Minimum Description Length Model Selection

de Rooij, S.

Publication date
2008

Bibliography

[1] P. Adriaans and J. van Benthem (editors). *Handbook of the Philosophy of Information*. Elsevier, 2008.

[2] H. Akaike. A new look at statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[3] V. Balasubramanian. Statistical inference, Occam's Razor, and statistical mechanics on the space of probability distributions. *Neural Computation*, 9:349–368, 1997.

[4] A. Barron, J. Rissanen, and B. Yu. The Minimum Description Length principle in coding and modeling. *IEEE Transactions on Information Theory*, 44(6):2743–2760, 1998.

[5] A.R. Barron. *Logically Smooth Density Estimation*. PhD thesis, Dept. of Electrical Engineering, Stanford University, Stanford, CA, 1985.

[6] A.R. Barron. Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems. In *Bayesian Statistics 6*, pages 27–52. Oxford University Press, 1998.

[7] A.R. Barron. Personal communication, 2008.

[8] A.R. Barron and T.M. Cover. Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 37(4):1034–1054, 1991.

[9] A.R. Barron and C. Sheu. Approximation of density functions by sequences of exponential families. *Annals of Statistics*, 19(3):1347–1369, 1991.

[10] A.R. Barron, Y. Yang, and B. Yu. Asymptotically optimal function estimation by minimum complexity criteria. In *Proceedings of the 1994 International Symposium on Information Theory*, page 38, Trondheim, Norway, 1994.

[11] J. Berger. Personal communication, 2004.

[12] J.O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. Springer-Verlag, New York, revised and expanded second edition, 1985.

[13] J.O. Berger and L.R. Pericchi. Objective Bayesian methods for model selection: introduction and comparison. *Institute of Mathematical Statistics Lecture Notes*, (Monograph series) 38:135–207, 1997.

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

[15] O. Bousquet. A note on parameter tuning for on-line shifting algorithms. Technical report, Max Planck Institute for Biological Cybernetics, 2003.

[16] O. Bousquet and M.K. Warmuth. Tracking a small set of experts by mixing past posteriors. *Journal of Machine Learning Research*, 3:363–396, 2002.

[17] M. Burrows and D.J. Wheeler. A block-sorting lossless data compression algorithm. Technical Report 124, Digital Equipment Corporation, Systems Research Center, May 1994.

[18] G. Chang, B. Yu, and M. Vetterli. Adaptive wavelet thresholding for image denoising and compression. *IEEE Transactions on Image Processing*, 9:1532–1546, 2000.

[19] N. Chomsky. Three models for the description of language. *IEEE Transactions on Information Theory*, 2(3), September 1956.

[20] R. Cilibrasi and P. Vitányi. Algorithmic clustering of music based on string compression. *Computer Music Journal*, 28:49–67, 2004.

[21] B. Clarke. Online forecasting proposal. Technical report, University of Dortmund, 1997. Sonderforschungsbereich 475.

[22] B. Clarke and A. Barron. Jeffreys' prior is asymptotically least favourable under entropy risk. *The Journal of Statistical Planning and Inference*, 41:37–60, 1994.

[23] B.S. Clarke and A.R. Barron. Information-theoretic asymptotics of Bayes methods. *IEEE Transactions on Information Theory*, IT-36(3):453–471, 1990.

[24] J.G. Cleary and I.H. Witten. Data compression using adaptive coding and partial string matching. *IEEE Transactions on Communications*, COM-32(4):396–402, April 1984.

[25] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Series in telecommunications. John Wiley, 1991.

[26] A.P. Dawid. Present position and potential developments: Some personal views, statistical theory, the prequential approach. *Journal of the Royal Statistical Society, Series A*, 147(2):278–292, 1984.

[27] A.P. Dawid. Statistical theory: The prequential approach. *Journal of the Royal Statistical Society B*, 147, Part 2:278–292, 1984.

