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

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

S. Dudoit (ed.), *Selected Works of Terry Speed*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4614-1347-9_7, © Springer Science+Business Media, LLC 2012

[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 as 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 [6, 9] 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. 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 is 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 is 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.

IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 38, NO. 2, MARCH 1992

Density Estimation by Stochastic Complexity

Jorma Rissanen, *Senior Member, IEEE*, Terry P. Speed, Bin Yu

Abstract—**The results by Hall and Hannan on optimization of histogram density estimators with equal bin widths by minimization of the stochastic complexity, are extended and sharpened in two separate ways. As the first contribution, two generalized histogram estimators are constructed. The first has unequal bin widths which, together with the number of the bins, are determined by minimization of the stochastic complexity with help of dynamic programming. The other estimator consists of a mixture of equal bin width estimators, each of which is defined by the associated stochastic complexity. As the main contribution in this paper, two theorems are proved, which together extend the universal coding theorems to a large class of data generating densities. The first gives an asymptotic upper bound for the code redundancy in the order of magnitude, achieved with a special predictive type of histogram estimator, which sharpens a related bound. The second theorem states that this bound cannot be improved upon by any code whatsoever.**

Index Terms—**MDL Principle, universal coding, histograms, asymptotic bounds, variable bin widths.**

I. Introduction

The MDL (minimum description length) principle to nonparametric density estimation is applied in this paper. This principle permits us to compare any two density estimators based upon a finite set of observed data by the codelength with which the data together with the estimator itself can be encoded. We prefer an estimator that achieves a short total codelength, which means that the best estimators are such that they assign high probabilities to clusters of the data points while at the same time the estimators themselves are not too complex to describe. Hence, for example, a histogram estimator with a large number of bins will not necessarily be good, because we have to describe, one way or another, the large number of counts of the observations falling in these bins. Similarly, the usual kernel estimators, which are formed as a sum of functions, one centered at each observed data point, are bad, because to describe them we need at least as many bits as for the description of the data points themselves. However, such estimators can be greatly simplified by retaining just enough functions to permit a good fit to the data, the number of them being subject to optimization. Such pruned-down kernel estimators turn out to have quite short codelengths [21].

In [11], an idealized codelength, the stochastic complexity, based upon the class of histogram estimators with equal-width bins was computed, which when minimized gave the optimal number of the bins and the associated density estimator. This estimator turns out to be good for data that are roughly uniformly distributed. However, when the distribution is strongly nonuniform, for instance having a long sparse tail, then many of the optimized number of bins may have very few data points or none at all, and one may then say that for the sparse portion of the data the density function is described with unnecessary detail. For such reasons, we extend the Hall–Hannan stochastic complexity calculation to the class of histogram estimators with variable-width bins, which can be calculated with dynamic programming. Despite an increased number of additional parameters to be encoded, the resulting codelength can be shorter than the Hall–Hannan stochastic complexity, while never exceeding it by more than about three bits for the entire data string. For small data sets histogram estimators lack smoothness. However, by constructing an estimator as a mixture of many equal bin width histograms we achieve a degree of "smoothness," not locally in terms of continuity or differentiability, but in a broader sense, without sacrificing efficiency. The analysis of the new estimators appears to be difficult, and we compare their performance with the equal bin width histogram estimator in an example.

It was shown in [11] that the optimal number of bins is also asymptotically of the correct magnitude to minimize the largest absolute deviation of the histogram estimator from the data generating density, in a fairly large family of nonparametric densities. Although such a result lends support to the idea of MDL principle providing good estimators, the support is somewhat indirect: the optimality is in terms of a sensible but still arbitrary distance measure. As the main analytic result in this paper, we prove another, stronger optimality property of a complexity based estimator, denoted $f^*(y \mid x^t)$, which is extended to a family of densities $f_n^*(x^n)$ for sequences $x^n = x_1, \cdots, x_n$ predictively by multiplication. In broad terms, we show that this estimator gives asymptotically the shortest codelength for the data in the order of magnitude that can be achieved by any density estimator, be it of histogram type or not, relative to the class $\mathcal{M}$ of densities $f(y)$ defined on the unit interval, which are uniformly bounded and also bounded away from zero and from infinity, and each having a bounded first derivative. These are extended to sequences by independence with the result $f^n(x^n)$. Moreover, we spell out the shortest mean

Manuscript received December 6, 1989. T. P. Speed and B. Yu are supported in part by NSF Grant DMS 88-02378.

J. Rissanen is with IBM Almaden Research Center, 650 Harry Road, Room K52/802, San Jose, CA 95120-6099.

T. P. Speed is with the Department of Statistics, Evans Hall, 3rd Floor, University of California, Berkeley, CA 94720.

B. Yu is with the Department of Statistics, University of Wisconsin, Madison, WI 53706.

IEEE Log Number 9104812.

0018-9448/92$03.00 © 1992 IEEE

codelength, which we define to be the asymptotic stochastic complexity of the data, relative to the nonparametric model class $\mathscr{M}$ in question. More specifically, our first theorem states that

$$\frac{1}{n} E_f \log \frac{f^n(x^n)}{f_n^*(x^n)} \leq A_f(n^{-2/3}), \qquad (1.1)$$

while the second theorem states that for any family of density estimators $\{g_t(x^t)\}$

$$\frac{1}{n} E_f \log \frac{f^n(x^n)}{g_n(x^n)} \geq Kn^{-2/3} \qquad (1.2)$$

holds for some positive constant K and all densities $f \in \mathscr{M}$, except for a set that is asymptotically ignorable in a suitable sense. We recall [17], [18] that for a model class with k free parameters, the right-hand side in (1.2), which represents the shortest codelength required to encode the optimal model with which the code for the data is designed, is $\frac{k}{2n} \log n$. Hence, we see that it takes a longer code to describe the estimator in the non-parametric family, as it should.

In [1], density estimators $\hat{p}$, minimizing a codelength criterion of the form $L(x^n, q) = -\log q(x^n) + L_n(q)$ were studied, where the second term denotes a prefix codelength for the estimator. Instead of providing an explicit construction for this length the authors specify it abstractly by certain properties. As their main contribution, the authors define the *index of resolvability* and show it to provide an upper bound both for the code redundancy, as well as for the Hellinger distance between the "true" density and its estimator of the form $\hat{p}(y \mid x^n)$, in probability. Further, an asymptotic formula is given for the index of resolvability. There are three main differences between their work and ours. First, we give an explicit construction for a density estimator, obtained with the MDL principle. Second, the class of estimators, provided by the two-stage codelength in [1], excludes those which do not satisfy the imposed condition that $L_n(\cdot)$ depends only on n as well as the important estimators obtained by a predictive coding process or by stochastic complexity. This is because the codelength for the data, resulting from these estimators, cannot be separated into codelengths for the model and the data, and hence the index of resolvability is inapplicable to them. By contrast, our second theorem does apply, not only to predictively constructed estimators but to estimators of any kind.

II. Histogram Estimators and Universal Codes

In [11], the *stochastic complexity* (for a general definition, see [21, Section 3.2]) of a set of observed data, relative to the class of histogram densities with equal size bins, their number to be optimized, was derived and the associated density estimator constructed. Another paper with similar ideas is [6], based on an earlier paper [5]. In the former, Dawid considers what is in effect a density estimate obtained from the equal bin width stochastic complexity in predictive form, and he demonstrates through simulations some of its desirable properties.

Although the histogram-like density estimators with the number of bins optimized will be shown to have strong asymptotic optimality properties among all density estimators whatsoever, other estimators may well perform better for small and medium size data or have other desirable properties such as a great degree of smoothness. For example, the various kernel estimators can be designed to provide any desired degree of smoothness, and the number of functions can be optimized with the MDL principle even though we cannot calculate the stochastic complexity for them in a closed form. In this section we study two generalizations of the usual histogram estimators, in both of which we can take advantage of the closed form solution for the stochastic complexity. The first class of estimators have variable-length bins, the lengths as well as the number of the bins determined by optimizing the stochastic complexity. This class shares the asymptotic optimality properties of the equal bin width histogram estimators. The other class consists of a mixture of a collection of ordinary equal bin width histograms, aimed at providing increased over-all "smoothness."

We begin by generalizing the Hall–Hannan complexity and the associated density estimators to histograms with variable-width bins. For convenience of notation we index the observed data so that $x_1 \leq x_2 \leq \cdots \leq x_n$, and note that the indexes need have no bearing on the time order of their arrival. Without loss of generality, we take all the observed data points as integers with the smallest $x_1 = 0$, and we write $x^n = x_1, \cdots, x_n$. Let $a = (a_1, \cdots, a_{m-1})$ denote an increasing sequence of the end points of m bins $[0, a_1]$, $(a_1, a_2], \cdots, (a_{m-1}, R]$, partitioning the range $[0, R]$; let $R_i = a_i - a_{i-1}$, $a_0 = 0$ and $a_m = R$, denote the length of the ith bin. Next, consider parametric histogram densities defined by $f(y \mid p, R, m, a) = \frac{p_i}{R_i}$, if y falls in the ith bin, where $p = (p_1, \cdots, p_m)$ denotes nonnegative parameters with sum unity.

With the uniform prior $\pi(p) = (m-1)!$ on the simplex defined by the parameters, we can evaluate the integral

$$f(x^n \mid R, m, a) = \int \prod_{t=1}^{n} f(x_t \mid p, R, m, a) \pi(p)\, dp$$

$$= \left(\prod_{i=1}^{m} R_i^{-n_i} \right) \frac{(m-1)! \Pi_i n_i!}{(m+n-1)!}. \qquad (2.1)$$

Then the stochastic complexity, $I(x^n \mid R, m, a) = -\log f(x^n \mid R, m, a)$, fixing n, R, m and a, is given by

$$I(x^n \mid R, m, a) = \sum_{i=1}^{m} n_i \log R_i + \log \binom{n}{n_1, \cdots, n_m}$$

$$+ \log \binom{n+m-1}{n}, \qquad (2.2)$$

in terms of the multinomial

$$\binom{n}{n_1, \cdots, n_m} = \frac{n!}{\Pi_i n_i!}$$

and the binomial

$$\binom{n+m-1}{n} = \frac{(n+m-1)!}{n!(m-1)!}$$

coefficients. Here, n_i denotes the number of observations that fall in the ith bin. In the special case $R = 1$ and $R_i = 1/m$ we get the Hall–Hannan complexity, which we write as $I(x^n \mid m)$.

We may interpret the three terms in (2.2) as codelengths corresponding to a particular encoding process. (Although a codelength is an integer-valued quantity it is convenient to regard the negative logarithm of a probability as a kind of idealized codelength, and with a further idealization we call even the negative logarithm of a density a codelength [17].) The last term is the length required to encode the m nonnegative integers n_i, when m is given; this is a special case of (2.4), derived next. The first two terms give the codelength for the observations when we imagine each to be specified by a pair (i, y), where i gives the bin (the ith) in which the observation falls, and y gives the position of the observation within this bin. Encoding of y, when we know that it belongs to the ith bin, clearly takes about $\log R_i$ bits, and the first term in (2.2) is the sum of these over all the n observations. Finally, the second term in (2.2) is the codelength required to encode the bin numbers (the first component in (i, y)) of all the n observations, for it is the logarithm of the number of all strings of length n in m symbols with the given counts. Another, predictive encoding process is defined by the conditional densities in (2.7), and taking the sum of their negative logarithms gives exactly the same codelength (2.2) for the same parameter values.

We can find the optimal sequence of the bins by dynamic programming. However, since the codelength required to encode the sequence of the bins, which must be added to (2.2), may be large, we generally get a shorter overall codelength if the end points of the bins are suitably restricted. We do this by introducing two parameters. The first is the precision, an integer d, with which the break points a_i are expressed; in other words, $a_i = k_i d$ and $\mathbf{a} = d\mathbf{k}$, where $\mathbf{k} = (k_1, \cdots, k_{m-1})$. The second new parameter is the minimum bin width we permit, say κd, also expressed as a multiple of d. To apply the dynamic programming argument, subdivide the interval $[0, R]$, where R is taken as a variable multiple of d, into $m + 1$ bins with the break points $\mathbf{a} = a_1, \cdots, a_{m-1}, \tau$ and write $\mathbf{a}' = a_1, \cdots, a_{m-1}$. Then from (2.2) we get by a straightforward calculation a decomposition of the form $I(x^{n(R)} \mid R, m+1, \mathbf{a}) = I(x^{n(\tau)} \mid \tau, m, \mathbf{a}') + \cdots$, where $x_1, \cdots, x_{n(R)}$ denotes the portion of the observed data falling within $[0, R]$, and similarly for $x^{n(\tau)}$. The remaining term, represented by the dots, is given by the last three terms in (2.3). Next, let

$$L_m(R) = \min_{\mathbf{a}} I\left(x^{n(R)} \mid R, m, \mathbf{a}\right),$$

where $\mathbf{a} = d\mathbf{k}$ with $k_{i+1} - k_i \geq \kappa$ and $k_{m-1} \leq k - \kappa$, and $k = R/d$. By the dynamic programming argument, we then get the recursion

$$L_{m+1}(R) = \min_{\tau} \left\{ L_m(\tau) + (n(R) - n(\tau)) \log (R - \tau) + \log \binom{n(R)+m}{n(\tau)+m} + \log \frac{n(\tau)+m}{m} \right\}, \tag{2.3}$$

where τ, besides being less than R, is also restricted to be a multiple of d as well as by the requirement of the minimum bin width. The recursive equations are solved for $m \geq 1$ and for $R = d\kappa, d(\kappa + 1), \cdots$, until the desired range including all the observations is reached. The initial value is $L_1(R) = n(R) \log R$ for all R. A recursive evaluation of (2.3) for the desired value of m and the range gives both the minimized stochastic complexity and the optimal sequence of the bin boundaries $\hat{\mathbf{a}}$ with about $(R/d)^2$ operations.

We need the codelengths required to encode the various integer-valued parameters, of which we first consider the increasing sequence $\mathbf{k}$ with $k_i - k_{i-1} \geq \kappa$ for $i = 1, \cdots, m-1$, $k_0 = 0$, and $k_m = k$. To get this length, associate with the sequence in a one-to-one fashion a binary string as follows. Begin with $k_1 - \kappa$ 0's and a 1, followed by $k_2 - k_1 - \kappa$ 0'*s* and a 1, and so on until $k - k_{m-1} - \kappa$ 0's are added, followed by the last 1. The string has $k - m\kappa$ 0's and m 1's, and it always ends with a 1. Hence, the codelength required for such a sequence is to within one bit

$$L(\mathbf{k}) = \log \binom{k - m(\kappa - 1) - 1}{m - 1}. \tag{2.4}$$

This (nonprefix) length estimate is valid provided that m, $k = R/d$, and κ are given. In fact, we need to encode the four parameters m, d, κ, and k, since in general the range R cannot be regarded as given. The code for these four integers must be a prefix code, for we must be able to decode them from a preamble in the entire code string without a separating comma. We recall that a positive integer i can be encoded in a prefix manner with about $L^*(i) = 1.5 + \log i + \log \log i + \cdots$ bits, where the series includes all the positive terms [8], [16]. Hence, we can encode the four parameters with the length $L(d, m, \kappa, k) = L^*(d) + L^*(m) + L^*(\kappa) + L^*(\max\{1, k - m\kappa\})$ bits. The best codelength we can get for the data sequence using variable-length bins by this procedure is then

$$L_V(x) = \min_{m, \mathbf{k}, d, \kappa} \{ I(x \mid \mathbf{k}, m, d\mathbf{k}) + L(\mathbf{k}) + L(d, m, \kappa, k) \}. \tag{2.5}$$

For each m, d, κ, and k only the first term in (2.5) depends on the sequence $\mathbf{k}$, and the minimization is done by the recursion (2.3). The minimization with respect to the remaining three bounded integer-valued parameters (k being determined by d and the range, which is not subject to optimization) is to be done by exhaustive search.

Consider the choice $d = 1$ and $\kappa = \lceil (k/m)$, where $\lceil x$ denotes the least integer upper bound for x. This forces the bins to have equal lengths, which means that $L(\mathbf{k}) = 0$, and with $k = R$ we have

$$L(1, m, \kappa, R) = L(m, R) + 2L^*(1),$$

where the first term denotes the prefix codelength for m and the range R, both to be added to the Hall-Hannan complexity to make it complete. We then see that the codelength $L_V(x^n)$ of the optimal variable bin width code never exceeds the length, say $L_E(x^n)$, of the optimal equal bin width code by more than about $2L^*(1) = 3$ bits. A further (trivial) subclass of uniform densities results from the choice $d = 1$ and $m = 1$.

Once the optimal parameters $\hat{m}$, $\hat{d}$, $\hat{\kappa}$, and $\hat{k}$, minimizing the stochastic complexity are found, we generally wish to construct a density estimate. One way is to calculate the natural histogram estimator

$$\hat{f}_V(y \mid x^n) = \frac{n_i}{n} \hat{R}_i^{-1}, \tag{2.6}$$

for y in the ith bin with length $\hat{R}_i = (\hat{k}_i - \hat{k}_{i-1})\hat{d}$. Another is defined by (2.1) as

$$\hat{f}(y \mid x^n) = \frac{f(x^n y \mid R, \hat{m}, \hat{a})}{f(x^n \mid R, \hat{m}, \hat{a})} = \frac{n_i + 1}{n + \hat{m}} \hat{R}_i^{-1}, \tag{2.7}$$

for y also in the ith bin; the pair $x^n y$ denotes the string $x_1, \cdots, x_n, y$ of length $n + 1$.

We next describe the estimator obtained as a mixture of the equal bin width histograms. Writing first

$$f(y \mid x^n, m) = \frac{n_i + 1}{n + m} \frac{m}{R}, \tag{2.8}$$

for the special case of (2.7) with equal bin widths, we define the mixture density estimator as

$$\begin{aligned} \hat{f}_M(y \mid x^n) &= \frac{1}{M} \sum_{m=1}^{M} f(y \mid x^n, m) \\ &= \frac{1}{RM} \sum_{m=1}^{M} m \frac{n_i + 1}{n + m}, \end{aligned} \tag{2.9}$$

where, again, i is the index of the bin in which y falls, and n_i is the number of the data points that fall within this bin. The number M is taken as a parameter to be optimized. With this estimator the data sequence can be encoded with the codelength

$$-\log \hat{f}_M(x^n) = -\sum_{t=0}^{n-1} \log f_M(x_{t+1} \mid x^t). \tag{2.10}$$

Example: We calculated the optimal codelength $L_V(x^n) = 572$, obtained with the parameters $\hat{m} = 4$, $\hat{d} = 6$, $\hat{\kappa} = 18$, $k = 18$, and $\hat{k} = 12, 14, 16$ for the set of 76 integers 0, 7, 18, 39, 49, 50, 61, 80, 82, 82, 82, 82, 84, 84, 85, 86, 88, 89, 89, 91, 91, 92, 92, 92, 92, 92, 93, 95, 96, 96, 101, 101, 101, 101, 105, 107, 107, 111, 112, 115, 115, 116, 117, 117, 118, 119, 119, 121, 122, 123, 124, 124, 125, 125, 125, 129, 129, 129, 130, 131, 196, 201, 201, 203, 212, 232, 236, 241, 241, 243, 243, 243, 245, 246, 248, 248. We also calculated the "fit," $-\log f(x^n \mid R, \hat{m}, \hat{a}) = 549$. Fig. 1 shows the corresponding density estimator marked "variable bin width." The optimal number of equal-length bins is $\hat{m} = 13$, which gives the total codelength $L_E(x^n) = 577$ and the "fit" 554. In Fig. 1, the associated density estimator is marked "equal bin width." Finally, we calculated the total codelength for the mixture estimator with the optimized number $\hat{M} = 16$ as $L_M(x^n) = 578$, and the "fit" $-\log \hat{f}_M(x^n) = 557$. The dependence of the codelength on the number of terms M in the mixture is very slight, and we can pick it in the form of an integer power of two. The associated density estimator is marked with "mixture" in Fig. 1.

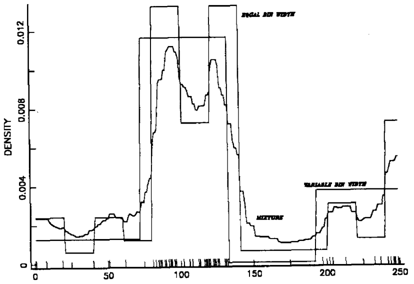

Fig. 1. Three density estimators.

Due to the relatively small data set the codelengths obtained with all the three estimators are virtually the same, despite the fact that the estimators differ considerably. We see in Fig. 1 that the large optimal number of bins, 13, in the equal bin width density estimator makes it somewhat "jumpy" in creating perhaps needlessly many local maxima and minima. The four bins in the variable bin width estimator, by contrast, give a less ragged density function. By far the best looking estimator, however, is the mixture density, in which, unlike in the usual kernel and spline estimators, "smoothness" is achieved without imposing analytic continuity. As a practical matter, both the equal bin width and the mixture estimators are easy to calculate, requiring only $O(n)$ number of operations for n observations, which makes them feasible to compute even for multidimensional data. By contrast, the variable bin width estimator requires $O(R^2)$ operations, which just about confines their calculations for scalar observations only.

III. Asymptotic Optimality

Ordinarily the goodness of a density estimator is expressed in terms of a suitably chosen distance measure between the estimated density and an assumed data generating one in some class. In this paper, in accordance with the MDL principle, we have taken the codelength with which the data and the estimator itself can be encoded as the yardstick for the quality of an estimator. The purpose of this section is to derive optimum asymptotic rates, in the order of magnitude, for the expected codelengths, relative to a class of smooth data generating densities, and to demonstrate an estimator and the associated universal code that achieves the optimal rate in the order of magnitude. The same measure also translates into the Kullback distance between the estimated and the data generating densities, which provides further support for the codelength criterion. Specifically, the class $\mathcal{M}$ consists of densities $f(y)$, defined on the unit interval such that $0 < c_0 \leq f(y) \leq c_1$, $c_0 < 1$, $c_1 > 1$, and each density having a bounded first derivative, say $|f'(y)| \leq c_f$. This class is larger than the class $\mathcal{M}'$ considered in [11], in which the absolute derivative was required to be bounded uniformly over the densities. It was shown in [11] that for each density function f in the class $\mathcal{M}'$ the number of bins minimizing (2.2) for $R = 1$ satisfies $\hat{m}(x^n)/(n/\log n)^{1/3} \to C_f$ in probability, where C_f is a constant for each f. Moreover, the corresponding optimal bin width $1/\hat{m}(x^n)$ is also of the correct order of magnitude for minimizing the largest absolute deviation of the histogram estimator from any data generating density in the same class. With this number of bins, the stochastic complexity (2.2), denoted now by $I(x^n \mid m_n)$, behaves asymptotically like

$$\frac{1}{n} I(x^n \mid m_n) \approx -\frac{1}{n} \log f(x^n) + K\left(\frac{\log n}{n}\right)^{2/3} \quad (3.1)$$

in probability, and the second term gives also the amount by which the mean-per-symbol stochastic complexity exceeds the entropy. Since the codelength provided by the variable bin width estimator, constructed in the previous section, exceeds the optimal equal bin width estimator $I(x^n \mid m_n)$ only by at most three bits, its mean-per-symbol length, too, is asymptotically no greater than the right-hand side.

However, we get a smaller excess term for a different density estimator, constructed from the stochastic complexity $I(x^n \mid m)$ in a predictive way. This estimator is defined in terms of the conditional densities (2.8), rewritten here for $R = 1$ and $0 \leq t$,

$$f(x_{t+1} \mid x^t, m) = \frac{t_i + 1}{(t + m)} m, \quad (3.2)$$

where we let the number of bins grow with t as $m_t^* = \lceil t^{1/3}$ to take advantage of an increasing information. Writing

$$f^*(y \mid x^t) = f(y \mid x^t, m_{t+1}^*), \quad (3.3)$$

where x^0 is the empty string, we obtain for any string $f_n^*(x^n) = \Pi_{t=1}^n f^*(x_t \mid x^{t-1})$, regardless of the ordering of the observations. Notice that the negative logarithm of $f_n^*(x^n)$ is not the stochastic complexity $I(x^n \mid m)$ for any single value of m, but rather it is the sum of the increments $I(x^{t+1} \mid m_t^*) - I(x^t \mid m_t^*)$. It is interesting that here the mean predictive codelength is asymptotically strictly shorter than the nonpredictive one. This is in contrast with all the parametric model classes studied, where the two mean lengths are asymptotically equal.

