References

Adriaans, P., and C. Jacobs (2006). Using MDL for grammar induction. In *Proceedings of the Eighth International Colloquium on Grammatical Inference (ICGI-2006)*. To appear.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki (Eds.), *Second International Symposium on Information Theory*, pp. 267–281.

Allen, T. V., and R. Greiner (2000). Model selection criteria for learning belief nets: An empirical comparison. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML-97)*.

Allen, T. V., O. Madani, and R. Greiner (2003). Comparing model selection criteria for belief networks. Submitted.

Anthony, M., and N. Biggs (1992). *Computational Learning Theory*. Cambridge, UK: Cambridge University Press.

Bain, L., and M. Engelhardt (1989). *Introduction to Probability and Mathematical Statistics*. Boston: PWS-Kent.

Balasubramanian, V. (1997). Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions. *Neural Computation 9*, 349–368.

Balasubramanian, V. (2005). MDL, Bayesian inference and the geometry of the space of probability distributions. In P. D. Grünwald, I. J. Myung, and M. A. Pitt (Eds.), *Advances in Minimum Description Length: Theory and Applications*. Cambridge, MA: MIT Press.

Barndorff-Nielsen, O. (1978). *Information and Exponential Families in Statistical Theory*. Chichester, UK: Wiley.

Barron, A. (1985). *Logically Smooth Density Estimation*. Ph. D. thesis, Department of Electrical Engineering, Stanford University, Stanford, CA.

Barron, A. (1986). Discussion on Diaconis and Freedman: the consistency of Bayes estimates. *Annals of Statistics 14*, 26–30.

Barron, A. (1990). Complexity regularization with application to artificial neural networks. In G. Roussas (Ed.), *Nonparametric Functional Estimation and Related Topics*, pp. 561–576. Dordrecht, the Netherlands: Kluwer Academic Publishers.

Barron, A. (1998). Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems. In A. D. J.M. Bernardo, J.O. Berger and A. Smith (Eds.), *Bayesian Statistics*, volume 6, pp. 27–52. Oxford: Oxford University Press.

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

Barron, A., and N. Hengartner (1998). Information theory and superefficiency. *Annals of Statistics 26*(5), 1800–1825.

Barron, A., J. Rissanen, and B. Yu (1998). The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory 44*(6), 2743–2760. Special Commemorative Issue: Information Theory: 1948-1998.

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

Barron, A., Y. Yang, and B. Yu (1994). Asymptotically optimal function estimation by minimum complexity criteria. In *Proceedings of the 1994 International Symposium on Information Theory*, pp. 38. Trondheim, Norway.

Bartlett, P., S. Boucheron, and G. Lugosi (2001). Model selection and error estimation. *Machine Learning 48*, 85–113.

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*, revised and expanded 2nd edition. Springer Series in Statistics. New York: Springer-Verlag.

Berger, J., and L. Pericchi (2001). Objective Bayesian methods for model selection: introduction and comparison. In P. Lahiri (Ed.), *Model Selection*, volume 38 of *IMS Lecture Notes – Monograph Series*, pp. 135–207. Beachwood, Ohio: Institute of Mathematical Statistics.

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

Blackwell, D., and L. Dubins (1962). Merging of opinions with increasing information. *Annals of Mathematical Statistics 33*, 882–886.

Blum, A., and J. Langford (2003). PAC-MDL bounds. In *Proceedings of the Sixteenth Conference on Learning Theory (COLT' 03)*, pp. 344–357.

Blumer, A., A. Ehrenfeucht, D. Haussler, and M. Warmuth (1987). Occam's razor. *Information Processing Letters 24*, 377–380.

Boucheron, S., O. Bousquet, and G. Lugosi (2005). Theory of classification: a survey of recent advances. *ESAIM: Probability and Statistics 9*, 323–375.

Breiman, L. (2001). Statistical modeling: the two cultures (with discussion). *Statistical Science 16*(3), 199 –215.

Brown, L. (1986). *Fundamentals of Statistical Exponential Families, with Applications in Statistical Decision Theory*. Hayward, CA: Institute of Mathematical Statistics.

Burnham, K., and D. Anderson (2002). *Model Selection and Multimodel Inference*. New York: Springer-Verlag.

Casella, G., and R. Berger. *Statistical Inference*. Belmont, CA: Wadsworth.

Cesa-Bianchi, N., Y. Freund, D. Helmbold, D. Haussler, R. Schapire, and M. Warmuth (1997). How to use expert advice. *Journal of the ACM 44*(3), 427–485.

Cesa-Bianchi, N., and G. Lugosi (2006). *Prediction, Learning and Games*. Cambridge, UK: Cambridge University Press.

Chaitin, G. (1966). On the length of programs for computing finite binary sequences. *Journal of the ACM 13*, 547–569.

Chaitin, G. (1969). On the length of programs for computing finite binary sequences: Statistical considerations. *Journal of the ACM 16*, 145–159.

Chernoff, H. (1952). A measure of asymptotic efficiency of test of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics 23*, 493–507.

Cilibrasi, R., and P. Vitányi (2005). Clustering by compression. *IEEE Transactions on Information Theory 51*(4), 1523–1545.

Clarke, B. (1997). Online forecasting proposal. Technical report, University of Dortmund. Sonderforschungsbereich 475.

Clarke, B. (2003). Combining model selection procedures for online prediction. *Sankhyā: The Indian Journal of Statistics, Series A 63*, 229–249.

Clarke, B. (2004). Comparing Bayes and non-Bayes model averaging when model approximation error cannot be ignored. *Journal of Machine Learning Research 4*(4), 683–712.

