Introduction to Minimum Length Encoding Inference

David L. Dowe
School of Computer Science and Software Engineering,
Monash University, Clayton, Vic. 3800, Australia

WWW: http://www.csse.monash.edu.au/∼dld/

July 13, 2002

1 Introduction - a prelude to the tutorial

1.1 A brief history of algorithmic info' theory, MDL, MML

The Minimum Message Length (MML) principle [47, 48, 54, 50] was initially developed in a series of at least 6 papers [47, sec. 2, p. 185][4][5, p. 64][6, 7, 48] by Chris Wallace and David Boulton from 1968 to 1975 and David Boulton's 1975 Ph.D. thesis [3], all beginning with the seminal (Wallace and Boulton, 1968)[47]. The (Wallace and Boulton, 1968)[47] work on mixture modelling - or clustering - of multinomial and Gaussian distributions using MML was the first application to statistics, econometrics and/or "data mining" of earlier independent work by Solomonoff [34] (see also [35, 36, 37]), Kolmogorov [22] and Chaitin [8] and subsequent work - on Minimum Description Length (MDL) - by Rissanen [28, 30, 29, 31, 33]. These information-theoretic approaches, MDL and MML, have been noted by various authors [28, p. 465][50, 37, 41, 42] to be related to Kolmogorov complexity (or algorithmic information theory).

1.2 Aesthetic properties - invariance, statistical consistency

Invariance is informally the property that a 1-to-1 re-parameterisation does not affect the answer. For example, if you estimate the length of a side of a cube and the volume of a cube, then your estimate of the volume will equal the cube of your estimate of the side length.

Consistency is informally the property that, as you get more and more data, then your estimate converges to any true underlying model with probability 1.

Various invariance results are given in [48, 50, 54] and various consistency results are given in [1, 45][54, p. 241]. Barron and Cover [1] also give results about *rates of convergence*.

Inconsistency can occur in problems where the amount of data per parameter is bounded above - e.g., Neyman-Scott problem [26, 15], fully-parameterised mixture modelling [53], (single [55] or multiple [44]) factor analysis.

1.2.1 Invariance and consistency conjecture

Dowe has conjectured [9, p. 93][50, p. 282][53, p. 78] that methods which are generally invariant and consistent must be Bayesian.

1.3 Is there a need for priors to be (Bayesian and) subjective?

The Jeffreys "prior" [20], when defined [51, Sec. 2.3.3], depends upon the sensitivity of the measuring instruments and the observational protocol used to obtain the data [23, 2][14, p. 217][51, Sec. 2.3] and even on the location of the measuring instrument [51, Sec. 2.3].

1.4 Approximations and efficiency of coding schemes

Although MML is consistent, the guaranteed rates of convergence [1] depend upon the efficiency of the coding scheme. See, e.g., [40, 21], as but one of many examples of an improvement occurring in MDL/MML results when the relevant coding scheme is made (even slightly) more efficient.

1.4.1 Improved approximations

As pointed out in [52, Sec. 2] and implicitly by the $o(log n)$ term in [32, eq. 8], the standardly used MML and MDL approximations are "only meant to be good for the exponential family and certain other families of functions"[52, Sec. 2]. Indeed, Grünwald et al. [19] give an example outside such families where coding the message in a certain order - not the order which C. S. Wallace would advocate - leads to problems with the (Wallace and Freeman, 1987) MML approximation [54][50, Sec. 6.1.2].

C. S. Wallace has published [55] and unpublished (in progress) improved approximations ([44], etc.); and Dowe with (in parts) L.J. Fitzgibbon has also developed an improved invariant approximation - see [17] (at ICML'2002) and [16] for some details.

1.5 Applications of MML

A general survey of both MML theory and MML applications is given in [50]. The seminal (Wallace and Boulton, 1968) mixture modelling program, Snob [47], has been