[28] A.P. Dawid. Prequential analysis, stochastic complexity and Bayesian inference. In J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, editors, *Bayesian Statistics 4*, pages 109–125. Oxford University Press, 1992.

[29] A.P. Dawid. Prequential data analysis. In M. Ghosh and P.K. Pathak, editors, *Current Issues in Statistical Inference: Essays in Honor of D. Basu*, Lecture Notes-Monograph Series, pages 113–126. Institute of Mathematical Statistics, 1992.

[30] X. De Luna and K. Skouras. Choosing a model selection strategy. *Scandinavian Journal of Statistics*, 30:113–128, 2003.

[31] P. Diaconis and D. Freedman. On the consistency of Bayes estimates. *The Annals of Statistics*, 14(1):1–26, 1986.

[32] P. Elias. Universal codeword sets and representations of the integers. *IEEE Transactions on Information Theory*, 21(2):194–203, 1975.

[33] M.R. Forster. The new science of simplicity. In A. Zellner, H. Keuzenkamp, and M. McAleer, editors, *Simplicity, Inference and Modelling*, pages 83–117. Cambridge University Press, Cambridge, 2001.

[34] D. Foster and E. George. The risk inflation criterion for multiple regression. *Annals of Statistics*, 22:1947–1975, 1994.

[35] L. Gerensce'r. Order estimation of stationary Gaussian ARMA processes using Rissanen's complexity. Technical report, Computer and Automation Institute of the Hungarian Academy of Sciences, 1987.

[36] P. Grünwald and S. de Rooij. Asymptotic log-loss of prequential maximum likelihood codes. In *Proceedings of the Eighteenth Annual Conference on Learning Theory (COLT 2005)*, pages 652–667. Springer, 2005.

[37] P. Grünwald and P. Vitányi. Shannon information and Kolmogorov complexity. Submitted to IEEE Transactions on Information Theory, available at www.cwi.nl/~paulv/publications.html, 2005.

[38] P.D. Grünwald. MDL tutorial. In P.D. Grünwald, I.J. Myung, and M.A. Pitt, editors, *Advances in Minimum Description Length: Theory and Application s*. MIT Press, 2005.

[39] P.D. Grünwald. *The Minimum Description Length Principle*. MIT Press, June 2007.

[40] M. Hansen and B. Yu. Minimum Description Length model selection criteria for generalized linear models. In *Science and Statistics: Festschrift for Terry Speed*, volume 40 of *IMS Lecture Notes – Monograph Series*. Institute for Mathematical Statistics, Hayward, CA, 2002.

[41] M.H. Hansen and B. Yu. Model selection and the principle of Minimum Description Length. *Journal of the American Statistical Association*, 96(454):746–774, 2001.

[42] J.A. Hartigan. *Bayes Theory*. Springer-Verlag, New York, 1983.

[43] D. Helmbold and M. Warmuth. On weak learning. *Journal of Computer and System Sciences*, 50:551–573, 1995.

[44] E.M. Hemerly and M.H.A. Davis. Strong consistency of the PLS criterion for order determination of autoregressive processes. *Ann. Statist.*, 17(2):941–946, 1989.

[45] M. Herbster and M.K. Warmuth. Tracking the best expert. In *Learning Theory: 12th Annual Conference on Learning Theory (COLT 1995)*, pages 286–294, 1995.

[46] M. Herbster and M.K. Warmuth. Tracking the best expert. *Machine Learning*, 32:151–178, 1998.

[47] H. Jeffreys. *Theory of Probability*. Oxford University Press, London, third edition, 1961.

[48] R. Kass and P. Vos. *Geometric Foundations of Asymptotic Inference*. Wiley, 1997.

[49] R.E. Kass and A.E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

[50] P. Kontkanen and P. Myllymäki. A linear-time algorithm for computing the multinomial stochastic complexity. *Information Processing Letters*, 103:227–233, September 2007.

[51] P. Kontkanen, P. Myllymäki, T. Silander, H. Tirri, and P.D. Grünwald. On predictive distributions and Bayesian networks. *Journal of Statistics and Computing*, 10:39–54, 2000.