Theorem 1: For all $f \in \mathcal{M}$,

$$\frac{1}{n} E_f \log \frac{f^n(x^n)}{f_n^*(x^n)} \leq A_f n^{-2/3}, \quad (3.4)$$

where $f^n(x^n) = \Pi_{t=1}^n f(x_t)$ and A_f a number dependent on f. Also,

$$E_f \int_0^1 f(y) \log \frac{f(y)}{f^*(y \mid x^n)} dy \leq B_f n^{-2/3}, \quad (3.5)$$

B_f is a number dependent of f. The expectation E_f is taken with respect to f^n over the data sequences x^n.

The proof is given in Appendix A.

The question arises whether any code exists with shorter mean length in the order of magnitude than given by the right-hand side of (3.4). Just as for parametric model classes [17], [18], we cannot expect this to be the greatest lower bound for *all* data generating densities, since one designed with f clearly reaches the entropy, but what we can expect is the right-hand side of (3.4) to represent, in the order of magnitude, the shortest possible mean codelength for all but a negligible subset of the densities. This turns out to be true, although the lack of nonsingular measures in function spaces forces us to invent a plausible way to capture the intuitive idea of "negligible subset" of densities. For this we need some notation. Consider a partition of the unit interval into m_n equal size bins, where

$$m_n = \lceil (n^{1/3}/\log n). \quad (3.6)$$

For a density function f in $\mathcal{M}$ let p_i denote the probability of the ith bin. Write $f_i = m_n p_i$ and denote by $\theta_f = (f_1, \cdots, f_{m_n})$ the collection of such linear functionals that act as parameters although they do not determine the density function completely. Further, write $\Omega_n = \{\theta_f \in R^{m_n} \mid f \in \mathcal{M}\}$.

Theorem 2: Let $g = \{g_n(x^n)\}$ be any family of densities on I^n, where I is the unit interval, such that the Kolmogorov consistency conditions are satisfied, and each member is positive except in a set of measure zero. Then, there exists a positive constant K such that for all sufficiently large values of n and all f in $\mathcal{M}$,

$$\frac{1}{n} E_f \log \frac{f^n(x^n)}{g_n(x^n)} \geq K n^{-2/3}, \quad (3.7)$$

except for f in a set $\{f \mid (f_1, \cdots, f_{m_n}) \in A_{g,n} \subset \Omega_n \subset R^{m_n}\}$ such that the ratio of the volume of $A_{g,n}$ to that of the entire set $\Omega_n = \{\theta_f \in R^{m_n} \mid f \in \mathcal{M}\}$ satisfies

$$\frac{V(A_{g,n})}{V(\Omega_n)} \to 0,$$

as $n \to \infty$.

The proof is given in Appendix B.

Remarks: The requirement that the family $\{g_n\}$ satisfies the consistency conditions for a random process is not really needed in this version of the theorem. However, in universal coding the main interest is in encoding sequences modeled as samples from random processes, for which the consistency requirement provides a collection of Kraft-inequalities for the symbols and hence, prefix codes. Further, just as in the case with parametric model classes we may interpret the right-hand side in (3.4) as the optimal model cost per observation in order of magnitude; i.e, the codelength per observation required to encode the density estimator itself. Since the estimator is defined predictively, this cost does not appear explicitly in the total codelength, but it may be visualized as resulting from the cumulative effect of the errors in the estimated counts n_i. This cost is greater than in the case with parametric models, namely, $(k \log n)/2n$, where k denotes the number of free parameters, reflecting the fact that the nonparametric model class here is richer and its members more difficult to estimate. The choice of $m_t^* = \lceil t^{1/3}$ is seen to be appropriate for the model class $\mathcal{M}$ with its specific smoothness conditions. For a class with different smoothness conditions and hence different ϵ-entropy, [7], another choice would be better leading to a different optimal rate. Extensions and variations of Theorems 1–2 have already been proved, including an a.s. approximation for the codelength and a minimax form of Theorem 2. These will be published separately. Finally, the second bound (3.5) serves to indicate that not only does the codelength obtained with the estimator (3.3) converge to the entropy, but also the estimator itself converges to the data generating density at the same rate, when the distance is measured in terms of the Kullback distance.

We may regard the theorems as the latest step in the series of statements about universal codelength, relative to model classes of steadily increasing generality. The very first such result is Shannon's coding theorem for the singelton class $\{P(x)\}$. It was followed by the theorems in [3], [4], [14], establishing worst case bounds for independent and Markov sources as well as for some gaussian classes. In [17], a sharper inequality of the type in Theorem 2, valid for all but a vanishing subset of parameters, were proved for general parametric classes, which was further strengthened for the Markov sources in [19]. A further generalization of the latter to the ARMA class became possible through the works [9] and [18].

The reachability of the lower bound with predictive coding has important implications in prediction theory. Indeed, the bound for the codelength translates naturally to a bound for the mean prediction error. Here, the early results in [17] and [20] have been vastly generalized in [12], [10], and [13].

Acknowledgment

The authors thank A. Barron for letting them see the two papers cited prior to their publication and for his valuable comments on our draft manuscript.

Appendix A

Partition the unit interval into m equal-length bins and let $I_k = \left(\frac{k}{m}, \frac{k+1}{m}\right]$; here $0 < m < n$ and $k = 0, \cdots, m$. To simplify the notation in (3.2) slightly we write

$$f(y \mid x^n, m) \equiv f_{n,m}(y) = \frac{1}{1 + m/n}\left(\frac{n_k}{n} m + \frac{m}{n}\right),$$

for $y \in I_k$, while bearing in mind that this density depends on the data x^n. Notice that $0 < \frac{1}{n+1} \leq f_{n,m}(y) \leq m$.

Lemma 1: For every $f \in \mathcal{M}$, $0 < c_0 \leq f(y) \leq c_1$, $c_0 < 1$, $c_1 > 1$,

$$\begin{aligned} E_f \int_0^1 f(x) \log \frac{f(x)}{f_{n,m}(x)}\, dx & \\ \leq \frac{2}{c_0} E_f \int_0^1 (f(x) - f_{n,m}(x))^2\, dx & \\ + 4m(n+1)(c_1^2 + m^2) e^{-Bc_0^2 n/(4m)}, & \end{aligned}$$

where B is a positive constant.

Proof of Lemma 1: Put $p_k = \int_{I_k} f(x)\, dx$, and we have $mp_k \geq c_0$. Further, [2, p. 10],

$$\begin{aligned} E \int_0^1 f(x) \log \frac{f(x)}{f_{n,m}(x)}\, dx & \\ \leq E \int_0^1 \frac{(f(x) - f_{n,m}(x))^2}{f_{n,m}(x)}\, dx & \\ = E\left\{ 1_{\{f_{n,m} > c_0/2\}} \int_0^1 \frac{(f(x) - f_{n,m}(x))^2}{f_{n,m}(x)}\, dx \right\} & \\ + E\left\{ 1_{\{f_{n,m} \not> c_0/2\}} \int_0^1 \frac{(f(x) - f_{n,m}(x))^2}{f_{n,m}(x)}\, dx \right\}. & \end{aligned}$$

The first term in the sum is bounded from above by the first term in the right-hand side of the inequality in the lemma. As to the second term, using the inequality $(f(x) - f_{n,m})^2 \leq 2(f^2(x) + f_{n,m}^2(x))$ together with the bounds for f and $f_{n,m}$ we get the upper bound

$$2(n+1)(c_1^2 + m^2) P\{f_{n,m}(x) \not> c_0/2\} \tag{A.1}$$

for it. Now, any sequence x^n for which $\{f_{n,m}(x^n) \leq c_0/2\}$ for some k (and, hence, $\{f_{n,m}(x^n) \not> c_0/2\}$ is true) also satisfies $\frac{n_k}{n} m \leq \frac{c_0}{2} - \frac{m}{n}\left(1 - \frac{c_0}{2}\right) < \frac{c_0}{2}$, and since $mp_k \geq c_0$ it further satisfies $mp_k - \frac{n_k}{n} m > \frac{c_0}{2}$. Therefore, $P\{f_{n,m}(x) \not> c_0/2\} \leq P\{n_k m/n - mp_k \geq c_0/2\}$ for some k. Further, by Bennett's inequality, [15],

$$\begin{aligned} P\left\{\frac{n_k}{n} m - mp_k \geq cm/\sqrt{n}\right\} \leq P\left\{\left|\frac{n_k}{n} m - mp_k\right| \geq cm/\sqrt{n}\right\} & \\ \leq 2e^{-Bmc^2}, & \end{aligned} \tag{A.2}$$

for any c, where B is a positive constant, independent of m, n, and k. Putting $\frac{cm}{\sqrt{n}} = \frac{c_0}{2}$ the upper bound (A.1) with (A.2) gives the second term in the right-hand side of the inequality in Lemma 1, which completes the proof. □

Lemma 2: For every $f \in \mathcal{M}$

$$E_f \int_0^1 (f(x) - f_{n,m}(x))^2\, dx \leq \frac{2m}{n} + 4\frac{c_f^2}{m^2} + 4(1 + c_1^2)\frac{m^2}{n^2}, \tag{A.3}$$

where $c_f = \max_y |f'(y)|$.

Proof of Lemma 2: We have first

$$
\begin{aligned}
&E\int_0^1 (f(x) - f_{n,m}(x))^2\,dx \\
&= E\sum_{k=1}^{m}\int_{I_k}\left(\frac{(n_k+1)m/n}{1+m/n} - f(x)\right)^2 dx \\
&= E\sum_{k=1}^{m}\int_{I_k}\left(\frac{n_k m/n - mp_k}{1+m/n}\right. \\
&\quad \left. + \frac{mp_k + m/n}{1+m/n} - f(x)\right)^2 dx \qquad \text{(A.4)} \\
&\le E\sum_{k=1}^{m}\left[\int_{I_k} 2(n_k m/n - mp_k)^2\right. \\
&\quad \left. + 2\left(\frac{mp_k + m/n}{1+m/n} - f(x)\right)^2\right] dx.
\end{aligned}
$$

For the first term in the right-hand side of the inequality we get

$$
\begin{aligned}
&E\sum_{k=1}^{m}\int_{I_k} 2(n_k m/n - mp_k)^2\,dx \\
&= 2\sum_{k=1}^{m} E(n_k m/n - mp_k)^2 \times \frac{1}{m} \\
&= 2\sum_{k=1}^{m} m^2\,\frac{p_k(1-p_k)}{nm} \le \frac{2m}{n}. \qquad \text{(A.5)}
\end{aligned}
$$

For the second term in the right-hand side of the inequality in (A.4), which does not depend on x^n, we get, again using $(a+b)^2 \le 2(a^2 + b^2)$ and the upper bounds for the densities and their derivatives

$$
\begin{aligned}
&4\sum_{k=1}^{m}\left[\int_{I_k}(mp_k - f(x))^2\,dx + \frac{1}{m}\left(\frac{mp_k + m/n}{1+m/n} - mp_k\right)^2\right] \\
&\qquad\qquad \le 4\left(\frac{c_f^2}{m^2} + c_1^2\,\frac{m^2}{n^2}\right). \qquad \text{(A.6)}
\end{aligned}
$$

This completes the proof. □

Returning to the proof of Theorem 1, we verify that the right-hand side of the inequality in Lemma 2 is minimized for m approximately $n^{1/3}$. At any rate, with the choice $m_{n-1}^* = \lceil n^{1/3}$ for m the right-hand side is bounded from above by

$$2(1 + 2c_f^2)n^{-2/3} + O(n^{-4/3}).$$

By Lemma 1, then, we get with this bound

$$
\begin{aligned}
E_f\int_0^1 f(x)\log\frac{f(x)}{f_{t,m_{t-1}^*}(x)}\,dx &\le A_f t^{-2/3} \\
&\quad + O(t^{5/3}e^{-Bt^{4/3}}) + O(t^{-4/3}),
\end{aligned}
$$

where $A_f = 2(1 + 2c_f^2)$, proving (3.5). Further, with the notation (3.3) and the subsequent convention

$$
\begin{aligned}
\frac{1}{n}E\log\frac{f_n(x^n)}{f_n^*(x^n)} &= \frac{1}{n}\sum_{t=1}^{n} E\int_0^1 f(x)\log\frac{f(x)}{f_{t,m_{t-1}^*}(x)}\,dx \\
&\le \frac{1}{n}(B_f n^{1/3} + O(1)),
\end{aligned}
$$

for a constant B_f, which concludes the proof. □

Appendix B

We begin with an estimate of the rate with which $\hat{f}_i = m_n\dfrac{n_i}{n}$ converges to f_i in probability, where m_n is given in (3.6). By Bennett's inequality, (A.2), we get

$$P\left\{\bigcup_i |\hat{f}_i - f_i| \ge \frac{cm_n}{\sqrt{n}}\right\} \le 2m_n e^{-Bm_n c^2}. \qquad \text{(B.1)}$$

We wish to select c so that the right-hand side gets smaller than some number α, $0 < \alpha < 1$, say $\alpha = 2/3$, to be specific. This is true with the choice

$$c = \frac{1}{\sqrt{B}}\sqrt{\frac{\log(3m_n)}{m_n}}.$$

This value, in turn, determines the threshold $cm_n/\sqrt{n}$ in (B.1), for which we pick

$$r_n = (3B)^{-1/2}n^{-1/3}, \qquad \text{(B.2)}$$

which for large n is slightly larger than what required. With these choices (B.1) gives the inequality

$$P\left\{\bigcup_i |\hat{f}_i - f_i| \ge r_n\right\} \le 2/3. \qquad \text{(B.3)}$$

Next, we generalize an inequality in [17], [21], valid for parametric classes of models, which links the Kullback distance and the estimation rate for parameters. Consider a partition of the compact set Ω_n into m_n-dimensional hypercubes of edge length r_n, given in (B.2). Write $\hat{\Omega}_n$ for the finite set of the centers of these cubes, and let $C(\theta)$ denote the cube with its center at θ. Further, let $X_n(\theta) = \{(x_1, \cdots, x_n) \mid \hat{\theta} \in C(\theta)\}$, where $\hat{\theta} = \left(\dfrac{m_n n_1}{n}, \cdots, \dfrac{m_n n_{m_n}}{n}\right)$. From (B.3)

$$P_n(\theta) \equiv P\{X_n(\theta)\} \ge 1/3. \qquad \text{(B.4)}$$

Next, consider the density function g_n, as specified in the theorem, and let $Q_n(\theta) = P_g(X_n(\theta))$ denote the probability mass g_n assigns to the set $X_n(\theta)$. Notice that for any two distinct points in $\hat{\Omega}_n$ these sets of strings are disjoint. The ratio $f^n(x^n)/P_n(\theta)$ defines a distribution on $X_n(\theta)$, as does of course $g_n(x^n)/Q_n(\theta)$. By the nonnegativity of the mutual information, applied to these two distributions, we get

$$T_n(\theta) \equiv \int_{X_n(\theta)} f^n(x^n)\log\frac{f^n(x^n)}{g_n(x^n)}\,dx^n \ge P_n(\theta)\log\frac{P_n(\theta)}{Q_n(\theta)}. \qquad \text{(B.5)}$$

Also,

$$E_f\log\frac{f^n(x^n)}{g_n(x^n)} \ge T_n(\theta) - 1, \qquad \text{(B.6)}$$

where we used the inequality $\log z \ge 1 - 1/z$ for $z = f^n(x^n)/g_n(x^n)$, whenever $g_n(x^n) > 0$, to get

$$
\begin{aligned}
&\int_{\overline{X}_n(\theta)} f^n(x^n)\log\left[f^n(x^n)/g_n(x^n)\right] dx^n \\
&\qquad\qquad \ge Q_n(\theta) - P_n(\theta) > -1;
\end{aligned}
$$

here $\bar{X}$ denotes the complement of X. Notice that for each hypercube with its center θ in $\hat{\Omega}_n$ we have a set of density functions associated with that θ, any one of which by (B.4) assigns a $O(1)$ probability to the cube. Let $f_{n,\theta}$ denote one of them. Now, if a single density function g_n succeeds in approximating all these density functions $f_{n,\theta}$ well, as θ runs through all the centers, then the probability mass it assigns to each cube cannot go to zero as n grows. But since there is just so much probability mass available for this density function, there can be only so many cubes where the approximation can be very good. A quantitative evaluation of the number of the cubes, where a very good approximation is possible, is what gives the desired inequality.

Putting the just sketched plan to work let K be a positive number and let $A_{g,n}$ be the set of θ's such that the left-hand side of (B.5) satisfies the inequality

$$\frac{1}{n} T_n(\theta) < Kn^{-2/3}, \tag{B.7}$$

which means that for these θ's we are trying to force the codelength $-\log Q_n(\theta)$ to be close to the ideal $-\log P_n(\theta)$. This with (B.4) and (B.5) implies

$$-\log Q_n(\theta) < T_n(\theta)\left[P_n^{-1}(\theta) - \frac{\log P_n(\theta)}{T_n(\theta)}\right] < 2Kn^{1/3}, \tag{B.8}$$

which holds for $\theta \in A_{g,n}$ and for all sufficiently large n. This gives a lower bound for $Q_n(\theta)$, which we write as $q_n(\theta)$ for short; in other words, forcing (B.7) causes us to "spend" a certain minimum amount of the available probability mass. Next, let $B_{g,n}$ be the smallest set of the centers of the hypercubes which cover $A_{g,n}$; and let ν_n be the number of the elements in $B_{g,n}$. Since the sets $X_n(\theta)$, $\theta \in \hat{\Omega}_n$, are disjoint, we have

$$1 \geq \sum_{\theta \in B_{g,n}} Q_n(\theta) \geq \nu_n q_n, \tag{B.9}$$

which with (B.8) gives the inequality $\log \nu_n < 2Kn^{1/3}$. The volume of $A_{g,n}$ is then bounded from above by

$$V(A_{g,n}) \leq \nu_n r_n^{m_n}, \tag{B.10}$$

which holds for all sufficiently large n.

We next calculate a lower bound for the volume of the m_n-dimensional set $\Omega_n = \{\theta_f = f_1, \cdots, f_{m_n} | f \in \mathcal{M}\}$. To do it, let $C = \frac{1}{3}\min\{1 - c_0, c_1 - 1\}$, and consider the set

$$D = \left\{\theta \in R^{m_n} \middle| \sum_{i=1}^{m_n} \theta_i = m_n, \ |\theta_j - 1| < C, \text{ all } j, \right\}$$

which has the volume $(2C)^{m_n}$. This will be the sought-for lower bound after we show that D is a subset of Ω_n. Hence, we must demonstrate that for each $\theta = (\theta_1, \cdots, \theta_{m_n})$, $\sum_i \theta_i = m_n$, in D there is a density function in $\mathcal{M}$ such that $f_i = \theta_i$. In fact, define a density function f_θ successively on $I_0, \cdots, I_{m_n-1}$, where $I_{i-1} = \left(\frac{i-1}{m_n}, \frac{i}{m_n}\right]$, as follows:

$$f_\theta(x) = \theta_i + (\theta_i - \theta_1)\sin\left[2\pi m_n\left(x - \frac{i-1}{m_n}\right) - \frac{\pi}{2}\right], \tag{B.11}$$

for $x \in I_{i-1}$. By a direct verification

$$m_n \int_{I_{i-1}} f_\theta(x)\, dx^n \equiv f_i = \theta_i, \tag{B.12}$$

so that the integral over the unit interval is unity. Further, the values of this function at the bin boundaries all equal θ_1, so the function is continuous. Its derivative at the bin boundaries vanishes, and the function has a first derivative in the entire unit interval. Also,

$$|f_\theta'(x)| \leq \max_i |2\pi m_n(\theta_i - \theta_1)| < \infty. \tag{B.13}$$

Finally,

$$f_\theta(x) \leq \max_i |\theta_i| + \max_j |\theta_j - \theta_1| \leq C + 1 + 2C \leq c_1$$

$$f_\theta(x) \geq -\max_i |\theta_i| - \max_j |\theta_j - \theta_1| \geq 1 - C - 2C \geq c_0. \tag{B.14}$$

By (B.12)–(B.14) f_θ belongs to $\mathcal{M}$.

The volume of Ω_n is then at least as large as the volume of D, or $(2C)^{m_n}$. Hence, with (B.10) we get

$$\log \frac{V(A_{g,n})}{V(\Omega_n)} \leq \log \nu_n - m_n \log(2C/r_n)$$

$$\leq n^{1/3}\left[2K - \frac{1}{3} - O\left(\frac{1}{\log n}\right)\right],$$

which goes to $-\infty$ for all K smaller than 1/6. Hence, for each such K we get by (B.7) $\frac{1}{n} T_n(\theta) \geq Kn^{-2/3}$, except for $\theta \in A_{g,n}$, and by (B.6) the claim in the theorem follows.

References

[1] A. R. Barron and T. M. Cover, "Minimum complexity density estimation," submitted for publication, 1989.

[2] A. R. Barron and C. Sheu, "Approximation of density functions by sequences of exponential families," submitted to *Ann. Statist.*, 1988.

[3] L. D. Davisson, "Minimax noiseless universal coding for Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 211–215, Mar. 1983.

[4] L. D. Davisson, R. J. McEliece, M. A. Pursley, and M. S. Wallace, "Efficient universal noiseless source code," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 269–279, May 1981.

[5] A. P. Dawid, "Present position and potential developments: Some personal views, statistical theory, The prequential approach," *J. Royal Statist. Soc. A*, vol. 147, pt. 2, pp. 278–292, 1984.

[6] ——, "Prequential data analysis," to appear in *Issues and Controversies in Statistical Inference; Essays in Honor of D. Basu's 65th Birthday*, M. Ghosh and P. K. Pathak, Eds. 1989.

[7] L. Devroye, *A Course in Density Estimation*. Boston, MA: Birkhauser.

[8] P. Elias, "Universal codeword sets and representations of the integers," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 194–203, Mar. 1975.

[9] L. Gerencse'r, "On a class of mixing processes," *Stochastics*, vol. 26, pp. 165–191, 1989.

[10] L. Gerencse'r and J. Rissanen, "A prediction bound for Gaussian ARMA processes," *Proc. 25th CDC*, vol. 3, Athens, Greece, 1986, pp. 1487–1490.

[11] P. Hall and E. J. Hannan, "On stochastic complexity and nonparametric density estimation," *Biometrika*, vol. 75, pp. 705–714, 1988.

[12] E. J. Hannan, A. J. McDougall, and D. S. Poskitt, "Recursive estimation of autoregressions," *J. Roy. Statist. Soc., Ser. B*, vol. 51, no. 2, pp. 217–233, 1989.

[13] E. M. Hemerly and M. H. A. Davis, "Strong consistency of the PLS criterion for order determination of autoregressive processes," *Ann. Statist.*, vol. 17, no. 2, pp. 941-946, 1989.

[14] R. E. Krishevskii and V. K. Trofimov, "The performance of universal encoding," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 199-207, Mar. 1981.

[15] D. Pollard, *Convergence of Stochastic Processes*. New York: Springer-Verlag, 1984, 215 pages.

[16] J. Rissanen, "A universal prior for integers and estimation by minimum description length," *Ann. Statist.*, vol. 11, no. 2, pp. 416-431, 1983.

[17] ____, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 629-636, July 1984.

[18] ____, "Stochastic complexity and modeling," *Ann. Statist.*, vol. 14, pp. 1080-1100, 1982.

[19] ____, "Complexity of strings in the class of Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 526-532, July 1986.

[20] ____, "A predictive least squares principle," *IMA J. Math. Contr. and Inform.*, vol. 3, no. 2-3, pp. 211-222, 1986.

[21] ____, *Stochastic Complexity in Statistical Inquiry*. Singapore: World Scientific Publ. Co., 1989, 175 pages.

