Chapter 7
Asymptotics and Coding Theory: One of the $n \to \infty$ Dimensions of Terry

Bin Yu

Terry joined the Berkeley Statistics faculty in the summer of 1987 after being the statistics head of CSIRO in Australia. His office was just down the hallway from mine on the third floor of Evans. I was beginning my third year at Berkeley then and I remember talking to him in the hallway after a talk that he gave on information theory and the Minimum Description Length (MDL) Principle of Rissanen. I was fascinated by the talk even though I did not understand everything. Terry pointed me to many papers, and before long Terry started to co-advise me (with Lucien Le Cam) as his first PhD student at Berkeley. It was truly a great privilege to work with Terry, especially as his first student at Berkeley since I had the luxury of having his attention almost every day – he would knock on my door to chat about research and to take me to the library to find references. Every Saturday I was invited to have lunch with him and his wife Sally at his rented house in the Normandy Village on Spruce Street – a cluster of rural European styled houses near campus (the most exotic part to me about the lunch was the avocado spread on a sandwich). Through my interactions with Terry, I was molded in $n \to \infty$ dimensions. In particular, I was mesmerised by the interplay shown to me by Terry of data, statistical models, and interpretations – it was art with rigor! I am able to pursue and enjoy this interplay in my current research, even though I ended up writing a theoretical PhD thesis.

The four papers under "asymptotics and coding theory" in this volume represent the MDL research done during my study with Terry (and Rissanen) and a paper after my PhD on Information Theory proper: lossy compression.

The Minimum Description Length (MDL) Principle was invented by Rissanen [7] to formalize Occam's Razor. Based on a foundation of the coding theory of Shannon, its most successful application to date is model selection – now a hot topic again under the new name of sparse modeling or compressed sensing in the high-dimensional situation. An idea closely related to MDL was Minimum Message Length (MML) first articulated in the context of clustering in Wallace and Boulton

Bin Yu
Departments of Statistics and Electrical Engineering & Computer Sciences, University of California, Berkeley, e-mail: binyu@stat.berkeley.edu

[13]. In a nutshell, MDL goes back to Kolmogorov's algorithmic complexity, a revolutionary concept, but not one that is computable. By rooting MDL in Shannon's information theory, Rissanen made the complexity (or code length) of a statistical or probabilistic model computable by corresponding a probability distribution to a prefix code via Kraft's inequality. At the same time, this coding interpretation of probability distribution removed the necessity of postulating a true distribution for data, since it can be viewed operationally a code-generating device. This seemingly trivial point is fundamental for statistical inference. Moreover, Rissanen put MDL on solid footing by generalizing Shannon's order source coding theorem to the second order to support the coding forms valid for use in MDL model selection. That is, he showed in Rissanen [8] that, for a nice parametric family of dimension k with n iid observations, they have to achieve a $\frac{k}{2}\log n$ lower bound asymptotically beyond the entropy lower bound when the data generating distribution is in the family. More information on MDL can be found in the review articles Barron et al. [3] and Hansen and Yu [5], and books Rissanen [9, 6] and Grünwald [4].

Not long before he and I started working on MDL in the late 1987, Terry had met Jorma Rissanen when Jorma visited Ted Hannan at the Australia National University (ANU). Hannan was a good friend of Terry's. Jorma's homebase was close by, the IBM Almaden Research Center in San Jose, so Terry invited him to visit us almost every month. Jorma would come with his wife and discuss MDL with us while his wife purchased bread at a store in Berkeley before they headed home together after lunch. We found Rissanen's papers original, but not always easy to follow. The discussions with him in person were a huge advantage for our understanding of MDL.

After catching up with the literature on MDL and model selection methods such as AIC [1] and BIC [11], we were ready to investigate MDL from a statistical angle in the canonical model of Gaussian regression and became among the first to explore MDL procedures in the nonparametric case using the convenient and canonical histogram estimate (which is both parametric and nonparametric). This line of research resulted in the first three papers on asymptotics and coding in this volume.

The research in Speed and Yu [12] started in 1987. The paper was possibly written in 1989, with many drafts including extensive comments by David Freedman on the first draft and it was a long story regarding why it took four years to publish. By then, it was well-known that AIC is prediction optimal and inconsistent (unless the true model is the largest model), while BIC is consistent when the true model is finite and one of the sub-regression models considered. Speed and Yu [12] addresses the prediction optimality question with refitting (causal or on-line prediction) and without refitting (batch prediction). A new lower bound on the latter was derived with sufficient achievability conditions, while a lower bound on the former had been given by Rissanen [8]. Comparisons of AIC, stochastic complexity , BIC, and Final Prediction Error (FPE) criteria [1] were made relative to the lower bounds and in terms of underfitting and overfitting probabilities. A finite-dimensional (fixed p to use modern terms) Gaussian linear regression model was assumed, as was common in other works around that time or before. The simple but canonical Gaussian regression model assumption made the technical burdens minimal, but it was sufficient to