[52] P. Kontkanen, P. Myllymäki, and H. Tirri. Comparing prequential model selection criteria in supervised learning of mixture models. In T. Jaakkola and T. Richardson, editors, *Proceedings of the Eighth International Conference on Articial Intelligence and Statistics*, pages 233–238. Morgan Kaufman, 2001.

[53] A.D. Lanterman. Hypothesis testing for Poisson versus geometric distributions using stochastic complexity. In Peter D. Grünwald, In Jae Myung, and Mark A. Pitt, editors, *Advances in Minimum Description Length: Theory and Applications*. MIT Press, 2005.

[54] V.I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

[55] J.Q. Li and A.R. Barron. Mixture density estimation. In *NIPS*, pages 279–285, 1999.

[56] K.C. Li. Asymptotic optimality of c_p , c_l, cross-validation and generalized cross-validation: Discrete index set. *Annals of Statistics*, 15:958–975, 1987.

[57] L. Li and B. Yu. Iterated logarithmic expansions of the pathwise code lengths for exponential families. *IEEE Transactions on Information Theory*, 46(7):2683–2689, 2000.

[58] F. Liang and A. Barron. Exact minimax predictive density estimation and MDL. In Peter D. Grünwald, In Jae Myung, and Mark A. Pitt, editors, *Advances in Minimum Description Length: Theory and Applications*. MIT Press, 2005.

[59] G. McLachlan and D. Peel. *Finite Mixture Models*. Wiley Series in Probability and Statistics, 2000.

[60] M.Li and P. Vitányi. *An Introduction to Kolmogorov Complexity and its Applications*. Springer-Verlag, New York, 2nd edition, 1997.

[61] D.S. Modha and E. Masry. Prequential and cross-validated regression estimation. *Machine Learning*, 33(1), 1998.

[62] A. Moffat. *Compression and Coding Algorithms*. Kluwer Academic Publishers, 2002.

[63] C. Monteleoni and T. Jaakkola. Online learning of non-stationary sequences. In *Advances in Neural Information Processing Systems*, volume 16, Cambridge, MA, 2004. MIT Press.

[64] K.R. Parthasarathy. *Probability Measures on Metric Spaces*. Probability and Mathematical Statistics. Academic Press, 1967.

[65] J. Poland and M. Hutter. Asymptotics of discrete MDL for online prediction. *IEEE Transactions on Information Theory*, 51(11):3780–3795, 2005.

[66] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, volume 77-2, pages 257–285, 1989.

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

[68] J. Rissanen. A universal prior for integers and estimation by Minimum Description Length. *Annals of Statistics*, 11:416–431, 1983.

[69] J. Rissanen. Universal coding, information, prediction, and estimation. *IEEE Transactions on Information Theory*, IT-30(4):629–636, 1984.

[70] J. Rissanen. A predictive least squares principle. *IMA Journal of Mathematical Control and Information*, 3:211–222, 1986.

[71] J. Rissanen. *Stochastic Complexity in Statistical Inquiry*, volume 15 of *Series in Computer Science*. World Scientific, 1989.

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

[73] J. Rissanen. MDL denoising. *IEEE Transactions on Information Theory*, 46(7):2537–2543, 2000.

[74] J. Rissanen, T.P. Speed, and B. Yu. Density estimation by stochastic complexity. *IEEE Transactions on Information Theory*, 38(2):315–323, 1992.

[75] G. Roelofs. *PNG – The Definitive Guide*. O'Reilly, 1999. Also available at www.faqs.org/docs/png.

[76] L.J. Savage. *The Foundations of Statistics*. Dover Publications, 1954.

[77] E.D. Scheirer. Structured audio, Kolmogorov complexity, and generalized audio coding. *IEEE Transactions on Speech and Audio Processing*, 9(8), november 2001.

[78] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

[79] C.E. Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–423, 623–656, 1948.