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

Clarke, B., and A. Barron (1994). Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference 41*, 37–60.

Clarke, B., and A. Dawid (1999). Online prediction with experts under a log-scoring rule. Unpublished manuscript.

Comley, J. W., and D. L. Dowe (2005). Minimum message length and generalized Bayesian nets with asymmetric languages. In P. D. Grünwald, I. J. Myung, and M. A. Pitt (Eds.), *Advances in Minimum Description Length: Theory and Applications*. Cambridge, MA: MIT Press.

Conway, J., and N. Sloane (1993). *Sphere Packings, Lattices and Groups*. New York: Springer-Verlag.

Cover, T., and J. Thomas (1991). *Elements of Information Theory*. New York: Wiley-Interscience.

Cramér, H. (1938). Sur un nouveau théorème-limite de la théorie des probabilités. *Actualités Scientifiques et Industrielles 736*, 5–23.

Csiszár, I. (1984). Sanov property, generalized I-projection and a conditional limit theorem. *Annals of Probability 12*(3), 768–793.

Csiszár, I., and P. Shields (2000). The consistency of the BIC Markov order estimator. *Annals of Statistics 28*, 1601–1619.

Davies, P., and A. Kovac (2001). Modality, runs, strings and multiresolution (with discussion). *Annals of Statistics 29*, 1–65.

Dawid, A. (1984). 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.

Dawid, A. (1992). Prequential analysis, stochastic complexity and Bayesian inference. In J. Bernardo, J. Berger, A. Dawid, and A. Smith

(Eds.), *Bayesian Statistics*, volume 4, pp. 109–125. Oxford: Oxford University Press.

Dawid, A. (1997). Prequential analysis. In S. Kotz, C. Read, and D. Banks (Eds.), *Encyclopedia of Statistical Sciences*, volume 1 (Update), pp. 464–470. New York: Wiley-Interscience.

Dawid, A. P., and V. G. Vovk (1999). Prequential probability: Principles and properties. *Bernoulli 5*, 125–162.

De Finetti, B. (1937). La prevision: ses lois logiques, ses sources subjectives. *Annales Institut H. Poincaré 7*, 1–68.

De Finetti, B. (1974). *Theory of Probability. A Critical Introductory Treatment*. London: Wiley.

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

De Rooij, S., and P. D. Grünwald (2006). An empirical study of MDL model selection with infinite parametric complexity. *Journal of Mathematical Psychology 50*(2), 180–192.

Devroye, L., L. Györfi, and G. Lugosi (1996). *A Probabilistic Theory of Pattern Recognition*. New York: Springer-Verlag.

Diaconis, P. (2003). The problem of thinking too much. *Bulletin of the American Academy of Arts and Sciences 16*(3), 26–38.

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

Domingos, P. (1999). The role of Occam's razor in knowledge discovery. *Data Mining and Knowledge Discovery 3*(4), 409–425.

Doob, J. (1949). Application of the theory of martingales. In *Le Calcul de Probabilités et ses Applications. Colloques Internationaux du Centre National de la Recherche Scientifique*, pp. 23–27.

Drmota, M., and W. Szpankowski (2004). Precise minimax redundancy and regret. *IEEE Transactions on Information Theory 50*, 2686–2707.

Duda, R., P. Hart, and D. Stork (2000). *Pattern Classification*. New York: Wiley.

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

Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *Quarterly Journal of Economics 75*, 643–649.

Feder, M. (1986). Maximum entropy as a special case of the minimum description length criterion. *IEEE Transactions on Information Theory 32*(6), 847–849.

Feller, W. (1968a). *An Introduction to Probability Theory and Its Applications*, 3rd edition, volume 1. New York: Wiley.

Feller, W. (1968b). *An Introduction to Probability Theory and Its Applications*, 3rd edition, volume 2. New York: Wiley.

Ferguson, T. (1967). *Mathematical Statistics – a decision-theoretic approach*. San Diego: Academic Press.

Figueiredo, M., J. Leitão, and A.K.Jain (2000). Unsupervised contour representation and estimation using b-splines and a minimum description length criterion. *IEEE Transactions on Image Processing 9*(6), 1075–1087.

Fisher, R. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A 222*, 309–368.

Floyd, S., and M. Warmuth (1995). Sample compression, learnability and the Vapnik-Chervonenkis dimension. *Machine Learning 21*, 269–304.

Forster, M. (2001). The new science of simplicity. In A. Zellner, H. Keuzenkamp, and M. McAleer (Eds.), *Simplicity, Inference and Modelling*, pp. 83–117. Cambridge: Cambridge University Press.

Foster, D., and R. Stine (1999). Local asymptotic coding and the minimum description length. *IEEE Transactions on Information Theory 45*, 1289–1293.

Foster, D., and R. Stine (2001). The competitive complexity ratio. In *Proceedings of the 2001 Conference on Information Sciences and Systems*. WP8 1-6.

Foster, D. P., and R. A. Stine (2005). The contribution of parameters to 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.

Freund, Y. (1996). Predicting a binary sequence almost as well as the optimal biased coin. In *Proceedings of the Ninth Annual ACM Conference on Computational Learning Theory (COLT' 96)*, pp. 89–98.

Friedman, N., D. Geiger, and M. Goldszmidt (1997). Bayesian network classifiers. *Machine Learning 29*, 131–163.

Gács, P., J. Tromp, and P. Vitányi (2001). Algorithmic statistics. *IEEE Transactions on Information Theory 47*(6), 2464–2479.

Gao, Q., and M. Li (1989). An application of minimum description length principle to online recognition of handprinted alphanumerals. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence (IJCAI-89)*, pp. 843–848.