Table 1: Test Results on Artificial Data Sets

	Decision Graph	C5
1st Monks		
accuracy	100±0.0%	75.2±4.3%
pr costing	9.8±0.0bits	278.4±12.1bits
XD6		
accuracy	89.2±1.8%	85.2±2.5%
pr costing	25.2±2.4bits	30.2±3.3bits
LED-7		
accuracy	72.3±3.2%	71.2±3.5%
pr costing	74.5±7.4bits	89.7±7.9bits
LED-24		
accuracy	68.5±4.4%	67.8±4.7%
pr costing	96.4±6.2bits	102.8±7.3bits
Car		
accuracy	91.5±1.7%	91.6±1.9%
pr costing	44.8±5.6bits	70.9±6.8bits
Scale		
accuracy	79.2±3.2%	61.9±2.8%
pr costing	54.6±5.9bits	83.4±4.7bits

extended (see [53] and references). Notes on some other (MML) topics: MML parameter estimation [53, 49, 15]; market (in)efficiency [12]; Turing (imitation game) test [11]; MML Bayesian/causal nets [25]; Vapnik and Chervonenkis's VC dimension [38] and preliminary comparisons with MML [39, 17]; support of Ockham's razor via MML [24]. Other good recent work is on MML decision graphs/trees (Tan and Dowe, submitted, 2002)[1]. Table 1 (see, e.g., [18][13, sec. 4][10, 24] and references therein re probabilistic prediction costing) shows an excerpt comparing this Tan-Dowe scheme somewhat favourably to C5 [27].

1.6 Unanswered questions

1.6.1 Question 1: Boosting (and bagging) in a coding message

How, if at all, can boosting (and bagging) be put in an MDL/MML framework? We should be watchful of boosting - and possibly moreso of bagging - because of the harm it can do to probabilistic predictions by, e.g., over-fitting noise.

1.6.2 Question 2: Acknowledgement

Given [43, 46] and various MML works, why doesn't Chris Wallace or one of the other major contributors to a closely-related area get at least one of the Turing Award and/or the ACM SIGKDD Innovation Award?

[1] A special "thanks" to the anonymous ICML'2002 reviewers who rejected this cutting-edge work.

2 Tutorial outline

Further to Section 1, the tutorial[2] will be on Minimum Length Encoding, encompassing both Minimum Message Length (MML) and Minimum Description Length (MDL) inductive inference, topics central to the 1999 special issue of the Computer Journal on Kolmogorov complexity (vol. 42, no. 4, 1999) - see, e.g., [33, 50] and other articles in that issue ([37, 32, 51, 52], etc.). This information-theoretic approach bridges many fields, and is yielding state-of-the-art solutions to at least many problems in machine learning, statistics, econometrics and "data mining". It has applications right across the sciences.

This work is information-theoretic in nature, with a broad range of applications in machine learning, statistics, knowledge discovery and "data mining". We discuss statistical parameter estimation and mixture modelling (or clustering) - via the Snob program [47, 53] - of continuous, discrete and circular data. We also discuss learning decision trees and decision graphs, both with standard multinomial leaf distributions and with more complicated models. We further discuss MML solutions of either cut-point problems [40, 21] or polynomial regression[39]; and, if time permits, possibly Support Vector Machines (SVMs), causal networks and finite state machines (Hidden Markov Models, HMMs) or other problems. The target audience is academics, machine learning and "data mining" practitioners and consultants, and/or others with at least first year university education in at least one of mathematics, statistics, econometrics or electrical engineering. The audience will learn introductory fundamentals of MML inference, and the many state-of-the art success of MML in statistics, machine learning and hybrid problems. The audience will also see some applications of MML to real-world data.

3 Other

[47, 50, 53].

References

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

[2] J.M. Bernardo and A.F.M. Smith. *Bayesian Theory*. Wiley, New York, 1994.

[3] D.M. Boulton. *The Information Criterion for Intrinsic Classification*. PhD thesis, Dept. Computer Science, Monash University, Australia, 1975.

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

[5] D.M. Boulton and C.S. Wallace. A program for numerical classification. *Computer Journal*, 13:63–69, 1970.

[2] This might perhaps also be called "**An introduction to Minimum Message Length inference**".

[6] D.M. Boulton and C.S. Wallace. An information measure for hierarchic classification. *The Computer Journal*, 16:254–261, 1973.

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

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

[9] D.L. Dowe, R.A. Baxter, J.J. Oliver, and C.S. Wallace. Point Estimation using the Kullback-Leibler Loss Function and MML. In *Proc. 2nd Pacific Asian Conference on Knowledge Discovery and Data Mining (PAKDD'98)*, pages 87–95, Melbourne, Australia, April 1998. Springer Verlag.

[10] D.L. Dowe, G.E. Farr, A.J. Hurst, and K.L. Lentin. Information-theoretic football tipping. In N. de Mestre, editor, *Proc. 3rd Conference on Mathematics and Computers in Sport*, pages 233–241, Bond University, Qld., Australia, 1996.