Ann. Inst. Statist. Math.
Vol. 45, No. 1, 35–54 (1993)

MODEL SELECTION AND PREDICTION: NORMAL REGRESSION*

T. P. SPEED[1] AND BIN YU[2]

[1] *Department of Statistics, University of California at Berkeley, CA 94720, U.S.A.*
[2] *Department of Statistics, University of Wisconsin-Madison, WI 53706, U.S.A.*

(Received September 27, 1991; revised April 27, 1992)

Abstract. This paper discusses the topic of model selection for finite-dimensional normal regression models. We compare model selection criteria according to prediction errors based upon prediction with refitting, and prediction without refitting. We provide a new lower bound for prediction without refitting, while a lower bound for prediction with refitting was given by Rissanen. Moreover, we specify a set of sufficient conditions for a model selection criterion to achieve these bounds. Then the achievability of the two bounds by the following selection rules are addressed: Rissanen's accumulated prediction error criterion (APE), his stochastic complexity criterion, AIC, BIC and the FPE criteria. In particular, we provide upper bounds on overfitting and underfitting probabilities needed for the achievability. Finally, we offer a brief discussion on the issue of finite-dimensional vs. infinite-dimensional model assumptions.

Key words and phrases: Model selection, prediction lower bound, accumulated prediction error (APE), AIC, BIC, FPE, stochastic complexity, overfit and underfit probability.

1. Introduction

This paper discusses the topic of model selection for prediction in regression analysis. We compare model selection criteria according to the quality of the predictions they give. Two types of prediction errors, prediction with and without refitting, will be considered. A lower bound on the former type of error was given by Rissanen (1986*a*), and in this paper (Section 2) we provide a lower bound for the latter. Moreover, also in Section 2 we specify a set of sufficient conditions for a model selection criterion to achieve these bounds. Roughly speaking, to achieve these bounds, a model selection criterion has to be consistent and satisfy some underfitting and overfitting probability constraints. Section 3 concerns the following model selection criteria: Rissanan's predictive "minimum description length"

* Support from the National Science Foundation, grant DMS 8802378 and support from ARO, grant DAAL03-91-G-007 to B. Yu during the revision are gratefully acknowledged.

(accumulated prediction error, or predictive least squares), stochastic complexity, AIC, BIC and FPE. We consider bounds on their overfitting and underfitting probabilities, and therefore their achievability of the prediction lower bounds. In particular, the selection rule based on the accumulated prediction error and BIC achieve the two prediction lower bounds, but AIC does not unless the largest model considered is the true model.

Detailed proofs are relegated to the last section 5. All of our results are obtained under the assumption that a finite dimensional normal model generates the data under discussion. This contrasts greatly with most previous discussions, notably Shibata (1983*a*, 1983*b*) and Breiman and Freedman (1983), where the "true" model is infinite-dimensional. More discussion on finite-dimensional models vs. infinite-dimensional models can be found in Section 4.

2. Model selection and prediction in regression

In order to compare model selection procedures a number of choices need to be made; these can be critical. Two objectives of regression analysis are data description and prediction. The focus will be on the second, prediction.

Write $y = (y_1, \ldots, y_n)'$ for the n-dimensional column vector of observations, and $X = (x_{ij})$ for the $n \times K$ matrix of covariates or regressors. Inner products and squared norms are denoted by $\langle y, z\rangle = \sum y_t z_t$ and $|y|^2 = \langle y, y\rangle$, respectively. For $1 \le t \le n$, $1 \le k \le K$, denote by $y(t)$ and $X_k(t)$ that $t \times 1$ and $t \times k$ subvector and submatrix of y and X respectively, consisting of the first t rows and, in the case of X, of the first k columns. The subscript k or the parenthetical t will be omitted when they are clear from the context, or when $k = K$ or $t = n$. The t-th row of X is denoted by x_t' and the j-th column by ξ_j, whilst $x_t'(k)$ denotes the t-th row of X_k, with an analogous convention regarding the dropping of t or k. Parameter vectors are denoted by $\beta = (\beta_1, \ldots, \beta_k)'$, written $\beta(k)$ when necessary.

The class of models to be discussed will be denoted by $\{M_k : 1 \le k \le K\}$, where M_k is the model prescribing that y is $N(X_k\beta, \sigma^2 I)$ for some $\beta \in \boldsymbol{R}^k$ and $\sigma^2 > 0$. The number K of models is supposed known, and for the present discussion is held fixed as the sample size $n \to \infty$.

One framework for prediction involving regression is the following: (y_1, x_1), $(y_2, x_2), \ldots, (y_t, x_t)$ are given. The object is to predict y_{t+1} from x_{t+1}. An obvious approach is to select a model on the basis of the data available at time t, and predict y_{t+1} from this model with $t + 1$ replacing t. The response y_t at time t is known before predicting y_{t+1}, so this framework is called *prediction with repeated refitting* because it allows model selection at each time.

A quite different framework assumes the existence of an initial data set $\{(y_1, x_1), \ldots, (y_n, x_n)\}$, often called a training sample, and the regressors $\tilde{x}_1, \ldots, \tilde{x}_m$ associated with a number of other units, the requirement being to predict the corresponding responses $\tilde{y}_1, \ldots, \tilde{y}_m$. A familiar variant on this would be when the "prediction" is in fact the *allocation* of units into predetermined groups. The standard solution to this problem is to select a model on the basis of the initial data set, and then predict or allocate using the model selected. This framework will be called *prediction without refitting*.

MODEL SELECTION AND PREDICTION: NORMAL REGRESSION

In this section, the above two frameworks for prediction will be discussed in detail: lower bounds are given in each case, and sufficient conditions for a model selection procedure to achieve them are obtained. However, we leave to Section 3 the achievability of these lower bounds by common selection procedures.

2.1 *Prediction with repeated refitting*

A natural measure of the quality of a sequence of predictions in the repeated refitting framework is the sum

$$\text{APE}_n = \sum_{t=1}^{n} (y_t - \hat{y}_{t|t-1})^2 \tag{2.1}$$

where $\hat{y}_{t|t-1}$ denotes a predictor of y_t made on the basis of data up to and including time $t-1$, and any covariates available at time t. Model selection is thus permitted at every stage. The predictors which we consider below are $\hat{y}_{t|t-1} = x_t'\hat{\beta}_{t-1}(\hat{k}_{t-1})$, where $\hat{\beta}_{t-1}(\hat{k}_{t-1})$ is the least squares estimator based on model $M_{\hat{k}_{t-1}}$ at time t, and we will compare selection procedures leading to different $\hat{k}_s$ according to the average size of APE which is achieved for large n. For the purposes of our asymptotic analysis, it is not necessary to specify how we define $\hat{k}_t$ for $t \leq K$. In practice a number of reasonable approaches exist.

Our comparison is based upon a general inequality derived by Rissanen ((1986*a*), p. 1087). As in Sections 3 and 4 we denote by k^* the dimension associated with the true model, and $\hat{y}_{t|t-1}$ is *any* predictor of y_t which is a measurable function of $y_1, \ldots, y_{t-1}$, and $x_1, \ldots, x_t$. Although all our discussions so far have supposed that the error variance σ^2 is known and equal to unity, we will state the inequality for an arbitrary unknown σ^2. It asserts that for all k^* there is a Lebesgue null subset $A(k^*)$ of $\boldsymbol{R}^{k^*}$ such that for $\beta^* \notin A(k^*)$:

$$\liminf_{n\to\infty} \frac{\boldsymbol{E}_{\beta^*}\{\sum_1^n (y_t - \hat{y}_{t|t-1})^2 - n\sigma^2\}}{k^* \log n} \geq \sigma^2. \tag{2.2}$$

We say that the lower bound (2.2) is achieved by a model selection criterion if it is achieved by the corresponding predictor $y_{t|t-1}$.

We need some assumptions before we can state our results on the achievability of the prediction lower bound (2.2).

Assume (cf. Lai *et al.* (1979)) that there exists a positive definite $K \times K$ matrix $C = C_K$ such that

$$\lim_{N\to\infty} N^{-1} \sum_{t=M+1}^{M+N} x_t x_t' = C \tag{2.3}$$

uniformly in $M \geq 0$. If $M = 0$, the left-hand side is just $\lim_N N^{-1}X(N)'X(N)$. A further specialization gives $\lim_N N^{-1}X_k(N)'X_k(N) = C_k$, where C_k denotes the principal $k \times k$ submatrix of C. Assume also that

(2.4) $M_{k^*} \subseteq M_K$ is the smallest true model, and $\beta(k^*)$ the true parameter.

With this background we can now state the following result, proved in Section 5 below.

THEOREM 2.1. *Suppose that* (2.3) *and* (2.4) *hold and that* $\hat{k}_n$, *the dimension defined by a model selection procedure, satisfies*:

(i) $\mathrm{pr}(\hat{k}_n < k^*) = O(n^{-2}(\log n)^{-c})$ *as* $n \to \infty$, *for some* $c > 1$, *and*

(ii) $\mathrm{pr}(\hat{k}_n > k^*) \leq O((\log n)^{-\alpha})$ *as* $n \to \infty$, *for some* $\alpha > 2$.

Then the predictor $\hat{y}_{t|t-1} = x_t'\hat{\beta}_{t-1}(\hat{k}_{t-1})$ *achieves the lower bound* (2.2).

2.2 *Prediction without refitting*

Now let us suppose that we have observed $(y_1, x_1), \ldots, (y_n, x_n)$ and are required to predict the responses $\tilde{y}_1, \ldots, \tilde{y}_m$ corresponding to units with covariate vectors $\tilde{x}_1, \ldots, \tilde{x}_m$. In most discussions of this aspect of model selection, see e.g. Nishi (1984) and Shibata (1986*a*), $m = n$ and $x_i = \tilde{x}_i$, $1 \leq i \leq n$. Our framework is more realistic and although the general conclusions do not seem to be different from Shibata's, this was not obvious a priori.

Our predictors will all be of the form $\tilde{x}_u'\hat{\beta}(\hat{k})$, $u = 1, \ldots, m$ where $\hat{k}$ corresponds to a model selected on the basis of $\{(y_t, x_t) : t = 1, \ldots, n\}$. Given that $\hat{k} = k$, a natural measure of the quality of our set of m predictions is given by the *prediction error*

$$\mathrm{PE}(k) = \boldsymbol{E}\{|\tilde{y} - \tilde{X}_k\hat{\beta}(k)|^2 \mid y\} = m\sigma^2 + |\tilde{X}_{k^*}\beta(k^*) - \tilde{X}_k\hat{\beta}(k)|^2,$$

which averages over the new observations and conditions on the initial data. Following this line of thought, an equally natural measure of the effectiveness of the model selection procedure leading to $\hat{k}$ is $\boldsymbol{E}\{\mathrm{PE}(\hat{k}) - m\sigma^2\}$, where this time the expectation is over the possible initial data sets. What we now do is give some results on the behaviour of this quantity under a range of assumptions about $\tilde{X}$.

Our results are asymptotic in both n, the size of the initial sample, and m, the number of predictions being made. For this reason we need to supplement assumption (2.3) with an analogous, but weaker hypothesis concerning $\tilde{X}$ namely: that there exists a $K \times K$ positive definite $\tilde{C} = \tilde{C}_K$ such that

$$\lim_{M\to\infty} M^{-1} \sum_{u=1}^{M} \tilde{x}_u\tilde{x}_u' = \tilde{C}. \tag{2.5}$$

In the theorems which follow, $\hat{k} = \{\hat{k}_n\}$ is the index resulting from a procedure selecting from the models $\{M_k : 1 \leq k \leq K\}$.

The components of condition (B) below are defined by the partitioning

$$C_{k+1} = \begin{bmatrix} C_k & D_{k,k+1} \\ D_{k,k+1} & E_{k,k+1} \end{bmatrix},$$

where C_k, $k \leq K$ is defined following (2.3).

THEOREM 2.2. *Assume conditions* (2.3), (2.4) *and* (2.5). *Then under any of the following conditions*:

MODEL SELECTION AND PREDICTION: NORMAL REGRESSION

(A) $\lim_{n\to\infty} \mathrm{pr}(\hat{k}_n < k^*) > 0$;
(B) $C_k^{-1} D_{k,k+1} = \tilde{C}_k^{-1} \tilde{D}_{k,k+1}$, $k^* \le k < K$;
(C) $\hat{k} = \hat{k}_{\mathrm{FPE}_\alpha}$ *for a sequence* $\alpha = (\alpha_n)$ *with* $n^{-1}\alpha_n \to 0$ *where* FPE_α *is the Final Prediction Error criterion defined in Section* 3, *we may conclude*

$$\lim_{m,n\to\infty} nm^{-1}\boldsymbol{E}\{\mathrm{PE}(\hat{k}_n) - m\sigma^2\} \ge \mathrm{tr}\{C_{k^*}^{-1}\tilde{C}_{k^*}\}\sigma^2. \tag{2.6}$$

The proof will be given in Section 5. It can be seen from the proof of this theorem that there will be other "symmetric" selection rules other than FPE_α for which the conclusion holds.

The next question of interest is the following: what kinds of selection rules attain the lower bound (2.6)?

THEOREM 2.3. *The lower bound* (2.6) *is attained for any consistent selection rule whose underfitting probability* $\mathrm{pr}(\hat{k}_n < k^*)$ *is* $o(n^{-2})$ *as* $n \to \infty$.

3. APE, stochastic complexity, and FPE

In this section, we consider the achievability of the two lower bounds in Section 2 of some commonly-used model selection criteria. We derive upper bounds on the underfitting and overfitting probabilities of these criteria and then use Theorem 2.1 or Theorem 2.3.

First, we consider the criterion based upon accumulated (one-step) prediction errors (APE) (or predictive least squares). This criterion is the predictive MDL criterion introduced in Rissanen (1984, 1986*b*). Many authors have discussed this criterion as detailed in the remark after Theorem 3.1.

We now introduce the definition of APE. Only ordinary least squares estimates will be used. For $1 \le k \le K$, $k+1 \le s \le n$, write

$$\hat{\beta}_s(k) = (X_k(s)'X_k(s))^{-1}X_k(s)'y(s)$$

and $\hat{\beta}(k) = \hat{\beta}_n(k)$. All of the matrices $X_k(t)$ will be assumed to have rank k when $t > k$. The *recursive residuals*, also called one-step prediction errors, based on M_k are $e_t(k) = y_t - x_t(k)'\hat{\beta}_{t-1}(k)$. The ordinary residuals are $r_{t,n}(k) = y_t - x_t(k)'\hat{\beta}_n(k)$. The parenthetical k will be dropped if its value is clear from the context.

For any fixed $k \le K$, consider the accumulated squared prediction error $\mathrm{APE}_n(k) = \sum_{t=k+1}^n e_t(k)^2$. Obviously, $\mathrm{APE}_n(k)$ is the same as the prediction error with refitting (2.2) when the model M_k is fixed through time t.

Expression $\mathrm{APE}_n(k)$ will lead us to a model selection criterion: choose that k which minimizes $\mathrm{APE}_n(k)$ over all $k \le K$.

For the remainder of this section σ^2 is supposed known and so, for simplicity, is taken to be 1. This is possible because, unlike many model selection criteria, the one based on APE does not require knowledge or an estimate of σ^2. The numbers $\{b_k\}$ which appear in the following theorem are normalized limiting (squared) bias terms defined by

$$b_k = \mathrm{tr}\{(E_{k,k^*} - D'_{k,k^*}C_k^{-1}D_{k,k^*})\zeta(k)\zeta(k)'\}$$

where for $k < k^*$ the principal submatrices C_k and C_{k^*} of C are written

$$C_{k^*} = \begin{bmatrix} C_k & D_{k,k^*} \\ D_{k,k^*} & E_{k,k^*} \end{bmatrix},$$

and $\beta(k^*) = (\beta(k)' \mid \zeta(k)')'$ is the corresponding partitioning of $\beta(k^*)$. It is shown in Section 5 (Lemma 5.3) that $b_1 \geq b_2 \geq \cdots \geq b_{k^*-1} > 0$.

THEOREM 3.1. *Under assumptions* (2.3) *and* (2.4), *as* $n \to \infty$, *let* $\hat{k}_n$ *denote the dimension selected by minimizing* $\mathrm{APE}_n(k)$. *Then we have the following bounds*:

(i) $\mathrm{pr}(\hat{k}_n < k^*) \leq O(\exp(-bn))$ *as* $n \to \infty$, *for* $b = \min(b_{k^*-1}/3, b_{k^*-1}^2/18)$.

(ii) $\mathrm{pr}(\hat{k}_n > k^*) \leq O(n^{-1/6})$ *as* $n \to \infty$.

Remark. The upper bound in (i) shows the interplay between the bias term b_k and the sample size n; the product of them determines the underfitting probability, not the sample size n alone.

COROLLARY 3.1. *The lower bounds* (2.2) *and* (2.6) *are attained for the* APE *selection rule.*

PROOF. Straightforward from Theorems 2.1, 2.2 and 3.1.

Remark. (a) Convergence in probability of the APE selection rule was established by Rissanen (1986*b*) under essentially the same conditions as we have used here. Other writers who have suggested the use of APE or a related criterion to select regression models include Hjorth (1982) and Dawid (1984, 1992). The latter describes a generalization of the use of APE as the prequential approach to statistical analysis. (b) There is no doubt that our assumptions could be weakened, but the derivations of the same results are expected to be much more involved. In the context of time series, Wax (1988) derived the weak consistency of an analogous estimator of the order of an autoregressive process without the Gaussian assumption, and Hemerly and Davis (1989) strengthened it to the a.s. consistency. Moreover, Wei (1992) obtained the a.s. consistency and asymptotic expansions of APE under stochastic regression models.

Now we turn to selection rules based on the residual sum of squares, which is $\mathrm{RSS}_n(k) = \sum_1^n r_{t,n}(k)^2$ where the ordinary residuals $r_{t,n}(k)$ are defined above. When $\sigma^2 = 1$ in the regression models M_k the *final prediction error* (FPE) criterion is $\mathrm{FPE}_{\alpha_n}(k) = \mathrm{RSS}_n(k) + \alpha_n k$ where (α_n) is a sequence of positive numbers. For AIC, $\alpha_n \equiv 2$. For BIC (Schwartz (1978)), $\alpha_n = \log n$. When σ^2 is not known, we may replace it by its usual estimate from the largest model M_K. Our results should still hold in that case.

Rissanen (1986*a*) introduced stochastic complexity (SC) of a set of data relative to a model as variant of his MDL and PMDL expressions, and in many cases it is asymptotically equivalent to the latter, whilst being easier to calculate. We refer to his paper for definitions of these quantities. For our regression models with error variance equal to unity, SC takes a particularly simple form if the prior distribution for the parameter $\beta(k)$ is taken to be $N(0, \tau I_k)$ where $\tau > 0$ is a scale parameter, $k = 1, \ldots, K$. A simple calculation yields the expression

$$\text{(3.1)} \qquad \mathrm{SC}_n(k) = \frac{1}{2} n \log 2\pi + \frac{1}{2} \log \det(I_n + \tau X_k X_k') + \frac{1}{2} y'(I_n + \tau X_k X_k')^{-1} y.$$

From Lemma 5.5 in Section 5 we see that as $n \to \infty$,

$$\mathrm{SC}_n(k) - \frac{1}{2} n \log 2\pi = k \log n + \mathrm{RSS}_n(k) + O(1) \qquad \text{a.s.}$$

and so any discussion of model selection based upon stochastic complexity is subsumed under that of BIC.

The FPE criterion has been discussed by Akaike (1970, 1974), Bhansali and Downham (1977), Atkinson (1980), and Shibata (1976, 1986*a*) amongst others. Geweke and Meese (1981) discuss the problem quite generally, but with random regressors, whilst Kohn (1983) considers selection in general parametric models. Shibata (1984) may be consulted for further details on some cases of FPE. The consistency of FPE's, with α_n's satisfying $\lim n^{-1}\alpha_n = 0$ and $\underline{\lim}(2 \log \log n)^{-1}\alpha_n > 1$, was established in a time-series context by Hannan and Quinn (1979). Moreover, the equivalence of BIC and APE has been shown by Hannan *et al.* (1989) for the finite-dimensional autoregressive models and by Wei (1992) for finite-dimensional stochastic regression models.

THEOREM 3.2. *Let $\hat{k}_n$ denote the dimension selected by* FPE_{α_n} *for some sequence α_n such that $n^{-1}\alpha_n \to 0$ as $n \to \infty$. Then*

(i) *$\hat{k}_n$ overfits with probability approaching unity as $n \to \infty$. More precisely, for any constant $0 < b < b_{k^*-1}/4$, $\mathrm{pr}(\hat{k}_n < k^*) \leq O(\exp(-bn))$ as $n \to \infty$.*

(ii) *If $k^* < K$, and $\liminf(2 \log \log n)^{-1}\alpha_n > 2$, we have, for some $\gamma > 2$, $\mathrm{pr}(\hat{k}_n > k^*) \leq O((\log n)^{-\gamma})$ as $n \to \infty$.*

We omit the proof of this theorem in this paper because Woodroofe (1982) and Haughton (1989) contain smilar bounds for BIC under more general models. Moreover, a lower bound, instead of an upper one, on the overfit probability (ii) is given in the Appendix II of Merhav *et al.* (1989) for BIC. Their result suggests that the overfit probability of BIC tends to zero slower than exponentially as n tends to infinity.

COROLLARY 3.2. (i) *The selection rules defined by* BIC *and* SC *all lead to predictors which achieve the lower bounds* (2.2) *and* (2.6);

(ii) *If* $\lim(2 \log \log n)^{-1}\alpha_n < 1$, *the selection rules defined by* FPE_{α_n} *do* not *achieve the lower bounds* (2.2) *and* (2.6) *unless $k^* = K$; in particular,* AIC *does* not *achieve the lower bounds unless $k^* = K$.*

4. Discussion

The results presented seem to suggest that if prediction is part of the objective of a regression analysis, then model selection carried out using APE, BIC, SC or an equivalent procedure has some desirable properties. Of course there is a qualification: in deriving these theorems we have assumed that the model generating our data is (i) fixed throughout the asymptotics; (ii) finite-dimensional; and (iii) belongs to the class of models being examined. Before commenting on these assumptions, let us see that our theorems are at least in general agreement with a number of analyses and simulations in the literature. The first paper to point out clearly that consistent model selection gives better predictions seems to be Shibata (1984), although he does not emphasize this conclusion. Atkinson's (1980) results also suggest the conclusion we have reached, but again this is not emphasized. The simulation results of Clayton *et al.* (1986) led them to conclude "that if the 'true' or 'approximately true' model is included among the alternatives considered, all reasonable model selection procedures will possess rather similar predictive capabilities". We feel that this conclusion is more a reflection of the limited scope of the simulations conducted rather than the true state of affairs. Indeed a close examination of the sample sizes and models these authors studied suggests that there was little opportunity for the procedures (not the models) to be distinguished, as far as the squared prediction error of the resulting choices is concerned. More recently, Rissanen (1989) reported clear differences between cross validation and SC, and to the extent that cross-validation and AIC perform similarly, Stone (1977), this is explained by Corollary 3.2.

Shibata (1981, 1983*a*, 1983*b*, 1984, 1986*a*, 1986*b*) presents a number of theorems demonstrating the optimality of AIC or other forms of FPE_{α_n} with bounded sequences (α_n), as well as arguments rebutting the criticisms that such procedures are unsatisfactory by virtue of their inconsistency under assumptions (i), (ii) and (iii). Shibata (1981), and Breiman and Freedman (1983) using random regressors, suppose the true model to be *infinite*-dimensional rather than *finite*-dimensional. Shibata (1981) also offers an optimality result for AIC valid under a "moving truth" assumption.

Clearly, the prediction optimality of BIC and its analogues like APE depend on the assumption that the true model is finite-dimensional, i.e., the bias term $b_k = 0$ for $k \geq k^*$. When the true model is assumed to be infinite-dimensional, i.e., $b_k > 0$ for all k, Breiman and Freedman (1983) showed that AIC's equivalent is optimal in terms of one-step further prediction. We now show by the following three simple examples that the decay rate of the bias term plays a determining role in the battle of AIC vs. BIC.