Gauss, C. (1957). *Gauss's Work on the Theory of Least Squares (1803-1826)*. Princeton, NJ: Princeton University Press. Translated by H.F. Trotter.

Gelman, A., B. Carlin, H. Stern, and D. Rubin (2003). *Bayesian Data Analysis*. Boca Raton, FL: CRC Press.

George, E., and D. Foster (2000). Calibration and empirical Bayes variable selection. *Biometrika 84*(4), 731–747.

Gerenscér, L. (1987). Order estimation of stationary Gaussian ARMA processes using Rissanen's complexity. Technical report, Computer and Automation Institute of the Hungarian Academy of Sciences.

Gerencsér, L. (1994). On Rissanen's predictive stochastic complexity for stationary ARMA processes. *Journal of Statistical Planning and Inference 41*, 303–325.

Gibbs, A., and F. Su (2002). On choosing and bounding probability metrics. *International Statistical Review 70*(3), 419–435.

Goodman, N. (1955). *Fact, Fiction, and Forecast*. Cambridge, MA: Harvard University Press.

Grünwald, P. D. (1996). A minimum description length approach to grammar inference. In S. Wermter, E. Riloff, and G. Scheler (Eds.), *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, Number 1040 in Lecture Notes in Artificial Intelligence, pp. 203–216. New York: Springer-Verlag.

Grünwald, P. D. (1998). *The Minimum Description Length Principle and Reasoning under Uncertainty*. Ph. D. thesis, University of Amsterdam, the Netherlands. Available as ILLC Dissertation Series 1998-03.

Grünwald, P. D. (1999). Viewing all models as "probabilistic". In *Proceedings of the Twelfth ACM Conference on Computational Learning Theory (COLT' 99)*, pp. 171–182.

Grünwald, P. D. (2000). Maximum entropy and the glasses you are looking through. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence (UAI 2000)*, pp. 238–246.

Grünwald, P. D. (2001). Strong entropy concentration, game theory and algorithmic randomness. In *Proceedings of the Fourteenth Annual Conference on Computational Learning Theory (COLT' 01)*, pp. 320–336.

Grünwald, P. D. (2005). A tutorial introduction to the minimum description principle. In P. D. Grünwald, I. Myung, and M. Pitt (Eds.), *Advances in Minimum Description Length: Theory and Applications*, pp. 3–79. Cambridge, MA: MIT Press.

Grünwald, P. D. (2007). Prediction is coding. Manuscript in preparation.

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

Grünwald, P. D., and J. Y. Halpern (2003). Updating probabilities. *Journal of Artificial Intelligence Research 19*, 243–278.

Grünwald, P. D., and J. Langford (2004). Suboptimality of MDL and Bayes in classification under misspecification. In *Proceedings of the Seventeenth Conference on Learning Theory (COLT' 04)*.

Grünwald, P. D., and J. Langford (2007). Suboptimal behavior of Bayes and MDL in classification under misspecification. *Machine Learning*. To appear.

Grünwald, P. D., I. J. Myung, and M. A. Pitt (Eds.) (2005). *Advances in Minimum Description Length: Theory and Applications*. MIT Press.

Hall, P., and E. Hannan (1988). On stochastic complexity and nonparametric density estimation. *Biometrika 75*, 705–714.

Halpern, J. (2003). *Reasoning about Uncertainty*. Cambridge, MA: MIT Press.

Hannan, E. (1980). The estimation of the order of an ARMA process. *Annals of Statistics 8*, 1071–1081.

Hannan, E., A. McDougall, and D. Poskitt (1989). Recursive estimation of autoregressions. *Journal of the Royal Statistical Society, Series B 51*, 217–233.

Hannan, E., and J. Rissanen (1982). Recursive estimation of mixed autoregressive-moving average order. *Biometrika 69*, 81–94.

Hansen, M., and B. Yu (2000). Wavelet thresholding via MDL for natural images. *IEEE Transactions on Information Theory 46*, 1778–1788.

Hansen, M., and B. Yu (2001). Model selection and the principle of minimum description length. *Journal of the American Statistical Association 96*(454), 746–774.

Hansen, M., and B. Yu (2002). 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*. Hayward, CA: Institute for Mathematical Statistics.

Hanson, A. J., and P. C.-W. Fu (2005). Applications of MDL to selected families of models. In P. D. Grünwald, I. J. Myung, and M. A. Pitt (Eds.), *Advances in Minimum Description Length: Theory and Applications*. Cambridge, MA: MIT Press.

Harremoës, P. (2004). The weak information projection. In *Proceedings of the 2004 International Symposium on Information Theory (ISIT 2004)*, pp. 28.

Harremoës, P. (2006). Interpretations of Rényi entropies and divergences. *Physica A 365*(1), 57–62.

Harremoës, P., and F. Topsøe (2001). Maximum entropy fundamentals. *Entropy 3*, 191–226. Available at `http://www.mdpi.org/entropy/`.

Hartigan, J. (1983). *Bayes Theory*. New York: Springer-Verlag.

Haussler, D. (1997). A general minimax result for relative entropy. *IEEE Transactions on Information Theory 43*(4), 1276–1280.

Haussler, D., and M. Opper (1997). Mutual information, metric entropy, and cumulative relative entropy risk. *Annals of Statistics 25*(6), 2451–2492.

Helmbold, D., and M. Warmuth (1995). On weak learning. *Journal of Computer and System Sciences 50*, 551–573.

Hemerly, E., and M. Davis (1989a). Recursive order estimation of stochastic control systems. *Mathematical Systems Theory 22*, 323–346.