[80] R. Shibata. Asymptotic mean efficiency of a selection of regression variables. *Annals of the Institute of Statistical Mathematics*, 35:415–423, 1983.

[81] E. Sober. The contest between parsimony and likelihood. *Systematic Biology*, 4:644–653, 2004.

[82] bzip2 and zzip. www.bzip2.org, debin.net/zzip. Lossless compression software.

[83] ImageMagick and NetPBM. www.imagemagick.org, www.netpbm.sourceforge.net. Open source packages of graphics software.

[84] R.J. Solomonoff. A formal theory of inductive inference, part 1 and part 2. *Information and Control*, 7:1–22, 224–254, 1964.

[85] T. Speed and B. Yu. Model selection and prediction: Normal regression. *Annals of the Institute of Statistical Mathematics*, 45(1):35–54, 1993.

[86] M. Stone. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society B*, 39:44–47, 1977.

[87] J. Takeuchi. On minimax regret with respect to families of stationary stochastic processes (in Japanese). In *Proceedings of IBIS 2000*, pages 63–68, 2000.

[88] J. Takeuchi and A. Barron. Asymptotically minimax regret for exponential families. In *Proceedings of SITA 1997*, pages 665–668, 1997.

[89] J. Takeuchi and A.R. Barron. Asymptotically minimax regret by Bayes mixtures. In *Proceedings of the 1998 International Symposium on Information Theory (ISIT 98)*, 1998.

[90] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.

[91] J. Tromp. Binary lambda calculus and combinatory logic. www.cwi.nl/ tromp/cl/cl.html, 2007.

[92] N. Vereshchagin and P. Vitányi. Algorithmic rate-distortion theory. arxiv.org/abs/cs.IT/0411014, 2005.

[93] N.K. Vereshchagin and P.M.B. Vitányi. Kolmogorov's structure functions and model selection. *IEEE Transactions on Information Theory*, 50(12):3265–3290, 2004.

[94] P.A.J. Volf and F.M.J. Willems. Switching between two universal source coding algorithms. In *Proceedings of the Data Compression Conference, Snowbird, Utah*, pages 491–500, 1998.

[95] V. Vovk. Derandomizing stochastic prediction strategies. *Machine Learning*, 35:247–282, 1999.

[96] E.J. Wagenmakers, P.D. Grünwald, and M. Steyvers. Accumulative prediction error and the selection of time series models. *Journal of Mathematical Psychology*, 2006. (this issue).

[97] C.Z. Wei. On predictive least squares principles. *Annals of Statistics*, 20(1):1–42, 1990.

[98] P. Whittle. Bounds for the moments of linear and quadratic forms in independent variables. *Theory of Probability and its Applications*, V(3), 1960.

[99] H. Wong and B. Clarke. Improvement over Bayes prediction in small samples in the presence of model uncertainty. *The Canadian Journal of Statistics*, 32(3):269–283, 2004.

[100] Q. Xie and A. Barron. Asymptotic minimax regret for data compression, gambling and prediction. *IEEE Transactions on Information Theory*, 46(2):431–445, 2000.

[101] Y. Yang. Model selection for nonparametric regression. *Statistica Sinica*, 9:475–499, 1999.

[102] Y. Yang. Mixing strategies for density estimation. *Annals of Statistics*, 28(1):75–87, 2000.

[103] Y. Yang. Can the strengths of AIC and BIC be shared? *Biometrica*, 92(4):937–950, 2005.

[104] Y. Yang. Consistency of cross-validation for comparing regression procedures. Submitted, 2005.

[105] Y. Yang and A.R. Barron. An asymptotic property of model selection criteria. *IEEE Transactions on Information Theory*, 44:117–133, 1998.

[106] Y. Yang and A.R. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27:1564–1599, 1999.

[107] B. Yu. Lower bounds on expected redundancy for nonparametric classes. *IEEE Transactions on Information Theory*, 42(1):272–275, 1996.