For simplicity, let us take the framework of Breiman and Freedman (1983) where an infinite-dimensional model with Gaussan $N(0,1)$ independent regressors is assumed with the error variance $\sigma^2 = 1$. Then the one-step ahead prediction error for the $(n+1)$-st observation based on model M_k is roughly $\mathrm{PE}(k) = b_k + kn^{-1}$. Moreover, AIC approximately minimizes $b_k + kn^{-1}$, while BIC minimizes $b_k + kn^{-1}\log n$. By the result of Breiman and Freedman (1983), asymptotically, $\mathrm{PE}(\hat{k}_{\mathrm{BIC}})/\mathrm{PE}(\hat{k}_{\mathrm{AIC}}) \geq 1$, where $\hat{k}_{\mathrm{AIC}}$ is the model selected by AIC, and similarly for $\hat{k}_{\mathrm{BIC}}$.

Example 1. Assume $b_k = k^{-\alpha}$. Straightforward calculation shows that, as $n \to \infty$, $\mathrm{PE}(\hat{k}_{\mathrm{BIC}})/\mathrm{PE}(\hat{k}_{\mathrm{AIC}}) \to \infty$.

Example 2. Assume $b_k = e^{-k}$. Then as $n \to \infty$, $\mathrm{PE}(\hat{k}_{\mathrm{BIC}})/\mathrm{PE}(\hat{k}_{\mathrm{AIC}}) \to 2$.

Example 3. Assume $b_k = e^{-e^k}$. Then as $n \to \infty$, $\mathrm{PE}(\hat{k}_{\mathrm{BIC}})/\mathrm{PE}(\hat{k}_{\mathrm{AIC}}) \to 1$.

To summarize, as the decay rate of the bias term increases, the prediction performance of BIC catches up with that of AIC. And, as we have seen, BIC out-performs AIC when $b_k = 0$ for $k > k^*$, i.e. when the model is finite.

Finally, all three of APE, BIC and SC derive from general approaches to the model selection problem and have extensions to situations where one or more of (i), (ii) and (iii) are dropped, see Sawa (1978) for some remarks about this situation. When something is known about these extensions, it will be of interest to compare them with AIC or, more generally FPE_{α_n}.

5. Proofs

Most of the arguments given below are straightforward. We have tried to be explicit wherever possible, and have included some proofs which may be found elsewhere in order to keep this paper self-contained.

The proofs are presented in the following order: Theorem 3.1, Corollaries 3.1 and 3.2, Theorem 2.2, Theorem 2.3 and Theorem 2.1. We continue to use the notation introduced in Section 2 above. It is straightforward to show

LEMMA 5.1. *For $k < s < t \leq n$ and $c \in R(X_k(t))$, we have* $\mathrm{cov}(e_{s+1}(k), c'y(t)) = 0$.

It follows from the lemma that

COROLLARY 5.1. (a) *For all $k < s < t \leq n$, we have* $\mathrm{cov}(e_s(k), e_t(k)) = 0$. (b) *For all $k < t \leq n$, and $c \in R(X_k)$,* $\mathrm{cov}(e_t(k), c'y) = 0$.

Let us write $\lambda_t(k) = \boldsymbol{E}\{e_t(k)\}$ and $\mu_t(k) = \mathrm{Var}\{e_t(k)\} - 1$, $\epsilon_t = y_t - \boldsymbol{E}\{y_t\}$ and $H_n(k) = X_n(k)(X_n(k)'X_n(k))^{-1}X_n(k)'$, and define the following quantities:

$$V_n(k) = \sum_{t=k+1}^{n} \mu_t(k), \qquad B_n(k) = \sum_{t=k+1}^{n} \lambda_t(k)^2, \qquad N_n(k) = |H_n(k)\epsilon|^2,$$

$$N_n^{\dagger}(k) = \sum_{t=k+1}^{n} \mu_t(k) \left[\frac{(e_t(k) - \lambda_t(k))^2}{\mu_t(k) + 1} - 1 \right],$$

$$B_n^{\dagger}(k) = 2 \sum_{t=k+1}^{n} (e_t(k) - \lambda_t(k))\lambda_t(k).$$

It is clear from the proof of the result we state shortly that V is a *variance* term, B is a *bias* term, and N is a *noise* term, whilst $N^\dagger$ is a second noise term and $B^\dagger$ a part-noise part-bias term.

LEMMA 5.2. *With the above notation*

$$\sum_{t=k+1}^{n} e_t(k)^2 - \sum_{t=1}^{n} \epsilon_t^2 = V_n(k) + B_n(k) - N_n(k) + B_n^\dagger(k) + N_n^\dagger(k). \tag{5.1}$$

PROOF. It follows from Corollary 5.1 that $\{e_{k+1}(k), \ldots, e_n(k)\}$ are pairwise uncorrelated, and uncorrelated with $c'y$ for all $c \in R(X_k)$. Thus we can make an orthogonal transformation and obtain

$$|\epsilon|^2 = |H(k)\epsilon|^2 + \sum_{t=k+1}^{n} \frac{[e_t(k) - \boldsymbol{E}\{e_t(k)\}]^2}{\operatorname{Var}\{e_t(k)\}}. \tag{5.2}$$

The lemma then follows from this equation and the comparing two sides of (5.1). □

In the lemmas which follow, (2.1) and (2.2) will be assumed without comment. Moreover, to state our next result we need a little further notation. For $k < k^*$, write the principal $k \times k$ submatrix C_k of C given by (2.4) in the form

$$C_{k^*} = \begin{bmatrix} C_k & D_{k,k^*} \\ D'_{k,k^*} & E_{k,k^*} \end{bmatrix}$$

and we write $\beta(k^*) = (\beta(k)' \mid \zeta(k)')'$ and $X_{k^*}(n) = [X_k(n) \mid Z_k(n)]$.

LEMMA 5.3. $n^{-1}B_n(k) \to b_k$ *as* $n \to \infty$, *where*

$$b_k = \operatorname{tr}\{(E_{k,k^*} - D'_{k,k^*}C_k^{-1}D_{k,k^*})\zeta(k)\zeta(k)'\}$$

satisfies $b_1 \geq b_2 \geq \cdots \geq b_{k^*-1} > 0$.

PROOF. We begin by observing that for $k < k^*$, $\lambda_t(k) = A_k(t)'\zeta(k)$, where

$$A_k(t)' = z_t(k)' - x_t(k)'(X_k(t-1)'X_k(t-1))^{-1}X_k(t-1)'Z_k(t-1).$$

It follows that $\lambda_t(k)^2 = \operatorname{tr}\{A_k(t)A_k(t)'\zeta(k)\zeta(k)'\}$ and so

$$n^{-1}\sum_{t=k+1}^{n} \lambda_t(k)^2 = \operatorname{tr}\left\{n^{-1}\sum_{t=k+1}^{n} A_k(t)A_k(t)'\zeta(k)\zeta(k)'\right\}.$$

Using (2.4) and the notation introduced above, $t^{-1}X_k(t)'X_k(t) \to C_k$, $t^{-1}X_k(t)' \cdot Z_k(t) \to D_{k,k^*}$, and $t^{-1}Z_k(t)'Z_k(t) \to E_{k,k^*}$ as $t \to \infty$, and so it follows that

$$n^{-1}\sum_{t=k+1}^{n} A_k(t)A_k(t)' \to E_{k,k^*} - D_{k,k^*}C_k^{-1}D_{k,k^*}$$

as $n \to \infty$, giving the expression for b_k stated. The monotonicity of b_k can then be checked using the partial order of positive definite matrices. □

For the next lemma we need some notation paralleling that used in Lemma 5.2 above. Write $\bar{\lambda}_t(k) = \boldsymbol{E}\{r_t(k)\}$ and $\bar{B}_n(k) = \sum_1^n \bar{\lambda}_t(k)^2$. Furthermore, put $\bar{B}_n^\dagger(k) = 2\sum_1^n \bar{\lambda}_t(k)\epsilon_t$. By variants of the proofs of Lemmas 5.2 and 5.3 and by the law of iterative algorithm, we obtain

LEMMA 5.4.

$$\sum_1^n r_t(k)^2 - \sum_1^n \epsilon_t^2 = \bar{B}_n(k) - N_n(k) + \bar{B}_n^\dagger(k) \tag{5.3}$$

where for $k < k^*$, $n^{-1}\bar{B}_n(k) \to b_k$, *and* $\bar{B}_n^\dagger(k) = O((n \log\log n)^{1/2})$ *a.s. as* $n \to \infty$.

LEMMA 5.5. *In the notation introduced prior to equation* (3.1)

$$\begin{aligned} &\log\det(I_n + \tau X_k(n)X_k(n)') + y(n)'(I_n + \tau X_k(n)X_k(n)')^{-1}y(n) \\ &\quad = k\log n + \sum_1^n r_t(k)^2 + O(1) \qquad a.s.\ n \to \infty. \end{aligned}$$

PROOF. Straightforward from assumption (2.3) and Rao ((1973), p. 33). □

In the following lemmas we use the notation $\rho_k = \xi_{k+1} - X_k\gamma_k$, $\tilde{\rho}_k = \tilde{\xi}_{k+1} - \tilde{X}_k\gamma_k$ and $\eta_k = X_k(X_k'X_k)^{-1}\tilde{X}_k'\tilde{\rho}_k$, where $\gamma_k = (X_k'X_k)^{-1}X_k'\xi_{k+1}$. It is evident that γ_k is the regression coefficient of the $(k+1)$-st variable on the previous k, and so ρ_k and $\tilde{\rho}_k$ are essentially residuals when the current model is M_k, whereas η_k is part residual and part fitted value.

LEMMA 5.6.

$$\tilde{X}_{k+1}(X_{k+1}'X_{k+1})^{-1}X_{k+1}\epsilon = \tilde{X}_k(X_k'X_k)^{-1}X_k\epsilon + |\rho_k|^{-2}\langle\rho_k, \epsilon\rangle\tilde{\rho}_k.$$

PROOF. This is a straightforward consequence of the formula for the inverse of a partitioned matrix, see e.g. Rao ((1973), p. 33). □

If we write $N_{m,n}(k) = |\tilde{X}_k(X_k'X_k)^{-1}X_k'\epsilon|^2$ by analogy with the noise term introduced just before Lemma 5.2, then we have

COROLLARY 5.2.

$$N_{m,n}(k+1) = N_{m,n}(k) + 2|\rho_k|^{-2}\langle\eta_k, \epsilon\rangle\langle\rho_k, \epsilon\rangle + |\rho_k|^{-4}|\tilde{\rho}_k|^2\langle\rho_k, \epsilon\rangle^2.$$

Now let us write $\tilde{X}_{k^*} = [\tilde{X}_k \mid \tilde{Z}_k]$ and $\tilde{R}_k = \tilde{Z}_k - \tilde{X}_k(X_k'X_k)^{-1}X_k'Z_k$. Furthermore, for $k > k^*$, write

$$C_{k+1} = \begin{bmatrix} C_k & D_{k,k+1} \\ D_{k,k+1} & E_{k,k+1} \end{bmatrix}$$

and similarly for $\tilde{C}_{k+1}$. Finally, denote by $\Delta_{k,k+1}$ and Δ_k, the differences $\tilde{C}_k^{-1} \cdot \tilde{D}_{k,k+1} - C_k^{-1} D_{k,k+1}$ and $\tilde{C}_k^{-1} \tilde{D}_{k,k^*} - C_k^{-1} D_{k,k^*}$, respectively.

The following formulae bear a close resemblance to ones obtained in a similar context by Box and Draper (1959, 1963). There, however, the emphasis is on design: the choice of x vectors. It should be clear from the context whether or not $k \leq k^*$ is required to give a non-trivial result.

Lemma 5.7. *As $m, n \to \infty$ we have*

(i) $m^{-1} \tilde{X}_k' \tilde{R}_k \to \tilde{C}_k \Delta_k$.

(ii) $m^{-1} \tilde{R}_k' \tilde{R}_k \to \tilde{E}_k - \tilde{D}_{k,k^*}' \tilde{C}_k^{-1} \tilde{D}_{k,k^*} + \Delta_k' \tilde{C}_k^{-1} \Delta_k$.

(iii) $m^{-1} |\tilde{\rho}_k|^2 \to \tilde{E}_{k,k+1} - \tilde{D}_{k,k+1}' \tilde{C}_k^{-1} \tilde{D}_{k,k+1} + \Delta_{k,k+1}' C_k^{-1} \Delta_{k,k+1}$.

(iv) $n^{-1} |\rho_k|^2 \to E_{k,k+1} - D_{k,k+1}' C_k^{-1} D_{k,k+1}$.

(v) $nm^{-2} |\eta_k|^2 \to \Delta_{k,k+1}' \tilde{C}_k C_k^{-1} \tilde{C}_k \Delta_{k,k+1}$.

Proofs. These are all straightforward consequences of the relevant definitions. □

Next we extend some earlier notation, writing $B_{m,n}(k) = \mathrm{tr}\{\tilde{R}_k' \tilde{R}_k \zeta(k) \zeta(k)'\}$, and $S_{m,n}(k) = 2\langle \tilde{R}_k \zeta(k), \tilde{X}_k (X_k' X_k)^{-1} X_k' \epsilon \rangle$. Clearly the first term is the analogue of the bias term introduced prior to Lemma 5.2, and reduces to it if $m = n$ and $\tilde{X} = X$. For the definition of PE(k), see Section 2 above.

Lemma 5.8. *In the notation just introduced, we have*

$$\mathrm{PE}(k) - m\sigma^2 = B_{m,n}(k) + N_{m,n}(k) - S_{m,n}(k).$$

Proof. $\mathrm{PE}(k) - m\sigma^2 = |\tilde{X}_{k^*} \beta(k^*) - \tilde{X}_k \hat{\beta}(k)|^2$, where we may write

$$\begin{aligned} \tilde{X}_{k^*} \beta(k^*) - \tilde{X}_k \hat{\beta}(k) &= \tilde{X}_{k^*} \beta(k^*) - \tilde{X}_k (X_k' X_k)^{-1} X_k' (X_{k^*} \beta(k^*) + \epsilon) \\ &= (\tilde{Z}_k - \tilde{X}_k (X_k' X_k)^{-1} X_k' Z_k) \zeta(k) - \tilde{X}_k (X_k' X_k)^{-1} X_k' \epsilon. \end{aligned}$$

The result now follows upon taking the squared norm of this vector. □

Lemma 5.9. *As $m, n \to \infty$ we have*

(i) $m^{-1} B_{m,n}(k) \to \mathrm{tr}\{(\tilde{E}_k - \tilde{D}_{k,k^*}' \tilde{C}_k^{-1} \tilde{D}_{k,k^*} + \Delta_k' \tilde{C}_k^{-1} \Delta_k) \zeta(k) \zeta(k)'\}$.

(ii) $m^{-1} n \boldsymbol{E}\{N_{m,n}(k)\} \to \mathrm{tr}(\tilde{C}_k C_k^{-1})$.

(iii) $m^{-1} n N_{m,n}(k) = O(\log \log n)$ *a.s.*

(iv) $m^{-1} n S_{m,n}(k) \to 0$ *a.s. if* $\Delta_k = 0$.

(v) $m^{-1} S_{m,n}(k) = O((n^{-1} \log \log n)^{1/2})$ *a.s. if* $\Delta_k \neq 0$.

Proof. (i) is an immediate consequence of Lemma 5.7(iv); (ii) and (iii) are straightforward calculations; (iv) follows from the definitions, whilst (v) is a now-familiar form of the law of the iterated logarithm. □

MODEL SELECTION AND PREDICTION: NORMAL REGRESSION

PROOF OF THEOREM 3.1. (i) We begin by obtaining some probability inequalities concerning the terms in $\text{APE}_n(k)$, cf. Lemma 5.2. Since $N_n(k) = |H_n(k)\epsilon|^2$ is a chi-squared r.v.,

$$\text{pr}(N_n(k) > \beta_n) \leq O(\exp(-\beta_n)) \quad \text{as} \quad n \to \infty.$$

Similarly, $B_n^\dagger(k)$ is a sum of independent zero mean normal r.v.'s whose variance is $O(n)$, and so $\text{pr}(|B_n^\dagger(k)| > \gamma_n) \leq O(\gamma_n^{-1} n^{1/2} \exp(-\gamma_n^2/2n))$.

Finally, $W_n(k) = V_n(k) + N_n^\dagger(k)$ is a sum of $n-k$ independent squared normals, the t-th of which is scaled by $\mu_t(k)$, and so

$$\begin{aligned}\text{pr}(W_n(k) > \delta_n) \leq \exp(-\delta_n) \prod_{k+1}^{n} (1 - 2\mu_t(k))^{-1/2} &\leq \exp\left\{-\delta_n + \sum_{k+1}^{n} \mu_t(k)\right\} \\ = \exp\{-\delta_n + k \log n + o(\log n)\} & \\ \leq n^{k+1} \exp(-\delta_n), \quad \text{as} \quad n \to \infty. &\end{aligned}$$

We now put these inequalities together, select (β_n), (γ_n) and (δ_n), and obtain (i). For simplicity, we drop subscripts n where no confusion will result. If $k < k^*$,

$$\begin{aligned}\text{pr}(\hat{k} = k) &\leq \text{pr}\{\text{APE}(k) < \text{APE}(k^*)\} \\ &= \text{pr}\{B(k) - N(k) + W(k) + B^\dagger(k) \\ &\qquad < B(k^*) - N(k^*) + W(k^*) + B^\dagger(k^*)\} \\ &\leq \text{pr}\{W(k^*) \geq B(k) + B^\dagger(k) - N(k)\} \\ &\qquad\qquad \text{since} \quad W(k) > 0 \quad \text{and} \quad N(k^*) > 0, \\ &\leq \text{pr}\{W(k^*) \geq nb_k + o(n) - \gamma_n - B_n\} \\ &\quad + P\{N(k) > B_n\} + P\{|B^\dagger(k)| > \gamma_n\} \\ &\leq n^{k+1} \exp(-nb_k + o(n) + \gamma_n + \beta_n) \\ &\quad + O(\exp(-\beta_n)) + O(\gamma_n^{-1} n^{1/2} \exp(-\gamma_n^2/2n)).\end{aligned}$$

We now see that if $\beta_n = b_k n/3$ and $\gamma_n = b_k n/3$, the desired conclusion follows since b_k decreases as k increases to $k^* - 1$.

(ii) For the overfitting probability, we estimate $\text{pr}(\hat{k} = k)$ for $k > k^*$, noting that in this case $\text{APE}(k) = V(k) - N(k) + N^\dagger(k)$, i.e. the bias terms disappear. In this proof we bound $-N^\dagger(k)$ and $N^\dagger(k^*)$ from below by the same quantity, β_n say, and calculate the tail probability as in the first part of the proof. We find that

$$\begin{aligned}\text{pr}(N^\dagger(k) < -\beta_n) &= \text{pr}(-N^\dagger(k) > \beta_n) \\ &\leq \exp(-\beta_n) \prod_{k+1}^{n} \{(1 + 2\mu_t(k))^{-1/2} \exp \mu_t(k)\} \\ &\leq O(\exp(-\beta_n)).\end{aligned}$$

Similarly we have $\text{pr}(N^\dagger(k^*) > \beta_n) \leq O(\exp(-\beta_n))$, and since $N(k) - N(k^*)$ is a chi-squared r.v. on $k - k^*$ degrees of freedom,

$$\text{pr}(N(k) - N(k^*) > \gamma_n) \leq O(\gamma_n^{-1+(k-k^*)/2} \exp(-\gamma_n/2)).$$

Thus if $k > k^*$,

$$\begin{aligned}\mathrm{pr}(\hat{k} = k) &= \mathrm{pr}\{\mathrm{APE}(k) < \mathrm{APE}(k^*)\} \\ &= \mathrm{pr}\{V(k) - N(k) + N^\dagger(k) < V(k^*) - N(k^*) + N^\dagger(k^*)\} \\ &\le \mathrm{pr}\{V(k) - \beta_n - (N(k) - N(k^*)) < V(k^*) + \beta_n\} \\ &\quad + \mathrm{pr}\{N^\dagger(k) < -\beta_n\} + \mathrm{pr}\{N^\dagger(k) > \beta_n\} \\ &\le O(\gamma_n^{-1+(k-k^*)/2} \exp(-\gamma_n/2)) + 2O(\exp(-\beta_n)),\end{aligned}$$

where $\gamma_n = (k - k^*)\log n + o(\log n) - 2\beta_n$, since $V(k) = k \log n + o(\log n)$, and similarly for $V(k^*)$. If we take $\beta_n = \beta \log n$ for $\beta = 6^{-1}$, say, then we deduce that $\mathrm{pr}(\hat{k} > k) \le O(n^{-1/6})$. □

Corollary 3.2 can be shown by an argument similar to Theorems 2.1 and 2.3. Note that when the selection rule is not consistent, the inequality is sharp since the prediction error based on M_k for some $k > k^*$ is strictly larger than the one based on M_{k^*}, and underfitting does not cause any problem since all FPE's underfit with a probability vanishing exponentially fast (Theorem 3.1(i)).

Let $\{H_j : j = 1, \dots, n\}$ be a set of pairwise orthogonal rank 1 projectors summing to the identity, such that for all $k = 1, \dots, K$ we have $\sum_{p=1}^k H_p = H(k)$, where $R(H(k)) = R(X_k(n))$. Let $\epsilon = (\epsilon_i)$ be an n-tuple of iid $N(0,1)$ random variables, F any function of $|H_i\epsilon|^2$ for a fixed $i \in \{1, \dots, n\}$, and ξ, η fixed vectors.

LEMMA 5.10. $\boldsymbol{E}\{\langle x_i, H_i\epsilon\rangle F(|H_i\epsilon|^2)\} = 0$.

PROOF. The lemma is an immediate consequence of the symmetry of the normal distribution. □

COROLLARY 5.3. *Let f be a function of $|H_1\epsilon|^2, \dots, |H_k\epsilon|^2$. Then if $1 \le i, j \le k$, we have*

$$\begin{aligned}\boldsymbol{E}\{\langle \xi, H_i\epsilon\rangle f(|H_1\epsilon|^2, \dots, |H_k\epsilon|^2)\} = 0, \\ \boldsymbol{E}\{\langle \xi, H_i\epsilon\rangle \langle \eta, H_j\epsilon\rangle f(|H_1\epsilon|^2, \dots, |H_k\epsilon|^2)\} = 0.\end{aligned}$$

PROOF. The identities follow from the lemma by a suitable conditioning. □

In the lemma which follows we use the expressions ρ_k and η_k defined prior to Lemma 5.6 above.

LEMMA 5.11. *Let $\hat{k}_n$ denote the dimension selected by* FPE_{α_n} *and suppose that $l > k \ge k^*$. Then we have*

$$\lim_{m,n} m^{-1} n |\rho_k|^{-2} \boldsymbol{E}\{\langle \rho_k, \epsilon\rangle \langle \eta_k, \epsilon\rangle 1_{\{\hat{k}_n = l\}}\} = 0. \tag{5.4}$$

PROOF. We begin by replacing $\hat{k}_n$ by $\tilde{k}_n$, that k which minimizes $\mathrm{FPE}(k)$ over the range $\{k^*, k^* + 1, \dots, K\}$. From Theorem 3.2 we know that $\mathrm{pr}(\hat{k}_n \ne \tilde{k}_n) \to 0$ as $n \to \infty$.

Now recall the definition of FPE(k) and note that if $k < l$, FPE$(k) \leq$ FPE(l) if and only if $\sum_{k+1}^{l} |H_p\epsilon|^2 \leq (l-k)\alpha$. Thus the event $\{\tilde{k} = l\}$ is the intersection of the two events: $\{\sum_{p=h+1}^{l} |H_p\epsilon|^2 \geq (l-h)\alpha; k^* \leq h < l\}$ and $\{\sum_{p=l+1}^{h} |H_p\epsilon|^2 \leq (h-l)\alpha, l < h \leq K\}$ whose indicators we denote by f_l and g_l respectively. Our aim is to show that

$$\boldsymbol{E}\{\langle\rho_k, \epsilon\rangle\langle\eta_k, \epsilon\rangle f_l g_l\} = 0 \tag{5.5}$$

and then deduce the conclusion of the lemma.