Hemerly, E., and M. Davis (1989b). Strong consistency of the PLS criterion for order determination of autoregressive processes. *Annals of Statistics 17*(2), 941–946.

Herbrich, R. (2002). *Learning Kernel Classifiers*. Cambridge, MA: MIT Press.

Herbrich, R., and R. C. Williamson (2002). Algorithmic luckiness. *Journal of Machine Learning Research 3*, 175–212.

Hertz, J., A. Krogh, and R. Palmer (1991). *Introduction to the theory of neural computation*. Lecture Notes of the Santa Fe Institute. Boston: Addison-Wesley.

Hjorth, U. (1982). Model selection and forward validation. *Scandinavian Journal of Statistics 9*, 95–105.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association 58*, 13–30.

Hoerl, A., and R. Kennard (1970). Ridge regression: Biased estimation of non-orthogonal components. *Technometrics 12*, 55–67.

Hutter, M. (2003). Optimality of universal Bayesian sequence prediction for general loss and alphabet. *Journal of Machine Learning Research 4*, 971–1000.

Hutter, M. (2004). *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Berlin: Springer-Verlag.

Hutter, M. (2006). On the foundations of universal sequence prediction. In *Proceedings of the Third Annual Conference on Theory and Applications of Models of Computation (TAMC 2006)*, pp. 408–420.

Jaynes, E. (1957). Information theory and statistical mechanics. *Physical Review 106*(4), 620–630.

Jaynes, E. (2003). *Probability Theory: The Logic of Science*. Cambridge, UK: Cambridge University Press. Edited by G. Larry Bretthorst.

Jeffereys, W., and J. Berger (1992). Ockham's razor and Bayesian analysis. *American Scientist 80*, 64–72.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Statistical Society (London) Series A 186*, 453–461.

Jeffreys, H. (1961). *Theory of Probability*, 3rd edition. London: Oxford University Press.

Jornsten, R., and B. Yu (2003). Simultaneous gene clustering and subset selection for classification via mdl. *Bioinformatics 19*(9), 1100–1109.

Kakade, S., M. Seeger, and D. Foster (2006). Worst-case bounds for Gaussian process models. In *Proceedings of the 2005 Neural Information Processing Systems Conference (NIPS 2005)*.

Kallenberg, O. (2002). *Foundations of Modern Probability*, 2nd edition. New York: Springer-Verlag.

Kalnishkan, Y., and M. Vyugin (2002). Mixability and the existence of weak complexities. In *Proceedings of the Fifteenth Conference on Computational Learning Theory (COLT' 02)*, pp. 105–120.

Kapur, J. N., and H. K. Kesavan (1992). *Entropy Optimization Principles with Applications*. San Diego: Academic Press.

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

Kass, R., and P. Voss (1997). *Geometrical Foundations of Asymptotic Inference*. New York: Wiley-Interscience.

Kass, R., and L. Wasserman (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association 91*, 1343–1370.

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

Kelly, J. (1956). A new interpretation of information rate. *Bell System Technical Journal*, 917–926.

Kolmogorov, A. (1941). Interpolation und Extrapolation von stationären zufälligen Folgen. *Isvestiia Akademii Nauk SSSR 5*, 3–14.

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

Kolmogorov, A. (1974a). Talk at the Information Theory Symposium in Tallinn, Estonia, 1974, according to P. Gács and T. Cover who attended it.

Kolmogorov, A. (1974b). Complexity of algorithms and objective definition of randomness. A talk at Moscow Mathematical Society meeting, April 16th, 1974. A 4-line abstract is available in *Uspekhi Matematicheskih Nauk* 29:4(1974), 155 (in Russian).

Kontkanen, P., W. Buntine, P. Myllymäki, J. Rissanen, and H. Tirri (2003). Efficient computation of stochastic complexity. In C. Bishop and B. Frey (Eds.), *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics (AISTATS 2003)*, pp. 181–188.

Kontkanen, P., and P. Myllymäki (2005a). Analyzing the stochastic complexity via tree polynomials. Unpublished manuscript.

Kontkanen, P., and P. Myllymäki (2005b). A fast normalized maximum likelihood algorithm for multinomial data. In *Proceedings of the Nine-*

teenth International Joint Conference on Artificial Intelligence (IJCAI-05), pp. 1613–1616.

Kontkanen, P., P. Myllymäki, W. Buntine, J. Rissanen, and H. Tirri (2005). An MDL framework for data clustering. In P. D. Grünwald, I. J. Myung, and M. A. Pitt (Eds.), *Advances in Minimum Description Length: Theory and Applications*. Cambridge, MA: MIT Press.

Kontkanen, P., P. Myllymäki, T. Silander, and H. Tirri (1999). On supervised selection of Bayesian networks. In K. Laskey and H. Prade (Eds.), *Proceedings of the Fifteenth International Conference on Uncertainty in Artificial Intelligence (UAI'99)*.

Kontkanen, P., P. Myllymäki, T. Silander, H. Tirri, and P. D. Grünwald (1998). Bayesian and information-theoretic priors for Bayesian network parameters. In C. Nedellec and C.Rouveirol (Eds.), *Machine Learning: ECML-98, Proceedings of the Tenth European Conference*, volume 1398 of *Lecture Notes in Artificial Intelligence*, pp. 89–94.

Kraft, L. (1949). A device for quantizing, grouping and coding amplitude modulated pulses. Master's thesis, Department of Electrical Engineering, MIT, Cambridge, MA.

Krichevsky, R., and V. Trofimov (1981). The performance of universal encoding. *IEEE Transactions on Information Theory 27*, 199–207.