reveal useful insights such as the orders of bias-variance trade-off when there was underfit or overfit, respectively. Related trade-offs are seen in analysis of modern model selection (sparse modeling) methods such as Lasso under high-dimensional regression models (large p large n). In fact, Speed and Yu [12] entertained the idea of a high-dimensional model through a discussion of finite dimensional models vs infinite dimensional models. In fact, much insight from this paper is still relevant today: BIC does well both in terms of consistency and prediction when the bias term drastically decreases to a lower level at a certain point (e.g. a "cliff" bias decrease when there are a group of major predictors and rest marginal). Working with Terry on this first paper of mine taught me lessons that I try to practice to this day: mathematical derivations in statistics should have meanings and give insights, and a good formulation of a problem are often more important than solving it.

The next two papers, Rissanen et al. [10] and Yu and Speed [14], are on histograms and MDL. They extend the MDL paradigm to the nonparametric domain. Around the same time Barron and Cover were working on other nonparametric MDL procedures through the resolvability index [2]. Rissanen spearheaded the first of the two papers, Rissanen et al. [10], to obtain a (properly defined) code length almost sure lower bound in the nonparametric case in the same spirit as the lower bound in the parametric case of his seminal paper [7]. This paper also showed that a histogram estimator achieve this lower bound. The second paper [14] introduced the minimax framework to address both the lower and upper code length bound questions for Lipschitz nonparametric families. Technically the paper was quite involved with long and refined asymptotic derivations, a Poissonization argument, and multinomial/Poisson cumulant calculations for which Terry showed dazzling algebraic power. A surprising insight from the second paper was that predictive MDL seemed a very flexible way to achieve the minimax optimal rate for expected code length. Working on the two histogram/MDL papers made me realize that there is no clear cut difference between parametric and nonparametric estimation: the so-called infinite dimensional models such as the Lipschitz family actually correspond to parametric estimation problems of dimensions increasing with the sample size. This insight holds for all nonparametric estimation problems and the histogram is a concrete example of sieve estimation.

The last of the four paper was on lossy compression of information theory proper. MDL model selection criteria are based on lossless code (prefix code) lengths. The aforementioned lower bound in Rissanen [7] was also fundamental for universal source (lossless) coding when the underlying data generating distribution has to be estimated, in addition to being the cornerstone of the MDL theory in the parametric case. It was natural to ask whether there is a parallel result for lossy compression where entropy is replaced by Shannon's rate-distortion function. Yu and Speed [15] showed it was indeed the case and there are quite a few follow-up papers in the information theory literature including Zhang et al. [16].

During my study with Terry, starting in the late 1987, Terry was moving full steam into biology as a visionary pioneer of statistical bioinformatics. To accommodate my interest in analysis and asymptotic theory and possibly pursue his other love for information theory rather than biology, Terry was happy to work with me

on theoretical MDL research and information theory, an instance of Terry's amazing intellectual versatility as amply clear from this volume.

References

[1] H. Akaike. A new look at the statistical model identification. *IEEE Trans. AC*, 19:716–723, 1974.

[2] A. R. Barron and T. M. Cover. Minimum complexity density estimation. *IEEE Trans. Inform. Theory*, 37:1034–1054, 1991.

[3] A. R. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Trans. Inform. Theory*, 44:2743–2760, 1998.

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

[5] M. H. Hansen and B. Yu. Model selection and the principle of minimum description length. *J. Am. Stat. Assoc.*, 96:746–774, 2001.

[6] J. Rissanen. *Information and Complexity in Statistical Modeling*. Springer, New York, 2007.

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

[8] J. Rissanen. Stochastic complexity and modeling. *Ann. Stat.*, 14:1080–1100, 1986.

[9] J. Rissanen. *Stochastic Complexity and Statistical Inquiry*. World Scientific, Singapore, 1989.

[10] J. Rissanen, T. P. Speed, and B. Yu. Density estimation by stochastic complexity. *IEEE Trans. Inform. Theory*, 38:315–323, 1992.

[11] G. Schwarz. Estimating the dimension of a model. *Ann. Stat.*, 6:461–464, 1978.

[12] T. P. Speed and B. Yu. Model selection and prediction: Normal regression. *Ann. Inst. Stat. Math.*, 45(1):35–54, 1993.

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

[14] B. Yu and T. P. Speed. Data compression and histograms. *Probab. Theory Relat. Fields*, 92:195–229, 1992.

[15] B. Yu and T. P. Speed. A rate of convergence result for a universal D-semifaithful code. *IEEE Trans. Inform. Theory*, 39:813–820, 1993.

[16] Z. Zhang, E. Yang, and V. K. Wei. The redundancy of source coding with a fidelity criterion. *IEEE Trans. Inform. Theory*, 43:71–91, 1997.