Since $\eta_k \in R(X_k)$, we may write $\langle\eta_k, \epsilon\rangle = \sum_{i=1}^{k}\langle\eta_k, H_i\epsilon\rangle$. Similarly, $\rho_k \in R(X_k)^{\perp}$ and so $\langle\rho_k, \epsilon\rangle = \sum_{j=k+1}^{n}\langle\rho_k, H_j\epsilon\rangle$. Thus our interim objective will be achieved if we can prove that for all i, j, $1 \leq i \leq k$, $k+1 \leq j \leq n$, we have

$$\boldsymbol{E}\{\langle\eta_k, H_i\epsilon\rangle\langle\rho_k, H_j\epsilon\rangle f_l g_l\} = 0. \tag{5.6}$$

Note that f_l is a function of $\{|H_p\epsilon|^2 : k^* < p \leq l\}$ whilst g_l is a function of $\{|H_p\epsilon|^2 : l < p \leq K\}$, and so if $i \leq k^*$ or $j > k$, (5.6) is trivially zero. If we take the case $k^* \leq i, j \leq l$, we can split off g_l by independence and use Corollary 5.3 to get the conclusion. Similarly if $k^* \leq i \leq l$ and $l < j \leq K$, we can again use independence this time splitting off $\langle\eta_k, H_i\epsilon\rangle f_l$, and again getting zero by the same corollary. Thus (5.6) and hence (5.5) are established.

The proof is completed by noting that $\lim_{m,n} m^{-1}n|\rho_k|^{-2}\boldsymbol{E}|\langle\eta_k, \epsilon\rangle\langle\rho_k, \epsilon\rangle|$ is finite, and so we can combine the result $\mathrm{pr}(\tilde{k}_n \neq \hat{k}_n) \to 0$ as $n \to \infty$ with (5.5) to obtain (5.4). □

Proof of Theorem 2.2. We obtain (2.6) under each of the three conditions in turn; in all cases making use of Lemmas 5.8 and 5.9. Then by Lemma 5.8, the left-hand side of (2.6) will be $O(n)$ as $m, n \to \infty$, since the bias terms $nB_{m,n}(k)$ for $k < k^*$ are not all eliminated, and these are $O(n)$ as $m, n \to \infty$, and cannot be canceled by either of the noise terms. Thus (2.6) is trivially true. Now let us assume (B). By virtue of the result just established, we may also suppose that $\mathrm{pr}(\hat{k}_n < k^*) \to 0$ as $n \to \infty$. Otherwise we make no assumptions concerning the selection procedure $\hat{k}$. On the set $\{\hat{k} > k^*\}$, $B_{m,n}(\hat{k}) = S_{m,n}(\hat{k}) = 0$, and so $\mathrm{PE}(\hat{k}) - m\sigma^2 = N_{m,n}(\hat{k})$.

Our proof begins by observing that

$$\begin{aligned}
&\lim_{m,n} nm^{-1}\boldsymbol{E}||\rho_k|^{-2}\langle\eta_k, \epsilon\rangle\langle\rho_k, \epsilon\rangle| \\
&\qquad \leq \lim_{m,n} nm^{-1}|\rho_k|^{-2}\{\boldsymbol{E}\langle\eta_k, \epsilon\rangle^2\boldsymbol{E}\langle\rho_k, \epsilon\rangle^2\}^{1/2} \\
&\qquad = \lim_{m,n} nm^{-1}|\rho_k|^{-2}\{|\eta_k|^2|\rho_k|^2\}^{1/2},
\end{aligned}$$

and this limit is zero by Lemma 5.7 and (B).

Repeated application of this result and Corollary 5.2 give a series of inequalities, which imply that for $k > k^*$:

$$\lim_{m,n} nm^{-1}\boldsymbol{E}\{N_{m,n}(k)1_{\{\hat{k}=k\}}\} \geq \lim_{m,n} nm^{-1}\boldsymbol{E}\{N_{m,n}(k^*)1_{\{\hat{k}=k\}}\},$$

whence $\lim_{m,n} nm^{-1}\boldsymbol{E}\{N_{m,n}(\hat{k})1_{\{\hat{k}\geq k^*\}}\} \geq \lim_{m,n} nm^{-1}\boldsymbol{E}\{N_{m,n}(k^*)1_{\{\hat{k}\geq k^*\}}\}$. Since $\text{pr}(\hat{k}_n \geq k^*) \to 1$ as $n \to \infty$, and $N_{n,m}(k^*) \geq 0$, $\lim_{m,n} nm^{-1}\boldsymbol{E}\{N_{m,n}(k^*)\} = \text{tr}\{\tilde{C}_{k^*}C_{k^*}^{-1}\}$ implies (2.6) in case (B).

Finally we consider case (C). The proof goes as for case (B), and in particular the selection rules $\hat{k}$ based on FPE_{α_n} for α_n such that $n^{-1}\alpha_n \to 0$ as $n \to \infty$, overfit with probability approaching unity by Theorem 3.2. The chain of inequalities leading to the final conclusion is also true, but this time the individual steps are justified by Theorem 3.1, and the proof is completed exactly as it was in case (B). Any other selection rule for which the same symmetry argument is valid also has the lower bound. □

PROOF OF THEOREM 2.3. (i) We begin by proving that the underfitting contribution to the left-hand side of (2.6) is asymptotically negligible. This follows from the readily checked fact that when $k < k^*$, $nm^{-1}\boldsymbol{E}\{(\text{PE}(k) - m\sigma^2)\} \leq O(n)$ as $m, n \to \infty$. Thus for all $k < k^*$,

$$nm^{-1}\boldsymbol{E}\{(\text{PE}(\hat{k}) - m\sigma^2)1_{\{\hat{k}=k\}}\} \leq O(n)\sqrt{\text{pr}(\hat{k}_n = k)} \to 0$$

as $m, n \to \infty$, and so $nm^{-1}\boldsymbol{E}\{\text{PE}(\hat{k}) - m\sigma^2)1_{\{\hat{k}<k^*\}}\} \to 0$ as $n, m \to \infty$.

Turning now to the overfitting contribution, we begin by proving that in the chain of inequalities used to prove the lower bound in cases (B) and (C), the terms dropped—the second and third terms of the right-hand side of Corollary 5.2—all have absolute expectations which are $O(mn^{-1})$. The argument at the beginning of the proof of case (B) of Theorem 2.2 shows this for the second term, for even without the hypothesis (B) we get a constant at that stage by Lemma 5.7(v). Similarly for the third terms,

$$\lim_{m,n} nm^{-1}\boldsymbol{E}\{|\rho_k|^{-4}|\tilde{\rho}_k|^2\langle\rho_k, \epsilon\rangle^2\} = O(1)$$

by Lemma 5.7. Thus we may use the consistency hypothesis and get

$$\begin{aligned}
&\lim_{m,n} nm^{-1}\boldsymbol{E}\{(\text{PE}(\hat{k}) - m\sigma^2)1_{\{\hat{k}>k^*\}}\} \\
&= \sum_{k=k^*+1}^{K} \lim_{m,n} nm^{-1}\boldsymbol{E}\{(\text{PE}(\hat{k}) - m\sigma^2)1_{\{\hat{k}=k\}}\} \\
&= \sum_{k=k^*+1}^{K} \lim nm^{-1}\boldsymbol{E}\{(\text{PE}(k^*) - m\sigma^2)1_{\{\hat{k}=k\}}\} \\
&= \lim_{m,n} nm^{-1}\boldsymbol{E}(\text{PE}(k^*) - m\sigma^2) = \text{tr}\{\tilde{C}_{k^*}C_{k^*}^{-1}\},
\end{aligned}$$

the second last step following from our assumption that $\text{pr}(\hat{k}_n = k) \to 0$ as $n \to \infty$ for all $k > k^*$. This completes the proof of (i).

(ii) Now we suppose that $\hat{k}$ is obtained by minimizing FPE_{α_n} for a sequence $\alpha_n < 2\log\log n$. We know from Theorem 3.2 that $\text{pr}(\hat{k} < k^*) = o(n^{-1})$ and so need only consider overfitting. By Shibata (1984), $\liminf \mathrm{pr}(\hat{k}_n = k^* + 1) > 0$. We next simplify $\lim_{m,n} nm^{-1}\boldsymbol{E}\{(\mathrm{PE}(\hat{k}) - m\sigma^2)\}$ in the now familiar way, noting that (as in the proof of Theorem 2.2) it coincides with

$$\begin{aligned}&\lim_{m,n} nm^{-1}\boldsymbol{E}\{(\mathrm{PE}(\hat{k}) - m\sigma^2)1_{\{\hat{k}\geq k^*\}}\}\\ &\quad\geq \mathrm{tr}\{\tilde{C}_{k^*}C_{k^*}^{-1}\} + \lim_{m,n} nm^{-1}\boldsymbol{E}\{|\rho_{k^*}|^{-4}|\tilde{\rho}_{k^*}|^2\langle\rho_{k^*},\epsilon\rangle^2 1_{\{\hat{k}=k^*+1\}}\}.\end{aligned}$$

Now the second term above is zero only if $\rho_{k^*} = 0$, which implies $k^* = K$, since we have assumed all design matrices to be of full rank. Thus the inequality (2.6) is strict for selection rules based on FPE_{α_n} with $\liminf(2\log\log n)^{-1}\alpha_n < 1$. □

PROOF OF THEOREM 2.1. Since ϵ_t is independent of $\hat{k}_{t-1}$ and $\hat{\beta}_{t-1}$ for all $t > 1$,

$$\boldsymbol{E}\left\{\sum_1^n (y_t - x_t'\hat{\beta}_{t-1}(\hat{k}_{t-1}))^2\right\} = n\sigma^2 + \sum_1^n \boldsymbol{E}(x_t'\beta^* - x_t'\hat{\beta}_{t-1}(\hat{k}_{t-1}))^2.$$

Write

$$\begin{aligned}U_n &= \sum_1^n \boldsymbol{E}\{(x_t'\beta^* - x_t'\hat{\beta}_{t-1}(\hat{k}_{t-1}))^2 1_{\{\hat{k}_{t-1}<k^*\}}\},\\ V_n &= \sum_1^n \boldsymbol{E}\{(x_t'\beta^* - x_t'\hat{\beta}_{t-1}(\hat{k}_{t-1}))^2 1_{\{\hat{k}_{t-1}=k^*\}}\},\\ W_n &= \sum_1^n \boldsymbol{E}\{(x_t'\beta^* - x_t'\hat{\beta}_{t-1}(\hat{k}_{t-1}))^2 1_{\{\hat{k}_{t-1}>k^*\}}\}.\end{aligned}$$

We deal with each of these three components in turn. Let us temporarily denote $x_t'(X_k(t-1)'X_k(t-1))^{-1}X_k(t-1)'\epsilon(t-1)$ by $d'\epsilon$. Then

$$\begin{aligned}U_n &= \sum_{k=1}^{k^*-1}\sum_{t=1}^n \boldsymbol{E}\{(x_t'\beta^* - x_t'\hat{\beta}_{t-1}(\hat{k}_{t-1}))^2 1_{\{\hat{k}_{t-1}=k\}}\}\\ &= \sum_{k=1}^{k^*-1}\sum_{t=1}^n \boldsymbol{E}\{(\lambda_t(k) - d'\epsilon)^2 1_{\{\hat{k}_{t-1}=k\}}\}\\ &\leq 2\sum_{k=1}^{k^*-1}\sum_{t=1}^n [\lambda_t(k)^2\mathrm{pr}(\hat{k}_{t-1} = k) + 2\boldsymbol{E}\{(d'\epsilon)^2 1_{\{\hat{k}_{t-1}=k\}}\}].\end{aligned}$$

Now for $k < k^*$, $\sum_1^n \lambda_t(k)^2 = b_k n + o(1)$ as $n \to \infty$, whilst $\mathrm{pr}(\hat{k}_{t-1} = k) \leq O(t^{-2}(\log t)^{-c})$ as $n \to \infty$, $c > 1$. Summing by parts we thus conclude that

$$\sum_{k=1}^{k^*-1}\sum_{t=1}^n \lambda_t(k)^2\mathrm{pr}(\hat{k}_{t-1} = k) = O(1) \quad \text{as} \quad n \to \infty.$$

Furthermore, $\boldsymbol{E}\{(d'\epsilon)^4\} = 3\boldsymbol{E}\{(d'\epsilon)^2\}$, and since $\boldsymbol{E}(d'\epsilon)^2 = |d|^2\sigma^2 = \mu_t(k)\sigma^2$,

$$\begin{aligned}\sum_{k=1}^{k^*-1}\sum_{t=1}^{n} \boldsymbol{E}\{(d'\epsilon)^2 1_{\{\hat{k}_{t-1}=k\}}\} &\le \sum_{k=1}^{k^*-1}\sum_{t=1}^{n} \sqrt{3}\sigma^2\mu_t(k)\{\mathrm{pr}(\hat{k}_{t-1}=k)\}^{1/2}\\ &= O(1) \quad \text{as} \quad n\to\infty,\end{aligned}$$

as argued above, but this time using $\sum_1^n \mu_t(k) = k\log n(1+o(1))$ as $n\to\infty$. Thus $U_n = O(1)$ as $n\to\infty$.

Turning now to the overfitting term V_n, we find only the quadratic form $(d'\epsilon)^2$, as the bias term vanishes. Thus we can argue as above, giving

$$\begin{aligned}W_n &= \sum_{k=k^*+1}^{K}\sum_{t=1}^{n} \boldsymbol{E}\{(d'\epsilon)^2 1_{\{\hat{k}_{t-1}=k\}}\}\\ &\le \sqrt{3}\sigma^2 \sum_{k=k^*+1}^{K}\sum_{t=1}^{n} \mu_t(k)\{\mathrm{pr}(\hat{k}_{t-1}=k)\}^{1/2} = O(1),\end{aligned}$$

since $\mathrm{pr}(\hat{k}_{t-1}=k) \le O(\log t^{-\alpha})$ as $t\to\infty$, where $\alpha > 2$.

Finally, we examine the term corresponding to getting the model correct. Since $\mathrm{pr}(\hat{k}_{t-1}\ne k^*) \le A(t^{-2}(\log t)^{-c}) + B(\log t)^{-\alpha}$ for large t,

$$\begin{aligned}V_n &= \sum_{t=1}^{n} \boldsymbol{E}\{(x_t'\beta^* - x_t'\hat{\beta}_{t-1}(k^*))^2 1_{\{\hat{k}_{t-1}=k^*\}}\}\\ &= \sum_{t=1}^{n} \boldsymbol{E}\{(d'\epsilon)^2\} - \sum_{t=1}^{n} \boldsymbol{E}\{(d'\epsilon)^2 1_{\{\hat{k}_{t-1}\ne k^*\}}\}\\ &= k^*\log n(1+o(1)) + O(1) \quad \text{as} \quad n\to\infty.\end{aligned}$$

□

Acknowledgements

We would like to thank Jorma Rissanen for his inspiration and for many useful discussions. Special thanks are due to David Freedman for his criticisms of the manuscript.

References

Akaike, H. (1970). Statistical predictor identification, *Ann. Inst. Statist. Math.*, **22**, 202–217.

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Trans. Automat. Control*, **19**, 716–723.

Atkinson, A. C. (1980). A note on the generalized information criterion for choice of a model, *Biometrika*, **67**, 413–418.

Bhansali, R. H. and Downham, D. Y. (1977). Some properties of the order of an autoregressive model selected by a generalization of Akaike's FPE criterion, *Biometrika*, **64**, 547–551.

Box, G. E. P. and Draper, N. R. (1959). A basis for the selection of a regression surface design, *J. Amer. Statist. Assoc.*, **54**, 622–654.

MODEL SELECTION AND PREDICTION: NORMAL REGRESSION

Box, G. E. P. and Draper, N. R. (1963). The choices of a second order rotatable design, *Biometrika*, **50**, 335–352.

Breiman, L. A. and Freedman, D. F. (1983). How many variables should be entered in a regression equation?, *J. Amer. Statist. Assoc.*, **78**, 131–136.

Clayton, M. K., Geisser, S. and Jennings, D. (1986). A comparison of several model selection procedures, *Bayesian Inference and Decision* (eds. P. Goel and A. Zellner), 425–439, Elsevier, New York.

Dawid, A. P. (1984). Present position and potential developments: some personal views, Statistical theory—The prequential approach (with discussion), *J. Roy. Statist. Soc. Ser. A*, **147**, 278–292.

Dawid, A. P. (1992). Prequential data analysis, *Current Issues in Statistical Inference: Essays in Honor of D. Basu, Institute of Mathematical Statistics, Monograph*, **17** (eds. M. Ghosh and P. K. Pathak).

Geweke, J. and Meese, R. (1981). Estimating regression models of finite but unknown order, *Internat. Econom. Rev.*, **22**, 55–70.

Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression, *J. Roy. Statist. Soc. Ser. B*, **41**, 190–195.

Hannan, E. J., McDougall, A. J. and Poskitt, D. S. (1989). Recursive estimation of autoregressions, *J. Roy. Statist. Soc. Ser. B*, **51**, 217–233.

Haughton, D. (1989). Size of the error in the choice of a model to fit data from an exponential family, *Sankhyā Ser. A*, **51**, 45–58.

Hemerly, E. M. and Davis, M. H. A. (1989). Strong consistency of the predictive least squares criterion for order determination of autoregressive processes, *Ann. Statist.*, **17**, 941–946.

Hjorth, U. (1982). Model selection and forward validation, *Scand. J. Statist.*, **9**, 95–105.

Kohn, R. (1983). Consistent estimation of minimal dimension, *Econometrica*, **51**, 367–376.

Lai, T., Robbins, H. and Wei, C. Z. (1979). Strong consistency of least squares estimates in multiple regression II, *J. Multivariate Anal.*, **9**, 343–361.

Merhav, N., Gutman, M. and Ziv, J. (1989). On the estimation of the order of a Markov chain and universal data compression, *IEEE Trans. Inform. Theory*, **39**, 1014–1019.

Nishi, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression, *Ann. Statist.*, **12**, 758–765.

Rao, C. R. (1973). *Linear Statistical Inference*, 2nd ed., Wiley, New York.

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

Rissanen, J. (1986*a*). Stochastic complexity and modeling, *Ann. Statist.*, **14**, 1080–1100.

Rissanen, J. (1986*b*). A predictive least squares principle, *IMA J. Math. Control Inform.*, **3**, 211–222.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry*, World Books, Singapore.

Sawa, T. (1978). Information criteria for discriminating among alternative regression models, *Econometrica*, **46**, 1273–1291.

Schwartz, G. (1978). Estimating the dimension of a model, *Ann. Statist.*, **6**, 461–464.

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

Shibata, R. (1981). An optimal selection of regression variables, *Biometrika*, **68**, 45–54.

Shibata, R. (1983*a*). Asymptotic mean efficiency of a selection of regression variables, *Ann. Inst. Statist. Math.*, **35**, 415–423.

Shibata, R. (1983*b*). A theoretical view of the use of AIC, *Times Series Analysis: Theory and Practice 4* (ed. O. D. Anderson), 237–244, Elsevier, Amsterdam.

Shibata, R. (1984). Approximate efficiency of a selection procedure for the number of regression variables, *Biometrika*, **71**, 43–49.

Shibata, R. (1986*a*). Selection of the number of regression variables; a minimax choice of generalized FPE, *Ann. Inst. Statist. Math.*, **38**, 459–474.

Shibata, R. (1986*b*). Consistency of model selection and parameter estimation, *Essays in Time Series and Allied Processes: Papers in Honour of E. J. Hannan, J. Appl. Probab.*, **23A**, 127–141.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion, *J. Roy. Statist. Soc. Ser. B*, **39**, 44–47.

Wax, M. (1988). Order selection for AR models by predictive least squares, *IEEE Trans. Acoust. Speech Signal Process.*, **36**, 581–588.

Wei, C. Z. (1992). On the predictive least squares principle, *Ann. Statist.*, **20**, 1–42.

Woodroofe, M. (1982). On model selection and the arc sine laws, *Ann. Statist.*, **10**, 1182–1194.

IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 39, NO. 3, MAY 1993

A Rate of Convergence Result for a Universal D-Semifaithful Code

Bin Yu and T. P. Speed

***Abstract*—The problem of optimal rate universal coding in the context of rate-distortion theory is considered. A D-semifaithful universal coding scheme for discrete memoryless sources is given. The main result is a refined covering lemma based on the random coding argument and the method of types. The average codelength of the code is shown to appraoch its lower bound, the rate-distortion function, at a rate $O(n^{-1} \log n)$, and this is conjectured to be optimal based on a result of Pilc. Issues of constructiveness and universality are also addressed.**

***Index Terms*— Discrete memoryless source, rate-distortion, D-semifaithful, universal coding, optimal rate, random coding, method of types.**

I. Introduction

ENTROPY has a central position in information theory, in part because in the limit it gives the shortest possible per-symbol average length of a noiseless code. If we consider a discrete memoryless source with distribution P_0, the entropy $H(P_0)$ serves as a nonasymptotic lower bound to the average expected codelength for data strings from this source. Moreover, the entropy lower bound can be achieved asymptotically at the rate $O(n^{-1})$ when the source distribution P_0 is known, and at the rate $O(n^{-1} \log n)$ when P_0 is not known.

Rissanen [19] improved the entropy lower bound by showing that entropy $\frac{1}{2} kn^{-1} \log n$ is an asymptotic lower bound to the average expected codelength. His bound holds for data strings from parametric statistical models satisfying mild regularity conditions, and the k in the lower bound is the dimension of the model. Discrete memoryless sources are covered by his result, with k there being the cardinality of the source alphabet minus one, and the rate $O(n^{-1}\log n)$ is optimal in this case, when P_0 is not known. The rate $O(n^{-1}\log n)$ has been shown to be achievable for various other statistical models, see for example Davisson [8], Rissanen [19], [20], Hannan and Kavalieris [10], Hemerly and Davis [11], Gerenscer and Rissanen [9], Clarke and Barron [5], and Weinberger, Lempel, and Ziv [27]. Extensions to nonparametric models can be found in Barron and Cover [1], Rissanen, Speed, and Yu [21], and Yu and Speed [28].

Rate-distortion theory was started by Shannon [23], and in that context we consider block-codes with a fidelity criterion, or semifaithful codes to use the term from a recent paper of Ornstein and Shields [15]. Instead of the expected codelength used in noiseless coding it is natural in rate-distortion theory to consider the log of the number of D-balls required to cover the n-tuple space of the source alphabet under some single-letter distance measure. The role of entropy in noiseless coding is then taken by the rate-distortion function, in the following sense: the rate-distortion function gives a lower bound to the log of the covering number, which we may also refer to as the expected code length of a D-semifaithful code, and this lower bound can be achieved in the limit by certain D-semifaithful codes. In particular, Ornstein and Shields [15] obtain D-semifaithful codes which achieve the rate-distortion function lower bound almost surely, for ergodic sequences, and Shields [22] uses Markov types for similar results. Earlier work for other classes of sources include Neuhoff, Gray, and Davisson [14], Mackenthun and Pursley [13] and Kieffer [12]. In the case of memoryless sources, the achievability proof can be found in standard texts, see for example, Cover and Thomas [7] for a recent exposition using the random coding argument. However, no results have yet been provided on the rate at which this lower bound is approached.

In this paper, we describe a D-semifaithful universal coding scheme of memoryless sources and obtain an associated rate result. We show, for a discrete memoryless source with a source alphabet of J elements and an unknown distribution P_0, that under some mild smoothness conditions on the rate-distortion function, a universal D-semifaithful code can be constructed such that the average expected length of this code tends to the rate-distortion function at the rate $n^{-1}\log n$. The techniques used are the method of types and random coding. The main result will be based on a refined coding lemma (Theorem 1) for type classes. It is "refined" because it improves the $o(1)$ term in the covering lemma in Csiszár and Körner [6] to an $O(n^{-1}\log n)$ term. In other words, we are able to give a better upper bound on (the log of) the number of D-balls needed to cover a type class, equivalently, on the number of D-semifaithful code words required to encode a type class. Then a two-stage code is constructed as the D-semifaithful code for all strings: first we encode the type class, and next we encode the elements of each type class using the refined covering lemma. The above results are contained in Section II.