Lai, T., and C. Lee (1997). Information and prediction criteria for model selection in stochastic regression and ARMA models. *Statistica Sinica 7*, 285–309.

Lanterman, A. (2001). Schwarz, Wallace, and Rissanen: Intertwining themes in theories of model order estimation. *International Statistical Review 69*(2), 185–212.

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.

Lee, M. (2002a). Generating additive clustering models with minimal stochastic complexity. *Journal of Classification 19*(1), 69–85.

Lee, M. (2002b). A simple method for generating additive clustering models with limited complexity. *Machine Learning 49*, 39–58.

Lee, P. (1997). *Bayesian Statistics — An Introduction*. London and Oxford: Arnold & Oxford University Press.

Lee, T. (2000). Regression spline smoothing using the minimum description length principle. *Statistics and Probability Letters 48*(71–82).

Lee, T. (2002c). Automatic smoothing for discontinuous regression functions. *Statistica Sinica 12*, 823–842.

Levenstein, V. (1968). On the redundancy and delay of separable codes for the natural numbers. *Problems of Cybernetics 20*, 173–179.

Li, J. (1999). *Estimation of Mixture Models*. Ph. D. thesis, Yale University, New Haven, CT.

Li, J., and A. Barron (2000). Mixture density estimation. In S. Solla, T. Leen, and K.-R. Müller (Eds.), *Advances in Neural Information Processing Systems*, volume 12, pp. 279–285.

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

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

Li, M., and P. Vitányi (1997). *An Introduction to Kolmogorov Complexity and Its Applications,* revised and expanded 2nd edition. New York: Springer-Verlag.

Liang, F., and A. Barron (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*. Cambridge, MA: MIT Press.

Liang, F., and A. R. Barron (2002). Exact minimax strategies for predictive density estimation, data compression, and model selection. In *Proceedings of the 2002 IEEE International Symposium on Information Theory (ISIT 2002)*.

Liang, F., and A. R. Barron (2004). Exact minimax strategies for predictive density estimation, data compression, and model selection. *IEEE Transactions on Information Theory 50*, 2708–2726.

Lindley, D., and A. Smith (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B 34*, 1–41. With discussion.

Littlestone, N., and M. Warmuth (1994). The weighted majority algorithm. *Information and Computation 108*(2), 212–261.

Liu, J., and P. Moulin (1998). A new complexity prior for multiresolution image denoising. In *Proceedings of IEEE Workshop on Time-Frequency Time-Scale Analysis*, pp. 637–640.

Lutwak, E., D. Yang, and G. Zhang (2005). Cramér-Rao and moment-entropy inequalities for Rényi entropy and generalized Fisher information. *IEEE Transactions on Information Theory 51*, 473–478.

MacKay, D. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge, UK: Cambridge University Press.

McAllester, D. (1998). Some PAC-Bayesian theorems. In *Proceedings of the Eleventh ACM Conference on Computational Learning Theory (COLT' 98)*, pp. 230–234.

McAllester, D. (1999). PAC-Bayesian model averaging. In *Proceedings of the Twelfth ACM Conference on Computational Learning Theory (COLT' 99)*, pp. 164–171.

McAllester, D. (2003). PAC-Bayesian stochastic model selection. *Machine Learning 51*(1), 5–21.

Mehta, M., J. Rissanen, and R. Agrawal (1995). MDL-based decision tree pruning. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD '95)*, pp. 216–221.

Meir, R., and N. Merhav (1995). On the stochastic complexity of learning realizable and unrealizable rules. *Machine Learning 19*, 241–261.

Merhav, N., and M. Feder (1998). Universal prediction. *IEEE Transactions on Information Theory IT-44*(6), 2124–2147. Special Commemorative Issue: Information Theory: 1948-1998.

Michalski, R., J. Carbonell, and T. Mitchell (1983). *Machine Learning, An Artificial Intelligence Approach*. San Francisco: Morgan Kaufmann.

Michie, D., D. Spiegelhalter, and C. Taylor (Eds.) (1994). *Machine Learning, Neural and Statistical Classification*. London: Ellis Horwood.

Modha, D. S., and E. Masry (1998). Prequential and cross-validated regression estimation. *Machine Learning 33*(1), 5–39.

Myung, I., V. Balasubramanian, and M. Pitt (2000). Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences USA 97*, 11170–11175.

Myung, I. J., M. A. Pitt, S. Zhang, and V. Balasubramanian (2000). The use of MDL to select among computational models of cognition. In

Advances in Neural Information Processing Systems, volume 13, pp. 38–44. Cambridge, MA: MIT Press.

Nannen, V. (2003). The paradox of overfitting. Master's thesis, University of Groningen, Groningen, the Netherlands.

Navarro, D. (2004). A note on the applied use of MDL appproximations. *Neural Computation 16*, 1763–1768.

Ndili, U., R. Nowak, and M. Figueiredo (2001). Coding-theoretic approach to image segmentation. In *Proceedings of the 2001 IEEE International Conference on Image Processing - ICIP'2001*.

Neal, R. (1996). *Bayesian learning for neural networks*. New York: Springer-Verlag.

Nowak, R., and M. Figueiredo (2000). Unsupervised segmentation of Poisson data. In *Proceedings of the International Conference on Pattern Recognition - ICPR'2000*, volume 3, pp. 159–162.

Osborne, M. (1999). MDL-based DCG induction for NP identification. In *Proceedings of the Third Conference on Computational Natural Language Learning (CoNLL '99)*, pp. 61–68.

Pednault, E. (2003). Personal communication, June 2003.