[11] D.L. Dowe and A.R. Hájek. A non-behavioural, computational extension to the Turing Test. In *Proc. International Conference on Computational Intelligence and Multimedia Applications (ICCIMA'98)*, pages 101–106, 1998.

[12] D.L. Dowe and K.B. Korb. Conceptual difficulties with the Efficient Market Hypothesis: Towards a naturalized economics. In D.L. Dowe, K.B. Korb, and J.J. Oliver, editors, *Proceedings of the Information, Statistics and Induction in Science (ISIS) Conference*, pages 212–223, Melbourne, Australia, August 1996. World Scientific.

[13] D.L. Dowe and N. Krusel. A decision tree model of bushfire activity. Technical report TR 93/190, Dept. of Computer Science, Monash University, Clayton, Vic. 3800, Australia, September 1993.

[14] D.L. Dowe, J.J. Oliver, and C.S. Wallace. MML estimation of the parameters of the spherical Fisher distribution. In A. Sharma et al., editor, *Proc. 7th Conf. Algorithmic Learning Theory (ALT'96), LNAI 1160*, pages 213–227, Sydney, Australia, October 1996.

[15] D.L. Dowe and C.S. Wallace. Resolving the Neyman-Scott problem by Minimum Message Length. In *Proc. Computing Science and Statistics - 28th Symposium on the interface*, volume 28, pages 614–618, 1997.

[16] L. J. Fitzgibbon, D. L. Dowe, and L. Allison. Change-point estimation using new minimum message length approximations. In *Proceedings of the 7th Pacific Rim International Conference on Artificial Intelligence (PRICAI'2002)*, Lecture Notes in Artificial Intelligence (LNAI). Springer-Verlag, 2002.

[17] L. J. Fitzgibbon, D. L. Dowe, and L. Allison. Univariate polynomial inference by Monte Carlo message length approximation. In *Proceedings of the 19th International Conference on Machine Learning (ICML'2002)*, pages 147–154. Morgan Kaufmann, 2002.

[18] I.J. Good. Explicativity, corroboration, and the relative odds of hypotheses. In *Good thinking— The Foundations of Probability and its applications*. University of Minnesota Press, Minneapolis, MN, U.S.A., 1983.

[19] P. Grunwald, P. Kontkanen, P. Myllymaki, T. Silander, and H. Tirri. Minimum encoding approaches for predictive modeling. In *Proceedings of the 14th International Conference on Uncertainty in Artificial Intelligence (UAI'98)*, pages 183–192, 1998.

[20] H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proc. of the Royal Soc. of London A*, 186:453–454, 1946.

[21] M. Kearns, Y. Mansour, A. Y. Ng, and D. Ron. An experimental and theoretical comparison of model selection methods. *Machine Learning*, 27:7–50, 1997.

[22] A.N. Kolmogorov. Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1:4–7, 1965.

[23] D.V. Lindley. Bayesian statistics, a review. *SIAM*, page 71, 1972.

[24] S.L. Needham and D.L. Dowe. Message length as an effective Ockham's razor in decision tree induction. In *Proc. 8th Int. Workshop on Artif. Intelligence and Statistics (AI+STATS 2001)*, pages 253–260, 2001.

[25] J.R. Neil, C.S. Wallace, and K.B. Korb. Learning Bayesian networks with restricted causal interactions. In *Proc. 15th Conf. on Uncertainty in Artif. Intelligence (UAI-99)*, pages 486–493. Morgan Kaufmann, July-August 1999.

[26] J. Neyman and E.L. Scott. Consistent estimates based on partially consistent observations. *Econometrika*, 16:1–32, 1948.

[27] J.R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993. C5 is available from http://www.rulequest.com .

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

[29] J. J. Rissanen. A universal prior for integers and estimation by minimum description length. *Annals of Statistics*, 11:416–431, 1983.

[30] J. J. Rissanen. Stochastic complexity. *Journal of the Royal Statistical Society (Series B)*, 49:223–239, 1987.

[31] J. J. Rissanen. Fisher Information and Stochastic Complexity. *IEEE Trans. on Information Theory*, 42(1):40–47, January 1996.

[32] J. J. Rissanen. Discussion of paper 'Minimum Message Length and Kolmogorov Complexity' by C.S. Wallace and D. L. Dowe. *Computer Journal*, 42(4):327–329, 1999.

[33] J. J. Rissanen. Hypothesis selection and testing by the MDL principle. *Computer Journal*, 42(4):260–269, 1999.

[34] R.J. Solomonoff. A formal theory of inductive inference. *Information and Control*, 7:1–22,224–254, 1964.