In Section III, we conjecture that the rate $n^{-1} \log n$ is asymptotically optimal. Our conjecture is based on a result of Pilc [16], [17], which is expressed in terms of the inverse of the rate-distortion function: the distortion-rate function. Pilc has upper bounds and lower bounds for noiseless channels and

Manuscript received June 10, 1991; revised August 17, 1992. This work was supported in part by the Wisconsin Alumni Research Foundation and the Army Research Office Grant DAAL03-91-G-007, and the National Science Foundation Grant DMS 8802378.

B. Yu is with the Department of Statistics, University of Wisconsin, Madison, WI 53706.

T. P. Speed is with the Department of Statistics, University of California, Berkeley, CA 94720.

IEEE Log Number 9206227.

for noisy channels, but we use his results only for noiseless channels. Unfortunately, although the rate $n^{-1}\log n$ in the upper bound of our two-stage code matches that in Pilc's lower bound, his lower bound is on the log cardinality of an *expected* D-semifaithful code (cf. the forthcoming definition) while our code is *pointwise* D-semifaithful with an upper bound on the expected codelength. Hence, we do not know at this stage if the rate $n^{-1}\log n$ is indeed optimal in terms of expected codelength. Moreover, his bound does not include Rissanen's since it holds only for nonzero distortion levels.

In Section IV, we compare our code with the code corresponding to Pilc's upper bound. The main point made there is that our code is universal, while the other one is not. In addition, the issue of construction versus pure existence is addressed in relation to our code and the one corresponding to Pilc's upper bound.

We start with some preliminaries on rate-distortion theory and the method of types. Our main reference on rate-distortion theory is Berger [2], and that on the method of types is Csiszár and Körner [6].

II. Preliminaries

Let $\mathcal{A}_0 = \{1, 2, \cdots, J-1, J\}$ be the source alphabet, and let $\mathcal{B}_0 = \{1, 2, \cdots, K\}$ be the reproducing alphabet. $\mathcal{B}_0$ could be the same as or a subset of $\mathcal{A}_0$. We assume our source is memoryless, i.e., that the letters $x_1, \cdots, x_n$, which make up our strings are mutually independent and identically distributed (i.i.d.) with distribution P_0 on $\mathcal{A}_0$. Without loss of generality we assume $P_0(j) > 0$ for all $j \in \mathcal{A}_0$. We use a single-letter fidelity criterion to measure the distortion between any nth order source string $x^n = (x_1, \cdots, x_n) \in \mathcal{A}_0^n$, and its code word $y^n \in \mathcal{B}_0^n$. More precisely, let

$$d_n(x^n, y^n) = n^{-1}\sum_{t=1}^{n} d(x_t, y_t),$$

where d is a bounded real nonnegative function on $\mathcal{A}_0 \times \mathcal{B}_0$, with maximum d_M and minimum d_m. Then the rate distortion function $R_n(P_0, D)$ for the distribution of $x_1, \cdots, x_n$ equals $nR(P_0, D)$ where the rate-distortion function $R(P_0, D)$ of P_0 can be formally defined as follows:

$$\begin{aligned} R(P_0, D) &= \min_W I(W, P_0) \\ &= \min_W \sum_{j=1}^{J}\sum_{k=1}^{K} P_0(j)W(k|j)\log\frac{W(k|j)}{Q(k)}, \end{aligned}$$

where the minimum is taken over the set of matrices W from $\mathcal{A}_0$ to $\mathcal{B}_0$ such that for any $j, k, W(k|j) \geq 0$, for all j, $\sum_{k=1}^{K} W(k|j) = 1$,

$$\sum_{j=1}^{J}\sum_{k=1}^{K} P_0(j)W(k|j)\,d(j,k) \leq D,$$

and Q is the marginal distribution on $\mathcal{B}_0$ induced by P_0 and W, i.e., for $k \in \mathcal{B}_0$,

$$Q(k) = \sum_{j=1}^{J} P_0(j)W(k|j).$$

The following properties of $R(P, D)$ can be found in Berger [2].

1) $R(P, \cdot)$ is convex, monotonically decreasing on $[0, D_{\max}]$ where $D_{\max} = \min_k \sum_{j=1}^{J} P(j)\,d(j,k)$. Moreover, for $D \geq D_{\max}$, $R(P, D) = 0$, and $R(P, 0) = H(P) = -\sum_{j=1}^{J} P(j)\log P(j)$. Hence $R'_D(P, D) \leq 0$ for any D, where $'$ denotes differentiation with respect to D.

2) If $I(W, P) = R(P, D)$, then for any j:

$$\frac{\partial I(W,P)}{\partial W(k|j)}|W = P(j)\log\frac{W(k|j)}{Q(k)},$$

where for any j, k:

$$W(k|j) = \frac{Q(k)e^{sd(j,k)}}{\sum_{\ell} Q(\ell)e^{sd(j,\ell)}},$$

with $s = R'_D(P, D) \leq 0$.

Definition (D-semifaithful code): A map $M_n : \mathcal{A}_0^n \to \mathcal{B}_0^n$ is called a *pointwise* D-semifaithful code if for any $x^n \in \mathcal{A}_0^n$

$$d_n(x^n, M_n(x^n)) \leq D.$$

Similarly a map M_n is called *expected* D-semifaithful with respect to a source distribution P_0 if whenever $(x_1, \cdots, x_n)$ are i.i.d. with common distribution P_0,

$$E_{P_0} d_n(x^n, M_n(x^n)) \leq D.$$

Since our main argument will be based on the method of types, we next introduce the definition of type and some of its properties. We will follow the notation of Csiszár and Körner [6], with $1\{A\}$ denoting the indicator of the event A and $\equiv$ meaning equal by definition.

Definition (Type): The type of a sequence $x^n \in \mathcal{A}_0^n$ is the distribution P_{x^n} on $\mathcal{A}_0$ defined for $j \in \mathcal{A}_0$ by

$$P_{x^n}(j) \equiv \frac{1}{n}N(j|x^n) \equiv \frac{1}{n}\sum_{t=1}^{n} 1\{x_t = j\}$$

that is, the empirical distribution of x^n on $\mathcal{A}_0$. We write $T_P^n = \{x^n : x^n \text{ has type } P\}$ for any P on $\mathcal{A}_0$ such that $\{nP(j)\}$ are integers.

For any given $x^n \in \mathcal{A}_0^n$, and a stochastic matrix $W : \mathcal{A}_0 \to \mathcal{B}_0$, we next define conditional types.

Definition (Conditional type): The conditional type W of a sequence $y^n \in \mathcal{B}_0^n$ given $x^n \in \mathcal{A}_0^n$ is defined for $j \in \mathcal{A}_0$, $k \in \mathcal{B}_0$ by

$$\begin{aligned} N(j|x^n)W(k|j) &\equiv N(j,k|x^n, y^n) \\ &\equiv \sum_{t=1}^{n} 1\{x_t = j \text{ and } y_t = k\}. \end{aligned}$$

We denote the set of sequences $y^n \in \mathcal{B}_0^n$ having the conditional type W given x^n by $T_W^n(x^n)$.

The cardinality of a type class, or a conditional type class can be bounded above and below as in the following results from Csiszár and Körner [6].

Lemma 1: For any type P of sequences in $\mathcal{A}_0^n$,

$$(n+1)^{-J}\exp(nH(P)) \leq |T_P^n| \leq \exp(nH(P)). \quad (1.1)$$

Lemma 2: For every $x^n \in \mathcal{A}_0^n$, and stochastic matrix $V : \mathcal{A}_0 \to \mathcal{B}_0$ such that $T_V^n(x^n)$ is nonempty,

$$(n+1)^{-JK} \exp\{nH(V|P_{x^n})\} \le |T_V^n(x^n)| \le \exp\{nH(V|P_{x^n})\}, \quad (1.2)$$

where $H(V|P) = \sum_{j=1}^{J} P(j)H(V(\cdot|j)) = -\sum_{j=1}^{J} P(j) \sum_{k=1}^{K} V(k|j) \log V(k|j)$.

Lemma 3: The total number of type classes is at most $(n+1)^J$.

III. A Universal Pointwise D-Semifaithful Code

In this section, we first use the random coding argument in Csiszár and Körner [6] to prove a refined covering lemma (Theorem 1). Then, we go on to give a two-stage universal D-semifaithful coding scheme (Theorem 2) with the rate $n^{-1} \log n$. We begin with a proposition extracted from Csiszár and Körner [6].

For a given type P on $\mathcal{A}_0^n$, positive constant D in $(0, D_{\max})$, and a subset B of $\mathcal{B}_0^n$, write

$$U_D(B) = \{x^n \in T_P^n : d_n(x^n, B) > D\},$$

where $d_n(x^n, B) := \min_{y^n \in B} d_n(x^n, y^n)$.

Proposition 1: Suppose $Z^{(m)} = \{Z_1, \cdots, Z_m\}$ are i.i.d. and uniform over a subset $G \subset \mathcal{B}_0^n$. If for some m_n we have $E|U_D(Z^{(m_n)})| < 1$, then there is a set $B_{P,D}$ such that $|B_{P,D}| \le m_n$ and $|U_D(B_{P,D})| < 1$. This implies that $U_D(B_{P,D}) = \phi$. In other words, that $B_{P,D}$ "covers" T_P^n within distance D.

See Csiszár and Körner [6] for the proof.

Moreover, note that,

$$\left|U_D\left(Z^{(m)}\right)\right| = \sum_{x^n \in T_P^n} 1\left\{U_D\left(Z^{(m)}\right)\right\}(x^n)$$

implying

$$E\big|\big(U_D\big(Z^{(m)}\big)\big| = \sum_{x^n \in T_P^n} \mathbb{P}_0\left(x^n \in U_D\left(Z^{(m)}\right)\right). \quad (2.1)$$

For any fixed $x^n \in T_P^n$, because the Z are i.i.d.,

$$\mathbb{P}_0\left(x^n \in U_D\left(Z^{(m)}\right)\right) = [\mathbb{P}_0(d_n(x^n, Z) > D)]^m, \quad (2.2)$$

where Z is uniformly distributed over G.

Furthermore, if we can find a subset $G_1(x^n) \subset G$ such that, for any $y^n \in G_1(x^n)$ we have $d_n(x^n, y^n) \le D$, then

$$\begin{aligned} \mathbb{P}_0(d_n(x^n, Z) > D) &= 1 - \mathbb{P}_0(d_n(x^n, Z) \le D) \\ &\le 1 - \mathbb{P}_0(Z \in G_1(x^n)) \\ &= 1 - |G_1(x^n)|/|G|. \end{aligned} \quad (2.3)$$

Combining (2.1), (2.2), and (2.3), we get

$$\begin{aligned} E\left|U_D\left(Z^{(m)}\right)\right| &\le \sum_{x^n \in T_P^n} \left(1 - \frac{|G_1(x^n)|}{|G|}\right)^m \\ &\le \sum_{x^n \in T_P^n} \exp\left(-\frac{|G_1(x^n)|}{|G|} m\right). \end{aligned} \quad (2.4)$$

The last inequality holds because $(1-t)^m \le \exp(-tm)$ for any $t > 0$. Next we choose a conditional type class as $G_1(x^n)$, and a type class as G. For the chosen $G_1(x^n)$ and G, we select m_n using (2.4) such that $E|U_D(Z^{(m)})| < 1$. For any type P and constant D in $(0, D_{\max})$, take W such that

$$\sum_{j,k} W(k|j)P(j)d(j,k) \le D^*,$$

and $I(W, P) = R(P, D^*)$, where $D^* = D - n^{-1}JKd_M$.

Note that this W depends on both P and D^*, but for simplicity, we do not indicate this dependence in any way. Because the $nP(j)$ are integers, we can find a stochastic matrix $[W]$, a truncation of W, such that for all j and k, $n[W](k|j)P(j)$ are integers, and

$$|W(k|j) - [W](k|j)| \le \frac{1}{nP(j)}, \qquad \text{for } j = 1, \cdots, J.$$

Let $[Q] = [W] \cdot P$, i.e., $[Q](k) = \sum_{j-1}^{J} [W](k|j)P(j)$. Then, the $n[Q](k)$ are also integers. Therefore, the type class $T_{[Q]}^n$ and the conditional type class $T_{[W]}^n(x^n)$ are well defined for all $x^n \in T_P^n$.

Let us take $G = T_{[Q]}^n$ and $G_1(x^n) = T_{[W]}^n(x^n)$. Then, for any $y^n \in G_1(x^n)$,

$$N(j, k|x^n, y^n) = [W](k|j)P(j)n,$$

and

$$\begin{aligned} N(k|y^n) &= \sum_j N(j, k|x^n, y^n) \\ &= \sum_j [W](k|j)P(j)n = [Q](k)n, \end{aligned}$$

that is, $y^n \in G = T_{[Q]}^n$. Hence, $G_1(x^n) \subseteq G$.

In addition, for any $y^n \in G_1(x^n)$, since $D^* = D - n^{-1}JKd_M$,

$$\begin{aligned} d_n(x^n, y^n) &= \sum_{j,k} [W](k|j)P(j)d(j,k) \\ &\le \sum_{j,k} (W(k|j)P(j) + n^{-1})d(j,k) \\ &\le \sum_{j,k} W(k|j)P(j)d(j,k) + n^{-1}JKd_M \\ &\le D^* + n^{-1}JKd_M = D. \end{aligned}$$

Recalling the bounds in (1.1) and (1.2), we have

$$\begin{aligned} |G| &\le \exp(nH(Q)), \\ |G_1(x^n)| &\ge (n+1)^{-JK} \exp\{nH([W]|P)\}. \end{aligned}$$

Thus, since $I([W], P) = H([Q]) - H([W]|P)$,

$$-\frac{|G_1(x^n)|}{|G|} \le -(n+1)^{-JK} \exp(-nI([W], P)). \quad (2.5)$$

Putting (2.5) into (2.4), we find

$$\begin{aligned} E\left| U_D\left(Z^{(m)}\right) \right| &\le \sum_{x^n \in T_P^n} \\ &\quad \cdot \exp\{-(n+1)^{-JK} \exp(-nI([W],P))m\} \\ &= |T_P^n| \exp\{-(n+1)^{-JK} \\ &\quad \cdot \exp(-nI([W],P))m\} \end{aligned}$$

Finally, we get by Lemma 1:

$$E\left| U_D\left(Z^{(m)}\right) \right| \le \exp(nH(P)) \exp\{-(n+1)^{-JK} \cdot \exp(-nI([W],P))m\}. \quad (2.6)$$

Now we can choose $m = m_n$ as an integer such that

$$\exp\{nI([W],P) + (JK+2)\log(n+1)\} \le m_n \le \exp\{nI([W],P) + (JK+4)\log(n+1)\}.$$

Then for such an m_n, (2.6) gives

$$E|U_D(Z^{(m)})| \le \exp(n \log J) \exp\left(-(n+1)^2\right) < 1, \qquad \text{for } n \text{ large.}$$

Applying Proposition 1, we obtain the following theorem.

Theorem 1 (Refined Covering Lemma for Type Classes): Given a type P on $\mathcal{A}_0$ and D in $(0, D_{\max})$, there is a subset $B_{P,D} \subset \mathcal{B}_0^n$ such that for any $x^n \in T_P^n$, $d_n(x^n, B_{P,D}) \le D$ and

$$|B_{P,D}| \le \exp\{nI([W],P) + (JK+4)\log(n+1)\},$$

where for any j, k

$$|[W](k|j) - W(k|j)| \le \frac{1}{nP(j)},$$

and $I(W,P) = R(P, D^*)$ for $D^* = D - n^{-1}JKd_M$.

Next, we show that we can replace $[W]$ by W in Theorem 1. Since $[W]$ is close to W, and D^* is close to D, we expect $I([W],P)$ to be close to $I(W,P) = R(P,D^*)$, hence close to $R(P,D)$. Formally, we expand $I([W],P)$ around $I(W,P)$ as follows:

$$I([W],P) = I(W,P) + \sum_{j=1}^{J} \sum_{k=1}^{K-1} \frac{\partial I}{\partial W(k|j)} \cdot (\cdot, P)|W_{(k|j)}([W](k|j) - W(k|j)) + \cdots, \quad (2.7)$$

where $(\cdots)$ denotes smaller order terms. Since $I(W,P) = R(P,D^*)$, by property 2) in Section I, for any k, j

$$\frac{\partial I(\cdot, P)}{\partial W(k|j)} |W = P(j) \log \frac{W(k|j)}{Q(k)}.$$

Note that for any j, k, $|[W](k|j) - W(k|j) \le (nP(j))^{-1}$, so we have from (2.7):

$$\begin{aligned} I([W],P) &= I(W,P) \\ &\quad + \left\{ \sum_{j=1}^{J} \sum_{k=1}^{K-1} \log \frac{W(k|j)}{Q(k)} \right\} n^{-1} \\ &\quad + o\left(n^{-1}\right). \end{aligned}$$

However, again by property 2) in Section I, for all j, k we have

$$W(k|j) = \frac{Q(k)e^{sd(j,k)}}{\sum_{\ell} Q(\ell) e^{sd(j,\ell)}},$$

where $s = R'_D(P, D^*) < 0$. Hence,

$$\frac{W(k|j)}{Q(k)} = \frac{e^{sd(j,k)}}{\sum_{\ell} Q(\ell)e^{sd(j,\ell)}} \le \frac{e^{-|s|d_m}}{\sum_{\ell} Q(\ell)e^{-|s|d_M}} = e^{|s|(d_M - d_m)}.$$

Similarly $W(k|j)/Q(k) \ge e^{-|s|(d_M - d_m)}$, and hence,

$$|\log \frac{W(k|j)}{Q(k)}| \le |s|(d_M - d_m).$$

Without loss of generality, assume $d_m = 0$. We then get

$$\begin{aligned} I([W],P) &\le I(W,P) + n^{-1}JK|s|d_M + 0\left(n^{-1}\right) \\ &= R(P,D^*) + n^{-1}JK|s|d_M + O\left(n^{-1}\right). \end{aligned}$$

On the other hand,

$$\begin{aligned} I(W,P) &= R(P,D^*) \\ &= R(P,D) + R'_D(P,D)(D^* - D) + \cdots \\ &= R(P,D) + |s|JKd_M n^{-1} + \cdots. \end{aligned}$$

Thus,

$$I([W],P) \le R(P,D) + 2JK|s|d_M n^{-1} + o\left(n^{-1}\right), \quad (2.8)$$

where $s = R'_D(P,D)$. We have proved the following

Corollary 1: Under the assumptions of Theorem 1

$$\log|B_{P,D}| \le nR(P,D) + 2JK|s|d_M + o(1) + (KJ+4)\log(n+1).$$

Theorem 2 (Universal Pointwise D-Semifaithful Coding): Let $\mathcal{F}$ be a class of distributions on $\mathcal{A}_0$ such that for some $D \in (0, D_{\max})$, the derivatives $\{\partial^2 R(P,D)/\partial P_j \partial P_{j'} : j, j' = 1, \cdots, J\}$ are uniformly bounded over $\mathcal{F}$ by a constant C, and $|E_{P_0} R'_D(P_{x^n}, D)| < \infty$ for all $P_0 \in \mathcal{F}$. Then there exists a two-stage code $M_n : \mathcal{A}_0^n \to \mathcal{B}_0^n$ such that

$$d_n(x^n, M_n(x^n)) \le D,$$

and for all $P_0 \in \mathcal{F}$, as $n \to \infty$,

$$n^{-1} E_{P_0} L(M_n(x^n)) \le R(P_0, D) + (KJ + J + 4)n^{-1} \cdot \log(n+1) + O\left(n^{-1}\right).$$

Proof of Theorem 2: For any $x^n \in \mathcal{A}_0^n$, our coding scheme has two stages: First, we encode P_{x^n}, which by Lemma 3 takes at most $J \log(n+1)$ bits. Next, we use Corollary 1, which asserts that for the type class $T_{P_{x^n}}^n$, there is a $B_{P_{x^n},D}$ which covers $T_{P_x^n}^n$ with radius D. We then take $M_n : T_{P_{x^n}}^n \to B_{P_{x^n},D}$ where $M_n(x^n) = y^n \in B_{P_{x^n},D}$ is such that $d_n(x^n, y^n) \le D$. This takes at most $\log|B_{P_{x^n},D}|$ bits, which, since $R'_D < 0$, is bounded by

$$nR(P_{x^n}, D) + (KJ+4)\log(n+1) - 2KJR'_D(P_{x^n}, D) + o(1).$$

For simplicity, denote P_{x^n} by $\overline{P}$. Taking the expectation of this last expression gives a bound on the expected codelength of

$$E_{P_0}R(\overline{P},D) + (KJ+4)n^{-1}\,\log(n+1) \\ - 2KJE_{P_0}R'_D(\overline{P},D)n^{-1} + o(n^{-1}).$$

To prove the theorem, it suffices to show

$$E_{P_0}R(\overline{P},D) = R(P_0,D) + O(n^{-1}), \qquad (2.9)$$

because $|E_{P_0}R'_D(\overline{P},D)| < \infty$ by assumption.

The idea to show (2.9) is simply a Taylor expansion of $R(P,D)$ around $P = P_0$, but because the Taylor expansion holds only in an $o(1)$ neighborhood of P_0, some effort has to be made to give a rigorous proof.

We split the set of types into two disjoint subsets:

$$\Omega_n = \{P : |P(j) - P_0(j)| \\ \le n^{-1/2}\log n, \text{ for all } j \in \mathcal{A}_0\},$$

and

$$\Omega_n^c = \{P : |P(j) - P_0(j)| \\ > n^{-1/2}\log n, \text{ for some } j \in \mathcal{A}_0\},$$

and we break the expectation $E_{P_0}R(\overline{P},D)$ up similarly, defining

$$E_1 = \sum_{P\in\Omega_n} \mathbb{P}_0(x^n \in T_P^n)R(P,D)$$
$$E_2 = \sum_{P\in\Omega_n^c} \mathbb{P}_0(x^n \in T_P^n)R(P,D),$$

where $\mathbb{P}_0$ denotes the probability measure on sequences defined by P_0. Before we go further, we need a good bound on $\sum_{P\in\Omega_n^c} \mathbb{P}_0(x^n \in T_P^n)$. Hoeffding's inequality, cf. Pollard [18, p. 191], implies that for all j,

$$\mathbb{P}_0(x^n : |\overline{P}(j) - P_0(j)| > n^{-1/2}\,\log\, n) \\ \le \exp(-2[\log n]^2 \cdot n/4n) \\ = \exp\Big(-\frac{1}{2}(\log n)^2\Big).$$

As a result,

$$\sum_{P\epsilon\Omega_n^c} \mathbb{P}_0(x^n \epsilon T_P^n) \\ \le \sum_{j=1}^{J} \mathbb{P}_0\Big(x^n : |\overline{P}(j) - P_0(j)| > n^{-1/2}\log n\Big) \\ \le J\,\exp\Big(-\frac{1}{2}(\log n)^2\Big) = Jn^{-\frac{1}{2}\log n}. \qquad (2.10)$$

Then (2.10) and the inequality $R(P,D) \le \log J$ together yield

$$E_2 = \sum_{P\in\Omega_n^c} \mathbb{P}_0(x^n \in T_P^n)R(P,D) \\ \le (\log J)Jn^{-\frac{1}{2}\log n} \\ = O(n^{-1}), \qquad \text{for } n \text{ large.}$$

On Ω_n, we can expand $R(P,D)$ as

$$R(P,D) = R(P_0,D) + \sum_{j=1}^{J-1} \frac{\partial R}{\partial P_j}(P_0,D)\cdot((P(j) - P_0(j)) \\ + \frac{1}{2}\sum_{j,j'}(P(j') - P_0(j'))\frac{\partial^2 R}{\partial P_j \partial P_{j'}}(P_0',D) \\ \cdot((P(j) - P_0(j)),$$

where P_0' is in between P_0 and P. Because the partial derivatives around P_0 are bounded by a constant C, the third term on the right is bounded by

$$2JC\sum_{j=1}^{J}(P(j) - P_0(j))^2,$$

and so its expectation is $O(n^{-1})$ by a known result concerning the multinomial variance.