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

Poland, J., and M. Hutter (2006). MDL convergence speed for Bernoulli sequences. *Statistics and Computing 16*, 161–175.

Qian, G., G. Gabor, and R. Gupta (1996). Generalised linear model selection by the predictive least quasi-deviance criterion. *Biometrika 83*, 41–54.

Qian, G., and H. Künsch (1998). Some notes on Rissanen's stochastic complexity. *IEEE Transactions on Information Theory 44*(2), 782–786.

Quinlan, J., and R. Rivest (1989). Inferring decision trees using the minimum description length principle. *Information and Computation 80*, 227–248.

Rasmussen, C., and Z. Ghahramani (2000). Occam's razor. In *Advances in Neural Information Processing Systems*, volume 13, pp. 294–300.

Rasmussen, C., and C. Williams (2006). *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press.

Rényi, A. (1960). On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 547–561.

Rice, J. (1995). *Mathematical Statistics and Data Analysis*. Duxbury Press.

Ripley, B. (1996). *Pattern Recognition and Neural Networks*. Cambridge, UK: Cambridge University Press.

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

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

Rissanen, J. (1984). Universal coding, information, prediction and estimation. *IEEE Transactions on Information Theory 30*, 629–636.

Rissanen, J. (1986a). Order estimation by accumulated prediction errors. In J. Gani and M. B. Priestley (Eds.), *Essays in Time Series and Allied Processes*, pp. 55–61. Sheffield, UK: Applied Probability Trust.

Rissanen, J. (1986b). A predictive least squares principle. *IMA Journal of Mathematical Control and Information 3*, 211–222.

Rissanen, J. (1986c). Stochastic complexity and modeling. *Annals of Statistics 14*, 1080–1100.

Rissanen, J. (1987). Stochastic complexity. *Journal of the Royal Statistical Society, Series B 49*, 223–239. Discussion: 252–265.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry*. Hackensack, NJ: World Scientific.

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

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

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

Rissanen, J. (2001). Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Transactions on Information Theory 47*(5), 1712–1717.

Rissanen, J. (2007). *Information and Complexity in Statistical Modeling*. New York: Springer-Verlag.

Rissanen, J., and E. Ristad (1994). Language acquisition in the MDL framework. In E. Ristad (Ed.), *Language Computations*. Philadelphia: American Mathematical Society.

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

Rissanen, J., and I. Tabus (2005). Kolmogorov's structure function in MDL theory and lossy data compression. In P. D. Grünwald, I. J. Myung, and M. A. Pitt (Eds.), *Advances in Minimum Description Length: Theory and Applications*. Cambridge, MA: MIT Press.

Rissanen, J., and B. Yu (1995). MDL learning. In D. Kueker and C. Smith (Eds.), *Learning and Geometry: Computational Approaches, Progress in Computer Science and Applied Logic*, volume 14, pp. 3–19. Boston: Birkhäuser.

Rockafellar, R. (1970). *Convex Analysis*. Princeton, NJ: Princeton University Press.

Roos, T. (2004). MDL regression and denoising. Unpublished manuscript.

Roos, T., P. Myllymäki, and H. Tirri (2005). On the behavior of MDL denoising. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics (AISTATS '05)*, pp. 309–316.

Roos, T., H. Wettig, P. Grünwald, P. Myllymäki, and H.Tirri (2005). On discriminative Bayesian network classifiers and logistic regression. *Machine Learning 59*(3), 267 – 296.

Rosenfeld, R. (1996). A maximum entropy approach to adaptive statistical language modeling. *Computer, Speech and Language 10*, 187–228.

Ross, S. (1998). *A First Course in Probability*. Upper Saddle River, NJ: Prentice-Hall.

Rubin, H. (1987). A weak system of axioms for "rational" behavior and the nonseparability of utility from prior. *Statistical Decisions 5*, 47–58.

Savage, L. (1954). *The Foundations of Statistics*. Dover Publications.

Schölkopf, B., and A. J. Smola (2002). *Learning with Kernels*. Cambridge, MA: MIT Press.

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

Seeger, M. (2004). Gaussian processes for machine learning. *International Journal of Neural Systems 14*(2), 69–106.

Seidenfeld, T. (1986). Entropy and uncertainty. *Philosophy of Science 53*, 467–491.

Shafer, G., and V. Vovk (2001). *Probability and Finance – It's Only a Game!* New York: Wiley.

Shaffer, C. (1993). Overfitting avoidance as bias. *Machine Learning 10*, 153–178.

Shannon, C. (1948). The mathematical theory of communication. *Bell System Technical Journal 27*, 379–423, 623–656.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association 88*, 486–494.

Shao, J. (1997). An asymptotic theory for linear model selection (with discussion). *Statistica Sinica 7*, 221–242.

Shawe-Taylor, J., P. Bartlett, R. Williamson, and M. Anthony (1998). Structural risk minimisation over data-dependent hierarchies. *IEEE Transactions on Information Theory 44*(5), 1926–1940.

Shawe-Taylor, J., and N. Cristianini (2004). *Kernel Methods for Pattern Analysis*. Cambridge, UK: Cambridge University Press.

Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika 63*(1), 117–126.

Shtarkov, Y. M. (1987). Universal sequential coding of single messages. *Problems of Information Transmission 23*(3), 3–17.

Sober, E. (2004). The contest between parsimony and likelihood. *Systematic Biology 4*, 644–653.

Solomonoff, R. (1964). A formal theory of inductive inference, part 1 and part 2. *Information and Control 7*, 1–22, 224–254.

Solomonoff, R. (1978). Complexity-based induction systems: comparisons and convergence theorems. *IEEE Transactions on Information Theory 24*, 422–432.