[35] R.J. Solomonoff. Complexity based induction systems: Comparisons and convergence theorems. *IEEE Transaction on Information Theory*, IT-24:422–432, 1978.

[36] R.J. Solomonoff. The discovery of algorithmic probability: A guide for the programming of true creativity. In P. Vitanyi, editor, *Computational Learning Theory: EuroCOLT'95*, pages 1–22. Springer-Verlag, 1995.

[37] R.J. Solomonoff. Two kinds of probabilistic induction. *Computer Journal*, 42(4):256–259, 1999. Special issue on Kolmogorov Complexity.

[38] V.N. Vapnik. *The Nature of Statistical Learning Theory.* Springer, 1995.

[39] M. Viswanathan and C.S. Wallace. A note on the comparison of polynomial selection methods. In *Proc. 7th Int. Workshop on Artif. Intelligence and Statistics*, pages 169–177. Morgan Kaufmann, 1999.

[40] M. Viswanathan, C.S. Wallace, D.L. Dowe, and K.B. Korb. Finding cutpoints in noisy binary sequences - a revised empirical evaluation. In *Proc. 12th Australian Joint Conf. on Artif. Intelligence*, pages 405–416. Springer-Verlag, 1999.

[41] P.M.B. Vitányi and Ming Li. Ideal MDL and Its Relation to Bayesianism. In D.L. Dowe, K.B. Korb, and J.J. Oliver, editors, *Proc. Information, Statistics and Induction in Science (ISIS) Conference*, pages 282–291, Melbourne, 1996. World Scientific.

[42] P.M.B. Vitányi and Ming Li. Minimum description length induction, Bayesianism, and Kolmogorov Complexity. *IEEE Trans. Inform. Theory*, IT-46(2):343–384, 2000.

[43] C.S. Wallace. A suggestion for a fast multiplier. *IEEE Trans. on Electron. Comp.*, EC-13:14–17, Feb. 1964.

[44] C.S. Wallace. Multiple Factor Analysis by MML Estimation. Technical Report 95/218, Dept. of Computer Science, Monash University, Clayton, Victoria 3168, Australia, 1995. accepted, to appear in J. Multiv. Analysis.

[45] C.S. Wallace. False Oracles and SMML Estimators. In *Proc. Information, Statistics and Induction in Science conference (ISIS'96)*, pages 304–316, Singapore, 1996. World Scientific. Was Tech Rept TR 89/128, Monash University, Australia, 1989.

[46] C.S. Wallace. Fast pseudorandom generators for normal and exponential variates. *ACM Transactions on Mathematical Software*, pages 119–127, 1996.

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

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

[49] C.S. Wallace and D.L. Dowe. MML estimation of the von Mises concentration parameter. Tech Rept TR 93/193, Dept. of Comp. Sci., Monash Univ., Clayton 3168, Australia, 1993. prov. accepted, Aust. and N.Z. J. Stat.

[50] C.S. Wallace and D.L. Dowe. Minimum Message Length and Kolmogorov Complexity. *Computer Journal*, 42(4):270–283, 1999. Special issue on Kolmogorov Complexity; http://www3.oup.co.uk/computer_journal/hdb/Volume_42/Issue_04/pdf/420270.pdf.

[51] C.S. Wallace and D.L. Dowe. Refinements of MDL and MML coding (reply to J. J. Rissanen computer Journal article). *Computer Journal*, 42(4):330–337, 1999. Special issue on Kolmogorov Complexity; http://www3.oup.co.uk/computer_journal/hdb/Volume_42/Issue_04.

[52] C.S. Wallace and D.L. Dowe. Rejoinder. *Computer Journal*, 42(4):345–347, 1999. Special issue on Kolmogorov Complexity; http://www3.oup.co.uk/computer_journal/hdb/Volume_42/Issue_04.

[53] C.S. Wallace and D.L. Dowe. MML clustering of multi-state, Poisson, von Mises circular and Gaussian distributions. *Statistics and Computing*, 10:73–83, 2000. http://www.wkap.nl/issuetoc.htm/0960-3174+10+1+2000 http://www.wkap.nl/sampletoc.htm?0960-3174+10+1+2000 http://www.wkap.nl/jrnlsample.htm?0960-3174 http://www.wkap.nl/sample.pdf?251190.

[54] C.S. Wallace and P.R. Freeman. Estimation and inference by compact coding. *J. Royal Statistical Society (Series B)*, 49:240–252, 1987.

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