Moreover, the fact that $E_{P_0}(P(j) - P_0(j)) = 0$ for all j, implies that

$$\Big|\sum_{P\in\Omega_n} \mathbb{P}_0(x^n \in T_P^n)(P(j) - P_0(j))\Big| \\ = \Big|\sum_{P\in\Omega_n^c} \mathbb{P}_0(x^n \in T_P^n)(P(j) - P_0(j))\Big| \\ \le 2\sum_{P\in\Omega_n^c} \mathbb{P}_0(x^n \in T_P^n) = O(n^{-1}). \qquad (2.11)$$

Similarly, we can show

$$\sum_{P\in\Omega_n} P_0(x^n \epsilon T_P^n)R(P_0,D) = R(P_0,D) + O(n^{-1}). \qquad (2.12)$$

Hence,

$$E_{P_0}R(\overline{P},D) \le R(P_0,D) + O(n^{-1}).$$

This completes the proof that

$$n^{-1}E_{P_0}L(M_n(x^n)) \le R(P_0,D) \\ + (KJ+J+4)n^{-1} \\ \cdot\log(n+1) + O(n^{-1}). \qquad \square$$

Remark: Note that it is very easy to check that the boundedness conditions on the derivatives of the rate-distortion funciton are satisfied by a Bernoulli(p) source with distortion measured by Hamming distance. In that case, the rate-distortion function is known, cf. Cover and Thomas [7], to be

$$R(p,D) = \begin{cases} H(p) - H(D), & \text{if } 0 \le D \le \min(p, 1-p), \\ 0, & \text{if } D > \min(p, 1-p). \end{cases}$$

Then,

$$R'_D(p,D) = \log\frac{1-D}{D}$$

and

$$\frac{\partial^2}{\partial p^2}R(p,D) = -\frac{1}{p(1-p)}$$

It is clear that the boundedness conditions are satisfied in this case if we choose $\mathcal{F}$ as the set of binary distributions with uniform bounds on p away from both 0 and 1. In general, if those derivatives exist, they are likely to be bounded. Thus, our conditions do not appear to be too stringent.

IV. Lower Bound

For a string of i.i.d. discrete source letters from $\mathcal{A}_0$, we have shown in the last section that under some smoothness conditions there is a pointwise D-semifaithful code with its average expected code length tending to $R(P_0, D)$ at the rate $n^{-1}\log n$. Recalling that $R(P_0, D)$ is a lower bound on the average codelength of such D-semifaithful codes, we may ask: is the rate $n^{-1} \log n$ the best possible?

Unfortunately, we have not been able to show that $n^{-1} \log n$ is the optimal rate, though we conjecture it is the case. The main reason for our conjecture is a lower bound due to Pilc [16], [17] in terms on the distortion-rate function $D(P_0, R)$, which is the inverse function of $R(P_0, D)$ in the variable D. Note that $D(P_0, \cdot)$ is defined on $[0, H(P_0)]$.

Theorem 3 (Pilc [17]): Assume that $x_1, \cdots, x_n$ are i.i.d. with distribution P_0 on $\mathcal{A}_0$, and that M_n is a map from $\mathcal{A}_0^n \to \mathcal{B}_0^n$. Given $R \in (0, H(P_0))$, if $|M_n(\mathcal{A}_0^n)| \le 2^{nR}$, then

$$\begin{aligned} E_{P_0} d_n(x^n, M_n(x^n)) &\ge D(P_0, R) \\ &\quad + \frac{1}{2}\,\frac{\log n}{|s_0(R)|n}\,(1 + o(1)), \qquad (3.1) \end{aligned}$$

where s_0 satisfies

$$\mu(s_0, P_0) - s_0\mu'(s_0, P_0) = -R$$

with

$$\mu(s, P_0) = \sum_{j=1}^{J} P_0(j) \log\Bigg(\sum_{k=1}^{K} Q(k) \exp\,(sd(j,k))\Bigg),$$
$$Q = WP_0,$$

and

$$I(W, P_0) = R(P_0, D_R) = R.$$

Moreover, for any $\epsilon > 0$, there exists $N(\epsilon) > 0$, such that if $n > N(\epsilon)$,

$$\begin{aligned} &\min E_{P_0} d_n(x^n, M_n(x^n)) \\ &\qquad \le D(P_0, R) + \frac{1}{2}\,(1 + \epsilon)\,\frac{\log n}{|s_0|n}\,(1 + o(1)), \qquad (3.2) \end{aligned}$$

where the minimum is taken over all codes (maps) M_n such that $|M_n(\mathcal{A}_0^n)| \le 2^{nR}$.

It is worth noting that the constant $\frac{1}{2}$ in front of the rate $n^{-1}\log n$ does not depend on the dimension J of source distribution P_0, whereas in the noiseless case the corresponding constant is $\frac{1}{2}\,(J - 1)$.

Applying the function $R(P_0, \cdot)$ to both sides of (3.1), we get the following corollary.

Corollary 2 (Lower Bound): Under the assumptions of Theorem 3, for any *expected* D-semifaithful code M_n, if $|M_n| = 2^{nR}$, then as $n \to \infty$,

$$R \ge R(P_0, D) + \frac{1}{2}\,n^{-1} \log n(1 + o(1)). \qquad (3.3)$$

Proof: By (3.1) and the fact that $R(P_0, \cdot)$ is decreasing,

$$\begin{aligned} R(P_0, E_{P_0} d_n(x^n, M_n(x^n)) &\le R(P_0, D(P_0, R) \\ &\quad + \frac{1}{2}\,\frac{\log n}{2|s_0|n}\,(1 + 0(1))) \\ &= R(P_0, D(P_0, R)) \\ &\quad + R'_D(P_0, D(P_0, R)) \\ &\quad \cdot \frac{\log n}{2|s_0|n}\,(1 + o'(1)), \end{aligned}$$

where the $o'(1)$ term represents the sum of $o(1)$ term in the previous expression and the smaller order terms from the Taylor expansion. From the parametric representation of $R(P_0, \cdot)$ (Berger [2]), it is easy to see that $s_0(R) = R'_D(P_0, D(P_0, R)) < 0$.

Also note that $R'_D < 0$, $E_{P_0} d_n(x^n, M_n(x^n)) \le D$, and $R(P_0, D(P_0, R)) = R$, so we have

$$R \ge R(P_0, E_{P_0} d_n(x^n, M_n(x^n)) + \frac{1}{2}\,n^{-1}\,\log n(1 + o'(1)).$$

Since $R(P_0, \cdot)$ is decreasing,

$$R(P_0, E_{P_0}(d_n(x^n, M_n(x^n)) \ge R(P_0, D).$$

This completes the proof of (3.3). □

Remark 1: Pilc's lower bound in Theorem 3 relies on some large deviation bounds from Shannon and Gallager [25], [26]. Those bounds are for tails of sums of i.i.d. variables and are accurate to the order $n^{-1/2}\ \exp(-cn)$ with the best constant c. Moreover, Pilc's original lower bound does not hold for noiseless coding because at $R = H(P_0)$, $s_0(R) = R'_D(P_0, D) = -\infty$. Hence, his lower bound does not include Rissanen's lower bound in the noiseless coding case as a special case.

Remark 2: From the previous section, we have a universal code which is *pointwise* D-semifaithful and $R(P_0, D) + O\big(n^{-1}\log n\big)$ in expected codelength. It would have been perfect if Pilc's result was in terms of expected code length and for pointwise D-semifaithful codes. However, Pilc's lower bound in the form of Corollary 2 is something like a dual to the result we seek; it says that for any *expected* D-semifaithful code, the log of the cardinality of the set of its code words is bounded below by $R(P_0, D) + O\big(n^{-1} \log n\big)$. Note that this log cardinality is not a random quantity, unlike the codelength of our universal code. For a pointwise D-semifaithful code, the log cardinality is likely to be bigger than the expected codelength.

V. Discussion

In this section, we compare the proofs of our Theorem 2 and Pilc's Theorem 3 from the points of view of constructiveness and universality. Both Pilc's upper bound (3.2) and our Theorem 2 involve a random coding argument. We might think that neither of them can give a code constructively, and that we can not really say which code is universal, since neither result looks constructive. On the other hand, we observe that the random coding argument in Theorem 2 does not need the true distribution P_0, since we proved the existence of a D-semifaithful code for each type class, while Pilc's random coding argument used a knowledge of P_0. This difference in random coding seems to suggest that our code might be universal, whereas Pilc's coding might not be so. We now show this to be the case.

A. Construction of a Universal D-Semifaithful Code when P_0 is Unknown

For each type class T_P^n, let $[Q] = [W] \cdot P$ as in Section II. $[W]$ can be obtained to any precision numerically (if not analytically) by Blahut's algorithm, Blahut [4]. We require a precision of order $(nP(j))^{-1}$. Then, the $n[Q](k)$ are integers, i.e., $[Q]$ is a type. Take m_n to be the integer part of $\exp\{nI([W], P) + 3JK\log(n+1)\}$. For this m_n, we could in principle search through all subsets of size m_n in $T_{[Q]}^n$ in order to find a subset $B_{P,D}$ that covers T_P^n within D distance. That is, for each subset B, we check whether $d(x^n, B) \leq D$ for all $x^n \in T_P^n$. For the m_n previously chosen, Theorem I guarantees the existence of such a pointwise D-semifaithful code. In other words, through exhaustive search, we can find at least one set $B_{P,D} \subset T_{[Q]}^n$ satisfying our D-covering requirement. We take the first such $B_{P,D}$ found as our codebook for T_P^n, and we have "constructed" a universal pointwise D-semifaithful code. Note that the code we just described has its code length approach the rate-distortion function lower bound at the rate $n^{-1}\log n$, and this rate is optimal in the noiseless coding case.

B. Construction of a D-Semifaithful Code when P_0 is Known

Pilc's upper bound (3.2) says that for any $R \in (0, H(P_0))$ and $\epsilon > 0$, we can use a random coding argument to find a map $M_n : \mathcal{A}_0^n \to \mathcal{B}_0^n$ such that as $n \to \infty$,

$$E_{P_0} d_n(x^n, M_n(x^n)) \leq D(P_0, R) + \left(\tfrac{1}{2} + \epsilon\right) \cdot \frac{\log n}{|s_0|n}(1 + o(1)).$$

When P_0 is known, for any fixed $D \in (0, D_{\max})$, we can take R_n to be $R(P_0, D) + \frac{1}{2}(1+\epsilon)n^{-1}\log n$. For this R_n, we can search through all subsets in $\mathcal{B}_0^n$ of size less than or equal to 2^{nRn}. We choose the codebook $\overline{B}_{p_0}$ as the set such that

$$E_{P_0} d\left(x^n, \overline{B}_{P_0}\right) = \min_B E_{P_0} d_n(x^n, B), \tag{4.1}$$

where the min is taken over all B with $|B| \leq 2^{nRn}$.

Pilc's result guarantees this codebook $\overline{B}_{P_0}$ satisfies

$$\begin{aligned} E_{P_0} d_n\left(x^n, \overline{B}_{P_0}\right) &\leq D(P_0, R_n) \\ &\quad + \frac{1}{2}(1+\epsilon)\frac{\log n}{|s_0|n}(1 + o(1)) \\ &= D(P_0, R(P_0, D)) - \frac{1}{2}(1+\epsilon)\frac{\log n}{|s_0|n} \\ &\quad + \frac{1}{2}(1+\epsilon)\frac{\log n}{|s_0|n} + o\left(n^{-1}\log n\right). \end{aligned}$$

The last equality holds because of the Taylor expansion of $D(P_0, \cdot)$, the fact that $D'(P_0, R(P_0, D)) = s_0^{-1}$, and $D' < 0$. It follows that

$$E_{P_0} d_n\left(x^n, \overline{B}_{P_0}\right) = D + o\left(n^{-1}\log n\right).$$

Without knowledge of the $o(1)$ term in Pilc's upper bound, $D + o\left(n^{-1}\log n\right)$ is the best level of distortion we can establish; we cannot deduce that the code is expected semifaithful at the exact level D.

The code $\overline{B}_{P_0}$ clearly depends on P_0 as we need to know P_0 to check (4.1). The $D + o\left(n^{-1}\log n\right)$-semifaithful code obtained from Pilc's upper bound is, therefore, not universal.

When P_0 is not known, a natural remedy would be to use the empirical distribution $\overline{P}$ instead of P_0 in the construction we have just outlined. But this does not work if we want to keep the rate $n^{-1}\log n$. The problem here is that when we replace P_0 by $\overline{P}$ in (4.1), we create an error of magnitude $(n^{-1}\log\log n)^{1/2}$ since $\|\overline{P} - P_0\| = O[(n^{-1}\log\log n)^{1/2}]$. This rate overwrites the desired rate $n^{-1}\log n$.

There is another difference between our code and Pilc's. The code $\overline{B}_{P_0}$ has the stated distortion on average, i.e., it is an *expected* D-semifaithful code, but the codelength is *pointwise* $R(P_0, D) + \frac{1}{2}(1+\epsilon)n^{-1}\log n$, not in expectation. On the other hand, our code $\{B_{P,D} : P \text{ any type}\}$ is *pointwise* D-semifaithful with the *expected* codelength $R(P_0, D) + (KJ + J + 4)n^{-1}\log n$. Ignoring the issue of nonuniversality, and the different constants in front of $n^{-1}\log n$, we might say that Pilc's result is dual to ours. We doubt that there exists a universal code that is *pointwise* D-semifaithful and whose log cardinality approaches the lower bound $R(P, D)$ at the rate $n^{-1}\log n$.

A technical difference of the two results concerns the mathematical tools employed. We both use the random coding argument, but the rate $n^{-1}\log n$ came out of the method of types for us, while for Pilc it came out of the large deviation results of Shannon and Gallager [25], [26]. This is not surprising, however, since large deviation results can be obtained using the method of types in the discrete memoryless source case. Due to the elegance of the method of types, our proofs are simpler and more direct than those of Pilc. Both results rely on the assumption that the source is i.i.d., although Pilc has results on noisy channels, too. Large deviation results do exist for independent not identical distributions, but we are not aware of any result as refined as that required by Pilc's bounds.

Acknowledgment

The authors would like to thank two referees for very helpful comments. The authors would like to thank Dr. T. Linder for pointing out that the boundedness condition on the second derivative of $R(D, P)$ in Theorem 2 is superfluous. This is because $-R'(P, D)$ is positive and bounded by $\log J/D$, as $R(P, D)$ is a convex function of D.

References

[1] A. R. Barron and T. M. Cover, "Minimum complexity density estimation," *IEEE Trans. Inform. Theory,* vol. 37, pp. 1034–1054, July 1991.

[2] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression.* Englewood Cliffs, NH: Prentice-Hall, 1971.

[3] T. Berger and L. D. Davisson, *Advances in Source Coding.* New York: Springer-Verlag, 1975.

[4] R. E. Blahut, *Principles and Practice of Information Theory.* Reading, MA: Addison-Wesley, 1978.

[5] B. S. Clarke and A. R. Barron, "Information theoretic asymptotics of Bayes methods," *IEEE Trans. Inform. Theory,* vol. 36, 453–471, May 1990.

[6] I. Csiszár and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems.* New York: Academic, 1981.

[7] T. M. Cover and J. A. Thomas, "Elements of information theory," Lecture Notes, Stanford Univ., CA, 1990.

[8] L. D. Davisson, "Minimax noiseless universal coding for Markov sources," *IEEE Trans. Inform. Theory,* vol. IT-29, pp. 211–215, 1983.

[9] L. Gerencser and J. Rissanen, "A prediction bound for Gaussian ARMA processes," *Proc. 25th CDC,* Athens, Greece, vol. 3, 1986, pp. 1487–1490.

[10] E. J. Hannan and L. Kavalieris, "Regression, autoregressive models," *J. Time Series Anal.,* vol. 7, pp. 27–49, 1986.

[11] E. M. Hemerly and M. H. A. Davis, "Strong consistency of the predictive least squares criterion for order determination of autoregressive processes," *Ann. Statist.,* vol. 17, pp. 941–946, 1989.

[12] J. Kieffer, "A unified approach to weak universal source coding," *IEEE Trans. Inform. Theory,* vol. IT-24, pp. 674–682, 1978.

[13] K. M. Mackenthun and M. B. Pursley, "Variable-rate universal block source coding subject to a fidelity constraint," *IEEE Trans. Inform. Theory,* vol. IT-24, pp. 349–360, 1978.

[14] D. L. Neuhoff, R. M. Gray, and L. D. Davisson, "Fixed rate universal block source coding with a fidelity criterion," *IEEE Trans. Inform. Theory,* vol. IT-21, pp. 511–523, 1975.

[15] D. S. Ornstein and P. C. Shields, "Universal almost sure data compression," *Ann. Probab.,* vol. 18, pp. 441–452, 1990.

[16] R. J. Pilc, "Coding theorems for discrete source-channel pairs," Ph.D. thesis, Dept. Elect. Eng., M.I.T., Cambridge, MA, 1967.

[17] ——, "The transmission distortion of a source as a function of the encoding block length," *Bell Syst. Tech. J.,* vol. 47, pp. 827–885, 1968.

[18] D. Pollard, *Convergence of Stochastic Processes.* New York: Springer-Verlag, 1984.

[19] J. Rissanen, "Stochastic complexity and modeling," *Ann. Statist.,* vol. 14, pp. 1080–1100, 1986.

[20] ——, "Complexity of strings in the class of Markov sources," *IEEE Trans. Inform. Theory,* vol. IT-34, pp. 526–532, July 1986.

[21] J. Rissanen, T. P. Speed, and B. Yu, "Density estimation by stochastic complexity," *IEEE Trans. Inform. Theory,* vol. 38, Pt. I, pp. 315–323, Mar. 1992.

[22] P. Shields, "Universal almost sure data compression using Markov types," *Probl. Contr. Inform. Theory,* vol. 19, pp. 269–277, 1990.

[23] C. Shannon and W. Weaver, *A Mathematical Theory of Communication.* Urbana, IL: Univ. Illinois Press, 1949.

[24] C. Shannon, "Coding theorems for a discrete source with a fidelity criterion," in *Information and Decision Processes,* R. E. Machol, Ed. New York: McGraw-Hill, 1959.

[25] C. Shannon and R. G. Gallager, "Lower bounds to error probability for coding on discrete memoryless channels I," *Inform. Contr.,* vol. 10, pp. 65–103, 1969.

[26] ——, "Lower bounds to error probability for coding on discrete memoryless channels II," *Inform. Contr.,* vol. 10, pp. 523–552, 1969.

[27] M. J. Weinberger, A. Lempel, and J. Ziv, "A sequential algorithm for the universal coding of finite memory sources," *IEEE Trans. Inform Theory,* vol. 38, pp. 1002–1014, May 1992.

[28] B. Yu and T. Speed, "Data compression and histograms," *Probability Theory Related Fields 92,* pp. 195–229, 1992.

The Annals of Statistics
1997, Vol. 25, No. 1, 169–185

INFORMATION AND THE CLONE MAPPING OF CHROMOSOMES

BY BIN YU[1] AND T. P. SPEED[2]

University of California, Berkeley

A clone map of part or all of a chromosome is the result of organizing order and overlap information concerning collections of DNA fragments called clone libraries. In this paper the expected amount of information (entropy) needed to create such a map is discussed. A number of different formalizations of the notion of a clone map are considered, and exact or approximate expressions or bounds for the associated entropy are calculated for each formalization. Based on these bounds, comparisons are made for four species of the entropies associated with the mapping of their respective cosmid clone libraries. All the entropies have the same first-order term $N \log_2 N$ (when the clone library size $N \to \infty$) as that obtained by Lehrach et al.

1. Introduction. The primary goal of the Human Genome Project is to sequence the entire human genome, which consists of about 3×10^9 base pairs (bp) of DNA. Current technology only permits sequencing of fragments of the order of a few hundred to a thousand base pairs of DNA in a single reaction. Consequently, much effort is devoted to fragmenting large DNA molecules, such as chromosomes, in such a way that the sequenced fragments can be readily assembled. Clone maps, which are one form of physical mapping, play a key role in this process, as well as providing a resource permitting the detailed study of chromosomal regions of biological interest.

A clone map of part or all of a chromosome is the result of organizing order and overlap information concerning collections of DNA fragments called clone libraries. Such libraries consist of many, typically thousands or tens of thousands, of DNA fragments from a chromosome or region of interest. Each fragment exists as an insert in an autonomously replicating DNA sequence, which resides within, and replicates with its host cells. In this manner it is possible to generate many copies of the fragment of interest, and the name clone is thus used as an abbreviation for the longer and more accurate name: cloned DNA fragment.

A large clone library might consist of 5000 cloned fragments of average length 100,000 base pairs, from a chromosome of length 100,000,000 base pairs. Assuming that the cloned fragments are randomly located along the

Received February 1995; revised April 1996.

[1] Research partially supported by ARO Grant DAAH04-94-G-0232 and NSF Grant DMS-93-22817.

[2] Research partially supported by NSF Grants DMS-91-13527 and DMS-94-04267.

AMS 1991 *subject classifications*. Primary 60K99; secondary 94A99.

Key words and phrases. Configuration variable, contig, entropy, random clone library.

chromosome, this would mean that any particular spot on the chromosome should be represented on an average of five cloned fragments, giving rise to the term fivefold coverage, or a five-hit library. We note that a library of fragments of this size is still not suitable for DNA sequencing. Typically, one or two further stages of subcloning are needed prior to sequencing, and there may be additional mapping at these stages as well. In such cases both the libraries and the fragments will be smaller, but the principles of mapping remain much the same. For details on clone mapping from the perspective of applied probability, see Lander and Waterman (1988). Nelson and Speed (1994) have a more statistical perspective, and give further references to these aspects of the topic.

In their paper comparing the relative merits of fingerprinting cloned fragments of DNA by hybridization of oligonucleotide probes and by digestion into restriction fragments, Lehrach et al. (1990) raised two interesting questions concerning the creation of clone maps of a chromosome: (1) how much information is needed? and (2) how much information is gained by the hybridization and restriction digestion methods, respectively? The answer to the first question offered by these authors was $\log_2(\frac{1}{2}N!)$ for a library of N clones. This figure corresponds to the average amount of information (the entropy, see the following discussion) required to identify the true ordering of N objects labeled $1, 2, \ldots, N$ when it is not possible to distinguish between the ordering $(i_1, i_2, \ldots, i_N)$ and its reverse $(i_N, \ldots, i_2, i_1)$, but otherwise all orderings are equally likely. However, it is not entirely clear why the ordering of objects in this way corresponds to any formal notion of a physical map, and even if there is such a correspondence, why all possible configurations should be equally likely.

To illustrate these points, let us briefly consider the cases of $N = 2$ and $N = 3$ clones, regarded mathematically as having identical length L bp and being randomly located along a chromosome of length G bp [cf. Lander and Waterman (1988)]. For two such clones we have two configurations, overlap or not, with quite unequal probabilities 2β and $1 - 2\beta$, respectively, where $\beta = L/G$. For three clones there are ten distinguishable configurations: one with no overlaps, three with exactly two clones overlapping, three with two different clone pairs overlapping, but no triple overlap and three distinguishable configurations involving a triple overlap. Again these can be seen to be far from equally probable. In practice, N will be in the hundreds or thousands.

In order to answer question (1) exactly, we would need to enumerate the set $\mathscr{X}$ of distinguishable configurations, calculate their probabilities $\{p(x): x \in \mathscr{X}\}$ and then go on to calculate the entropy $H(\mathbf{X}) = -\sum_{x \in \mathscr{X}} p(x) \log_2 p(x)$ of a random configuration $\mathbf{X}$. The first part of this program has been completed [see Newberg (1993)], but to our knowledge no one has carried the calculation of the probabilities beyond $N = 3$, although this is, in principle, possible. We do not know how to obtain the entropy $H(\mathbf{X})$ exactly, but in the following discussion we will find bounds on entropies of various configuration variables which are relevant to clone mapping.