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

Spiegelhalter, D., N. Best, B. Carlin, and A. van der Linde (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B 64*(4), 583–639.

Starkie, B. (2001). Programming spoken dialogs using grammatical inference. In *Advances in Artificial Intelligence (AI 2001)*. Berlin: Springer.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B 36*(2), 111–147.

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

Strang, G. (1988). *Linear Algebra and its Applications*, 3rd edition. Philadelphia: W.B. Saunders.

Su, Y., I. J. Myung, and M. A. Pitt (2005). Minimum description length and cognitive modeling. In P. D. Grünwald, I. J. Myung, and M. A. Pitt (Eds.), *Advances in Minimum Description Length: Theory and Applications*. Cambridge, MA: MIT Press.

Szpankowski, W. (1998). On asymptotics of certain recurrences arising in universal coding. *Problems of Information Transmission 34*(2), 142–146.

Tabus, I., J. Rissanen, and J. Astola (2002). Normalized maximum likelihood models for Boolean regression with application to prediction and classification in genomics. In W. Zhang and I. Shmulevich (Eds.), *Computational and Statistical Approaches to Genomics*.

Tabus, I., J. Rissanen, and J. Astola (2003). Classification and feature gene selection using the normalized maximum likelihood model for discrete regression. *Signal Processing 83*(4), 713–727. Special issue on Genomic Signal Processing.

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

Takeuchi, J., and A. Barron (1997). Asymptotically minimax regret for exponential families. In *Proceedings SITA '97*, pp. 665–668.

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

Takeuchi, J., and A. R. Barron (1998b). Robustly minimax codes for universal data compression. In *Proceedings of the Twenty-First Symposium on Information Theory and Its Applications (SITA '98)*.

Topsøe, F. (1979). Information-theoretical optimization techniques. *Kybernetika 15*(1), 8–27.

Topsøe, F. (2007). Information theory at the service of science. In I. Csiszár, G. Katona, and G. Tardos (Eds.), *Entropy, Search, Complexity*, volume 16 of *Bolyai Society Mathematical Studies*. New York: Springer-Verlag.

Townsend, P. (1975). The mind-body equation revisited. In C.-Y. Cheng (Ed.), *Psychological Problems in Philosophy*, pp. 200–218. Honolulu: University of Hawaii Press.

Uffink, J. (1995). Can the maximum entropy principle be explained as a consistency requirement? *Studies in History and Philosophy of Modern Physics 26B*, 223–261.

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge, UK: Cambridge University Press.

Vapnik, V. (1982). *Estimation of Dependencies Based on Empirical Data*. Berlin: Springer-Verlag.

Vapnik, V. (1998). *Statistical Learning Theory.* New York: Wiley.

Vapnik, V., and A. Chervonenkis (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications 16*(2), 264–280.

Vereshchagin, N., and P. Vitányi (2002). Kolmogorov's structure functions with an application to the foundations of model selection. In *Proceedings Fourty-Seventh IEEE Symposium on the Foundations of Computer Science (FOCS'02)*.

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

Viswanathan., M., C. Wallace, D. Dowe, and K. 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)*, pp. 405–416.

Vitányi, P. M. (2005). Algorithmic statistics and Kolmogorov's structure function. In P. D. Grünwald, I. J. Myung, and M. A. Pitt (Eds.), *Advances in Minimum Description Length: Theory and Applications*. Cambridge, MA: MIT Press.

von Neumann, J. (1928). Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen 100*, 295–320.

Vovk, V. (1990). Aggregating strategies. In *Proceedings of the Third Annual ACM Conference on Computational Learning Theory (COLT' 90)*, pp. 371–383.

Vovk, V. (2001). Competitive on-line statistics. *International Statistical Review 69*, 213–248.

Wagenmakers, E., P. D. Grünwald, and M. Steyvers (2006). Accumulative prediction error and the selection of time series models. *Journal of Mathematical Psychology 50*(2), 149–166.

Wallace, C. (2005). *Statistical and Inductive Inference by Minimum Message Length*. New York: Springer-Verlag.

Wallace, C., and D. Boulton (1968). An information measure for classification. *Computer Journal 11*, 185–195.

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

Wallace, C., and P. Freeman (1987). Estimation and inference by compact coding. *Journal of the Royal Statistical Society, Series B 49*, 240–251. Discussion: pages 252–265.

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

Watanabe, S. (1999a). Algebraic analysis for non-regular learning machines. In *Advances in Neural Information Processing Systems*, pp. 356–363.

Watanabe, S. (1999b). Algebraic analysis for singular statistical estimation. In *Tenth International Conference on Algorithmic Learning Theory (ALT'99)*, volume 1720 of *Lecture Notes in Computer Science*, pp. 39–50.

Wax, M. (1988). Order selection for AR models by predictive least squares. *IEEE Transactions on Acoustics, Speech and Signal Processing 36*(4), 581–588.

Webb, G. (1996). Further experimental evidence against the utility of Occam's razor. *Journal of Artificial Intelligence Research 4*, 397–417.

Wei, C. (1992). On predictive least squares principles. *Annals of Statistics 20*(1), 1–42.

Weisstein, E. (2006). Gamma function. From MathWorld–A Wolfram Web Resource.

Wiener, N. (1949). *Extrapolation, Interpolation and Smoothing of Stationary Time Series*. Cambridge, MA: MIT Press.

Wilks, S. (1938). The large sample distribution of the likelihood ratio for testing composite hypothesis. *Annals of Mathematical Statistics 9*, 60–62.

Willems, F., Y. Shtarkov, and T. Tjalkens (1995). The context-tree weighting method: basic properties. *IEEE Transactions on Information Theory 41*, 653–664.

