Theory, 46*(2), 446–464.

Vovk, V., and A. Gammerman (1999). Complexity approximation principle. *Computer Journal, 42*(4), 318–322. [special issue on Kolmogorov complexity].

Wallace, C.S. (1986). An improved program for classification. In *Proceedings of the Ninth Australian Computer Science Conference (ACSC-9)*, Volume 8, pp. 357–366.

Wallace, C.S. (1995). Multiple factor analysis by MML estimation. Technical report 95/218, Deptartment of Computer Science, Monash University, Clayton, Victoria, Australia.

Wallace, C.S. (1996). False oracles and SMML estimators. In D. Dowe, K. Korb, and J. Oliver (Eds.), *Proceedings of the Information, Statistics and Induction in Science (ISIS '96) Conference*, Melbourne, Australia, pp. 304–316. Singapore: World Scientific. Also Technical Rept 89/128, Department of Computer Science, Monash University, Clayton, Victoria, Australia, June 1989.

Wallace, C.S. (1997). On the selection of the order of a polynomial model. Technical report, Department of Computer Science, Royal Holloway College, London, England.

Wallace, C.S. (1998). Intrinsic classification of spatially correlated data. *Computer Journal, 41*(8), 602–611.

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

Wallace, C.S., and D.M. Boulton (1975). An invariant Bayes method for point estimation. *Classification Society Bulletin, 3*(3), 11–34.

Wallace, C.S., and D.L. Dowe (1993). MML estimation of the von Mises concentration parameter. Technical report 93/193, Department of Computer Science, Monash University, Clayton, Victoria, Australia.

Wallace, C.S., and D.L. Dowe (1994). Intrinsic classification by MML — the Snob program. In *Proceedings of the Seventh Australian Joint Conference on Artificial Intelligence*, University of New England, Armidale, Australia, pp. 37–44.

Wallace, C.S., and D.L. Dowe (1999a). Minimum message length and Kolmogorov complexity. *Computer Journal, 42*(4), 270–283. [special issue on Kolmogorov Complexity].

Wallace, C.S., and D.L. Dowe (1999b). Refinements of MDL and MML coding. *Computer Journal, 42*(4), 330–337. [special issue on Kolmogorov complexity].

Wallace, C.S., and D.L. Dowe (1999c). Rejoinder. *Computer Journal, 42*(4), 345–347.

Wallace, C.S., and D.L. Dowe (2000). MML clustering of multi-state, Poisson, von Mises circular and Gaussian distributions. *Statistics and Computing, 10*(1), 73–83.

Wallace, C.S., and P.R. Freeman (1987). Estimation and inference by compact coding. *Journal of the Royal Statistical Society Series B, 49*, 240–252.

Wallace, C.S., and P.R. Freeman (1992). Single factor analysis by MML estimation. *Journal of the Royal Statistical Society Series B, 54*(1), 195–209.

Wallace, C.S., and M.P. Georgeff (1983). A general objective for inductive inference. Technical report 83/32, Department of Computer Science, Monash University, Clayton, Victoria, Australia.

Wallace, C.S., and K.B. Korb (1999). Learning linear causal models by MML sampling. In A. Gammerman (Ed.), *Causal Models and Intelligent Data Management*, pp. 89–111. Berlin: Springer Verlag.

Wallace, C.S., and J.D. Patrick (1993). Coding decision trees. *Machine Learning, 11*, 7–22.

Wettig, H., P. Grünwald, T. Roos, P. Myllymäki, and H. Tirri (2003). When discriminative learning of Bayesian network parameters is easy. Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI 2003), Acapulco, Mexico, pp. 491–496.