Woodroofe, M. (1982). On model selection and the arcsine laws. *Annals of Statistics 10*, 1182–1194.

Xie, Q., and A. Barron (1997). Minimax redundancy for the class of memoryless sources. *IEEE Transactions on Information Theory 43*, 646–657.

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

Yamanishi, K. (1998). A decision-theoretic extension of stochastic complexity and its applications to learning. *IEEE Transactions on Information Theory 44*(4), 1424–1439.

Yamazaki, K., and S. Watanabe (2003). Singularities in mixture models and upper bounds of stochastic complexity. *International Journal of Neural Networks 16*, 1029–1038.

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

Yang, Y. (2005a). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrica 92*(4), 937–950.

Yang, Y. (2005b). Consistency of cross-validation for comparing regression procedures. Submitted for publication.

Yang, Y., and A. Barron (1998). An asymptotic property of model selection criteria. *IEEE Transactions on Information Theory 44*, 117–133.

Yang, Y., and A. Barron (1999). Information-theoretic determination of minimax rates of convergence. *Annals of Statistics 27*, 1564–1599.

Yu, B. (1994). Lower bound on the expected redundancy for classes of continuous Markov sources. In S. Gupta and J. Berger (Eds.), *Statistical Decision Theory and Related Topics*, volume V, pp. 453–466.

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

Yu, B., and T. Speed (1992). Data compression and histograms. *Probability Theory and Related Fields 92*, 195–229.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In P. Goel and A. Zellner (Eds.), *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, pp. 223–243. Amsterdam: North–Holland.

Zhang, J., and J. Myung (2005). A note on informative normalized maximum likelihood with data prior. Manuscript in preparation.

Zhang, T. (2004a). Learning bounds for a generalized family of Bayesian posterior distributions. In S. Thrun, L. K. Saul, and B. Schölkopf (Eds.), *Advances in Neural Information Processing Systems*, volume 16. Cambridge, MA: MIT Press.

Zhang, T. (2004b). On the convergence of MDL density estimation. In Y. Singer and J. Shawe-Taylor (Eds.), *Proceedings of the Seventeenth Conference on Learning Theory (COLT' 04)*, Lecture Notes in Computer Science. New York: Springer-Verlag.

List of Symbols

$\inf$: infimum, 44

J: observed information, 112

K: Kolmogorov complexity, 571
k: number of parameters, 61
k^*, *see* α^*
$\ddot{k}$, *see* $\ddot{\alpha}$
κ: kernel function, 390

$\mathcal{L}$: set of codelength functions, 99
L: codelength function (of code or coding system), 83
L_U: length function of uniform code, 88
$\bar{L}_{\text{IND}}$ with IND $\in$ { meta, plug-in, nml, Bayes, Jeffreys, two-part }, *see* $\bar{P}_{\text{IND}}$
$\bar{L}$: length function of universal code, 174
$\bar{L}_{\text{CUP}}$: CUP univ. code, 409
$\dot{L}_n$: first part of two-part code, 272
$\bar{L}_{\text{cond-2-p}}$: conditional two-part codelength, 285
LREG: luckiness regret, 311
$\ln$: natural logarithm, 45
$\log$: base 2 logarithm, 45

$\mathcal{M}$: model, 57
$\mathcal{M}^{(k)}$: k-dimensional model, **60**
$\langle\mathcal{M}\rangle$: information closure, 376
$\mathcal{M}_{\text{loc}}$: normal location family, 298
$\ddot{\mathcal{M}}_n$: discretization of $\mathcal{M}$, 476
$\mathcal{M}_\Theta$: model indexed by Θ, 57
$\mathcal{M}^{\mathbf{x}}$: linear model, 349
$M(\Theta_0, \epsilon)$: packing number, 222
μ: mean-value parameter, 610
$\hat{\mu}$, *see* $\hat{\alpha}$
$\max$: maximum, 44

$\min$: minimum, 44

$N(\Theta_0, \epsilon)$: covering number, 222
N: normal distribution, 51
$\mathbb{N}$: natural numbers, 42

O, o, ORDER: order notation, 45

$\mathcal{P}^{(k)}$: polynomial linear regression model, 64
P: probability distr./source/mass fn., 46
$P(\cdot \mid \cdot)$: conditional prob., 50
$P^{(n)}$: marginal on $\mathcal{X}^n$, 54
$\bar{P}_{\text{Bayes}}$: Bayesian univ. model, 175
P^*, *see* α^*
$\ddot{P}$, *see* $\ddot{\alpha}$
$\bar{P}_{\text{Cnml-1}}$: conditional NML universal model, 322
$\bar{P}_{\text{Cnml-2}}$: conditional NML universal model, 321
$\bar{P}_{|n}$: prequential estimator, 463
$\bar{P}_{\text{Jeffreys}}$: Bayesian universal model with Jeffreys' prior, 235
$\bar{P}_{\text{LB}}$: Liang-Barron universal model, 325
$\bar{P}_{\text{Lnml-1}}$: luckiness LNML universal model, 309
$\bar{P}_{\text{Lnml-2}}$: luckiness LNML universal model, 311
$\bar{P}_{\text{lucky-Jeffreys}}$: luckiness Bayesian universal model, 313
P_{me}: maximum entropy distribution, 638
$\bar{P}_{\text{meta}}$: metauniversal model, 304
P_{mre}: minimum relative entropy distribution, 640
$\bar{P}_{\text{nml}}$: NML universal model, 181
$\bar{P}_{\text{plug-in}}$: plug-in univ. model, 197

Subject Index