INFORMATION AND CLONE MAPPING

The reason that the entropy $H(\mathbf{X})$ is the appropriate measure of information is explained in texts on information theory [see, e.g., Craig et al. (1990) and Rényi (1984)]. We content ourselves here with a brief informal explanation, applicable when the elements of $\mathscr{X}$ *are* equally likely, each having probability $1/|\mathscr{X}|$, in which case $H(\mathscr{X})$ achieves its upper bound $\log_2|\mathscr{X}|$. The argument goes like this: to identify any particular element $x \in \mathscr{X}$, we consider successive subdivisions of $\mathscr{X}$ into halves, quarters, eighths, and so on, and if we were told at each stage which half, quarter, eighth, and so forth contained the particular element, we would gain one bit of information each time. Clearly this process cannot finish in less than k steps, where $2^k \leq |\mathscr{X}| < 2^{k+1}$, and this k is thus a lower bound to the number of such questions, equivalently bits of information, necessary to identify the particular element in question. More refined procedures can limit the amount of information necessary to $\log_2|\mathscr{X}| + \varepsilon$, where $\varepsilon > 0$ is as small as we wish [see, e.g., Rényi (1984)]. A similar but more complicated argument applies when the elements of $\mathscr{X}$ are not equiprobable [see the discussion of the noiseless coding theorem in Cover and Thomas (1991)].

In this paper we study the entropy $H(\mathbf{X})$ of a random configuration $\mathbf{X}$ most appropriate to the clone mapping problem. The study is done through seven other random structures, $\mathbf{P}$, $\mathbf{Q}$, $\mathbf{U}$, $\mathbf{V}$, $\mathbf{W}$, $\mathbf{Y}$ and Z, each of which can be regarded as embodying a greater or lesser amount of the structure implicit in $\mathbf{X}$, but whose entropies are more accessible. We derive a variety of exact and approximate expressions and lower and upper bounds for the entropies of these quantities. We compute these bounds for clone libraries of interest and the bounds are reasonable for all configuration variables considered and very tight for some. Based on these computations, comparisons are made for four "model" species in terms of information needed for the mapping of their respective cosmid clone libraries. It is somewhat surprising that all the entropies have the same first-order term $N\log_2 N$ when $N \to \infty$, as that obtained in Lehrach et al. (1990). We end the paper with some remarks concerning the more difficult question 2.

In closing this brief introduction we note that in the analysis which follows we essentially ignore the role of distances, although we do consider the placement variable $\mathbf{W}$ in units of thousands of base pairs. Many physical mapping methods produce some information concerning distances as well as clone order, and such information can be very useful in practice, even when (as is often the case) there are large error bounds attached. In particular, it would be misleading to compare the hybridization and restriction digest methods mentioned previously, solely on the basis of the information they produce concerning clone order. The restriction digest method produces fairly precise information about distances, whereas the hybridization method does not. An analysis, which incorporates distance as well as order and overlap information, is beyond us at this time.

2. What is a clone map? We now introduce several different but related abstractions of the notion of a clone map of a chromosome, this being informally an ordering of a library of cloned fragments of the chromosome in question. As noted previously, we adopt the mathematical model for a clone library used in Lander and Waterman (1988), namely, that the N cloned fragments can be identified with N randomly located subintervals of equal length L of a genome of length G. More formally, the left-hand endpoints (say) of the N intervals corresponding to the cloned fragments are independently located uniformly along $[0, G - L]$. It will be convenient at points in the argument to take an alternative, effectively equivalent view of the left-hand endpoints as being the points on $[0, G - L]$ of a homogeneous Poisson process with rate $\lambda = N/G$ per base pair.

2.1. *Fully ordered configurations.* Following the terminology of Alizadeh, Karp, Newberg and Weisser (1993), we use the term *placement* to describe a configuration of positions of the clones along the chromosome, that is, a specification $\mathbf{W} = (W_1, W_2, \dots, W_N)$, where $W_i \in [0, G - L]$ is the location of the left-hand endpoint of the ith cloned fragment, $i = 1, 2, \dots, N$. The units here are base pairs (bp) or kilobase pairs (kb); see the following discussion. Experimental procedures exist which could precisely determine these locations for a clone library, but most clone mappings have more modest aims, seeking to single out a less completely specified configuration from among a class of a priori equivalent alternatives. Before we turn to a discussion of such "coarser" configurations, we make a connection with the work of Lehrach et al. (1990), which stimulated this research. By the *linear ordering* of a clone library, we mean the sequence $\mathbf{V} = (V_1, V_2, \dots, V_N)$ of labels of the ordered left-hand endpoints of the clones; equivalently, the vector of *ranks* of $\mathbf{W} = (W_1, W_2, \dots, W_N)$ listed in reverse order. This variable seems to be the one considered in Lehrach et al. (1990).

2.2. *Island configurations.* We turn now to a second class of clone configurations, those based on the notion of an *island*, which is either a single clone, not overlapping with any other clone in the library, or a set of clones, each pair of which is connected by a chain of overlapping pairs of clones. Islands of two or more clones are usually called *contigs*, and many clone mapping projects have as their initial objective the determination of all contigs in their library and the ordering, up to inversion, of clones within contigs. This is usually the objective of *fingerprint-based* clone mapping, which attempts to infer clone order and overlap from information concerning each of the clones in the library, such as the list of fragment lengths following digestion by restriction enzymes, or the pattern of hits and misses following hybridization with a panel of probes. Fingerprint-based clone mapping projects usually turn to quite different techniques such as radiation hybrid or fluorescence in situ hybridization (FISH) mapping [see, e.g., Cox, Burmeister, Price and Myers (1990) and Trask (1991)].

The most basic island configuration variable is Z, the *number* of islands. More informative is the variable $\mathbf{U} = (U_1, U_2, \dots, U_N)$ of *island sizes*, which is a *partition of the integer N*, that is, $\Sigma_1^N U_i = Z$, $\Sigma_1^N iU_i = N$; or, equivalently,

U_i is the number of islands containing i clones. The components of $\mathbf{U}$ are the multiplicities of the *block sizes* of the *partition* $\mathbf{Q}$ *of the set* $\{1, 2, \ldots, N\}$ of clone labels into islands. Here $\mathbf{Q}$ is the unordered list of disjoint subsets of $\{1, 2, \ldots, N\}$, usually called blocks or equivalence classes, but called islands in this context, whose union is $\{1, 2, \ldots, N\}$.

More informative again than $\mathbf{Q}$ is the configuration variable we term the *distinguishable orderings* of the clones and denote by $\mathbf{Y}$, namely, the variable which refines $\mathbf{Q}$ by including information on the ordering of clones within contigs, up to inversion. Thus $\mathbf{Y}$ tells us which clones are together in a contig and, up to a flip, the order in which they appear, but it contains no information on the relative positions of distinct islands along the genome.

There is one last refinement which we mention, namely, the configuration variable discussed in Newberg (1993), which includes information on the depth of coverage within contigs. We denote this configuration variable by $\mathbf{X}$, and note that it may be regarded as refining $\mathbf{Y}$ by containing not just information on the labels of the left-hand endpoints of the clones within each contig, up to inversion, but the labels of the interleaved sequence of the left-hand and right-hand endpoints of the clones, again up to inversion. Newberg (1990) calls two configurations of clones *topologically similar* if one can be transformed into the other by permuting the islands and/or reflecting some of the islands. An adjustment of the amount by which any pair of clones overlap leaves one with a topologically similar clone ordering, if no endpoint of a clone is moved past an endpoint of another clone. With this definition, $\mathbf{X}$ is the set of equivalence classes of topologically distinct configurations, called *interleavings* in Newberg (1993) and Alizadeh, Karp, Newberg and Weisser (1993).

2.3. *Pairwise overlaps.* Many fingerprint-based clone mapping projects take as their starting points the determination of pairwise overlaps among the clones in their library [see, e.g., Branscomb et al. (1990), Craig et al. (1990) and Fu, Timberlake and Arnold (1992)]. For this reason we define the *pairwise overlap* variable $\mathbf{P} = (P_{ij}\colon 1 \le i < j \le N)$, where $P_{ij} = 1$ if clones i and j overlap, and $P_{ij} = 0$ otherwise. It is clear that $\mathbf{P}$ can be obtained from $\mathbf{X}$ but not from $\mathbf{Y}$. In seeking to estimate $H(\mathbf{P})$ we do not mean to imply that pairwise comparisons are the best, or even an effective way to ascertain pairwise overlap information. Indeed, many of the most common clone mapping methods, such as STS-content mapping [Green and Green (1991)] and restriction mapping [Olson et al. (1986)], do not attempt to determine pairwise overlaps at all. Nevertheless, it seems to us of interest to ask just how large $H(\mathbf{P})$ is in relation to the entropies of other, more refined configuration variables.

This concludes our discussion of the different abstractions of the notion of a clone map of a chromosome based on a library of cloned DNA fragments from that chromosome. As with all mathematical idealizations, our variables all fail to account for many features of real clone mapping projects. Our hope is that the features we do retain are the important ones, and that our results are at least qualitatively correct and useful. We now illustrate the different variables just introduced in a simple case.

EXAMPLE. Suppose that $G = 150$, $L = 20$ and $N = 8$. We list the set of configuration variables refining $\mathbf{W} = (120, 50, 10, 45, 105, 55, 20, 76)$. The vector of *ranks* of these values, viewed as observations on $[0, 150]$, is $(1, 5, 8, 6, 2, 4, 7, 3)$, and so $\mathbf{V} = (3, 7, 4, 2, 6, 8, 5, 1)$. Using the values in $\mathbf{W}$, it is easy to ascertain that

$$\mathbf{X} = \{(373'7')^*, (4264'2'6')^*, (88')^*, (515'1')^*\},$$

where 3 (resp. 3′) denotes the left-hand (resp. right-hand) end of clone 3 or vice versa, and * indicates the fact that the ordering is only unique up to reversal. In a similar notation we have

$$\mathbf{Y} = \{(37)^*, (426)^*, (8), (51)^*\},$$

while $\mathbf{Q} = 15|246|37|8$, $\mathbf{U} = (1^1, 2^2, 3^1)$ and $Z = 4$.

3. Results. In this section we present our approximations to the entropy of the configurations just described. All proofs are collected in the appendices.

We have sought close nonasymptotic upper and lower bounds to the entropy expressions of interest, and have been quite successful in this regard with $H(\mathbf{Q})$ and $H(\mathbf{Y})$, and somewhat less so with $H(\mathbf{X})$ and $H(\mathbf{P})$. Exact calculations of $H(\mathbf{W})$ and $H(\mathbf{V})$ are straightforward. It is also of interest to consider our results asymptotically as $N \to \infty$. In so doing, we could keep L/G fixed and let $c = NL/G$ increase, or we could keep c fixed and let L/G decrease. A value of c in the range 3–10 is typical, with $c = 5$ being quite common, although values in the range 40–50 have been used. Our figures and tables have c fixed at 5.

The easiest entropy to evaluate is $H(\mathbf{W})$ which is just $N \log_2(G - L) \approx N \log_2 G$. This last expression can be rewritten as

$$H(\mathbf{W}) = N \log_2 N + N \log_2(L/c)$$

by making the substitution $c = NL/G$. It is clear that the leading term is $N \log_2 N$, and also that the second term depends on the units in which L is measured. The most reasonable choice would seem to be kilobase pairs (kb), in which the values $G = 100{,}000$ kb, $L = 40$ kb (corresponding to a cosmid library) and $c = 5$ give $N = 12{,}500$ and $H(\mathbf{W}) = 2.1 \times 10^5$, compared with $N \log_2 N = 1.7 \times 10^5$.

As pointed out in Lehrach et al. (1990), we may use Stirling's formula to get

$$\begin{aligned} H(\mathbf{V}) &= \log_2(\tfrac{1}{2}N!) \\ &\approx N \log_2 N + \tfrac{1}{2}\log_2 N - (\log_2 e)N - \log_2(\sqrt{2\pi}) - 1. \end{aligned}$$

INFORMATION AND CLONE MAPPING

Now let us define

$$\overline{L}(\mathbf{U}) = \mathbb{E}\{Z\}[\log_2 N - \log_2 e] + \tfrac{1}{2}\log_2 \mathbb{E}\{Z\} + \log_2\left(\sqrt{2\pi}e^{1/12}\right),$$

$$\overline{L}(\mathbf{U}) = \mathbb{E}\{Z\}\left(\log_2 \frac{Np^2}{q(1-q^N)} + (\log_2 q)c_N\right),$$

$$\underline{M}(\mathbf{U}) = Ne^{-c}(a_N + b_N) - (\log_2 e)N,$$

$$\overline{M}(\mathbf{U}) = Ne^{-c}(a_N + b_N) - (\log_2 e)N + \left(\log_2\left(\sqrt{2\pi}e^{1/12}\right)\right)\mathbb{E}\{Z\},$$

where $a_N = \mathbb{E}\{F^N \log_2 F^N\}$, $b_N = \frac{1}{2}\mathbb{E}\{\log_2 F^N\}$ and $c_N = \mathbb{E}\{F^N\}$, and F^N is a truncated geometric random variable with $p = e^{-c}$ and truncation at N. That is, for $q = 1 - p$, $P(F^N = j) = pq^{j-1}/(1 - q^N)$, $j = 1, 2, \ldots, N$. We have the following bounds on the entropies.

Result A (Finite-sample entropy bounds). Let us introduce the following abbreviations:

$$\underline{H}(\mathbf{Y}) = \log_2 N! - \overline{L}(\mathbf{U}) - Np(1-p),$$

$$\overline{H}(\mathbf{Y}) = \log_2 N! - \underline{L}(\mathbf{U})^+ - Np(1-p) + \mathbb{E}\{Z\}H(F^N) + \log_2 N,$$

$$\underline{H}(\mathbf{Q}) = \log_2 N! - \overline{L}(\mathbf{U}) - \overline{M}(\mathbf{U}),$$

$$\overline{H}(\mathbf{Q}) = \log_2 N! - \underline{L}(\mathbf{U})^+ - \underline{M}(\mathbf{U}) + \mathbb{E}\{Z\}H(F^N) + \log_2 N,$$

$$\overline{H}(\mathbf{X}) = N\log_2 N + N\log_2(4/e) - \log_2 N,$$

$$\underline{H}(\mathbf{X}) = \underline{H}(\mathbf{Y}),$$

$$\overline{H}(\mathbf{P}) = \overline{H}(\mathbf{X}),$$

$$\underline{H}(\mathbf{P}) = \underline{H}(\mathbf{Q}),$$

where

$$H(F^N) = \sum_{j=1}^{N} -P(F^N = j)\log_2 P(F^N = j).$$

Then our main bounds may be expressed as

$$\underline{H}(S) \le H(S) \le \overline{H}(S),$$

where S may be $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Q}$ or $\mathbf{P}$.

Result B (Asymptotic expansions for entropies). The following expressions are valid as $N \to \infty$:

(i) $$H(\mathbf{W})/N\log_2 N = 1 + o(1),$$

(ii) $$H(\mathbf{V})/N\log_2 N = 1 + o(1),$$

(iii) $$(1 - e^{-c}) + o(1) \le H(\mathbf{X})/N\log_2 N \le 1 + o(1),$$

$$
\begin{aligned}
&\text{(iv)} && H(\mathbf{Y})/N\log_2 N = (1 - e^{-c}) + o(1),\\
&\text{(v)} && H(\mathbf{Q})/N\log_2 N = (1 - e^{-c}) + o(1),\\
&\text{(vi)} && H(\mathbf{P})/N\log_2 N = (1 - e^{-c}) + o(1).
\end{aligned}
$$

The finite-sample bounds in Result A are really only useful when they are not very far apart. Fortunately, they are reasonably close for all four configuration variables considered here and very close for **Y** and **Q**. Figure 1 is the log-log plot of the entropy bounds for $c = 5$ and $N = 100,\dots,20{,}000$, and it is clear that the bounds are very tight for $H(\mathbf{Y})$ and $H(\mathbf{Q})$, tight for $H(\mathbf{X})$, but not so close for $H(\mathbf{P})$. It is also comforting to see that **W**, **X** and **Y**, which are all reasonable definitions of a clone map, turn out to have very similar entropies. The other interesting and useful observation is that $H(\mathbf{V})$ is numerically very close to $H(\mathbf{Y})$ for the range of N that we considered and for $c = 5$. Therefore, the simple Stirling expansion for $H(\mathbf{V})$ can be used as a valid short-hand formula for $H(\mathbf{Y})$ when $c = 5$. This shows that Lehrach et al.'s intuition works well here since the coverage is high enough that most of the randomness in the configuration variable **Y** comes from the permutation which is captured in **V**.

It is perhaps remarkable that the entropies of **W**, **V**, **X**, **Y**, **Q** and **P** all turn out to have the first-order term $N\log_2 N$, asymptotically, as obtained in Lehrach et al. (1990) (cf. Result B). Moreover, the constant for the first-order

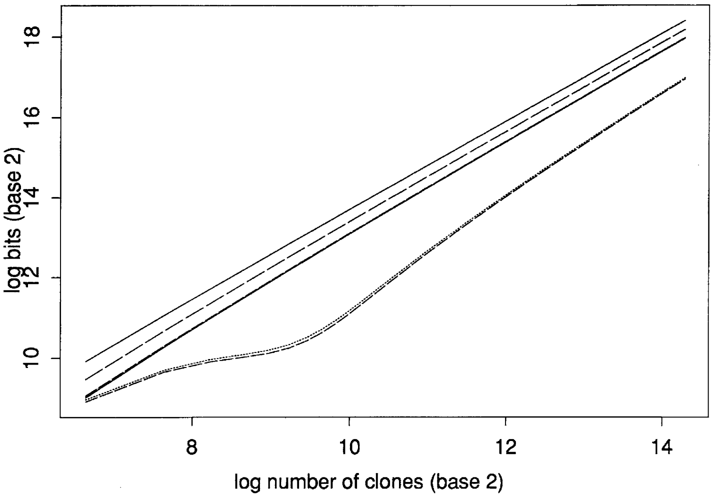

FIG. 1. $\log_2 H(\mathbf{W})$ *(top line)*; $\log_2 \overline{H}(\mathbf{X})$ *(second line from top)*; $\log_2 \overline{H}(\mathbf{Y})$, $\log_2 H(\mathbf{V})$ *and* $\log_2 \underline{H}(\mathbf{Y})$ *in the third line (cluster) and in that order from top*; $\log_2 \overline{H}(\mathbf{Q})$ *and* $\log_2 \underline{H}(\mathbf{Q})$ *in the bottom line (cluster) and in that order from top. Here the basic unit for* **W** *is* kb, $L = 40$ kb *and* $c = 5$. $\log_2 \underline{H}(\mathbf{Y})$ *and* $\log_2 \overline{H}(\mathbf{X})$ *serve as lower and upper bounds for* $\log_2 H(\mathbf{X})$ *and* $\log_2 \underline{H}(\mathbf{Q})$ *and* $\log_2 \overline{H}(\mathbf{X})$ *serve as lower and upper bounds for* $\log_2 H(\mathbf{P})$.

INFORMATION AND CLONE MAPPING

terms of $H(\mathbf{Y})$, $H(\mathbf{Q})$ and $H(\mathbf{P})$ is the same, namely, $1 - e^{-c}$. Unfortunately, this asymptotic result is not so useful for the values of N which are relevant here, because the term which makes the difference between $H(\mathbf{Y})$ and $H(\mathbf{Q})$ (cf. Figure 1) is $M(\mathbf{U})$, which is $O(N)$. The problem is that $\log_2 N$ is asymptotically larger than any constant term, but in this case it is much smaller than the corresponding constant ($\sim$ 260) in the $O(N)$ term.

An interesting fact which follows from the entropy bounds is that $H(\mathbf{P})/H(\mathbf{X}) \geq 0.20$ for $c = 5$ and $N = 100, 200, \ldots, 20{,}000$ (cf. Figure 2). (Note that the turns on the ratios for small N are probably artifacts of our bounds, not indicative of the true ratios of the entropies.) This implies that the pairwise variable **P** contains a substantial proportion of the information in the interleaving variable **X**. However, although the pairwise mapping approach is definitely a good starting point for any clone mapping effort, recovering the pairwise variable **P** efficiently may well be improved by using multiple comparisons.

Table 1 lists the entropy bounds for specific cosmid ($L = 40$ kb) clone libraries corresponding to the G for a bacterium *E. coli*, yeast *S. cerevisiae*, roundworm *C. elegans* and humans. Here we observe behavior similar to that found in the figures. Table 2 gives the bounds on $H(\mathbf{W})$, $H(\mathbf{X})$ and $H(\mathbf{Y})$ for the last three species in relation to those of the bacterium *E. coli*. The ratios are seen to be species specific rather than specific to the configuration variables. We conclude that it makes sense to say, for example, that cosmid clone mapping for the roundworm requires about 40 times as much information as that for the bacterium *E. coli*, and that such mapping for humans

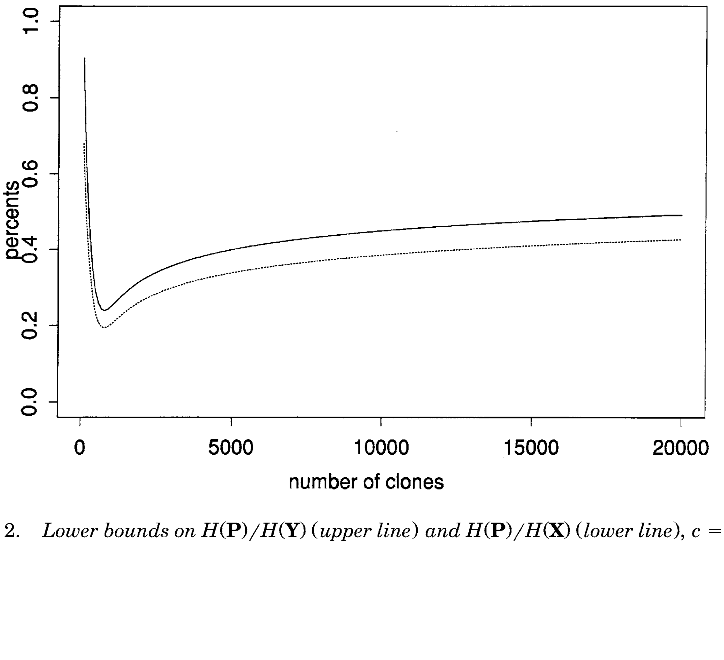

FIG. 2. *Lower bounds on* $H(\mathbf{P})/H(\mathbf{Y})$ *(upper line) and* $H(\mathbf{P})/H(\mathbf{X})$ *(lower line),* $c = 5$.

TABLE 1
Entropies and ratios for fivefold cosmid clone libraries of four species (ratios based on unrounded figures)

	Bacterium $N = 500$	**Yeast** $N = 1{,}875$	**Roundworm** $N = 12{,}500$	**Human** $N = 375{,}000$
$H(\mathbf{W})$	6.0×10^3	2.6×10^4	2.1×10^5	8.1×10^6
$H(\mathbf{V})$	3.8×10^3	1.8×10^4	1.5×10^5	6.4×10^6
$\underline{H}(\mathbf{X})$	3.7×10^3	1.8×10^4	1.5×10^5	6.4×10^6
$\overline{H}(\mathbf{X})$	4.8×10^3	2.1×10^4	1.8×10^5	7.2×10^6
$\underline{H}(\mathbf{X})/\overline{H}(\mathbf{X})$	0.79	0.82	0.85	0.89
$\underline{H}(\mathbf{Y})$	3.7×10^3	1.8×10^4	1.5×10^5	6.4×10^6
$\overline{H}(\mathbf{Y})$	3.8×10^3	1.8×10^4	1.5×10^5	6.4×10^6
$\underline{H}(\mathbf{Y})/\overline{H}(\mathbf{Y})$	0.99	0.99	0.99	0.99
$\underline{H}(\mathbf{Q})$	1.1×10^3	5.5×10^3	7.0×10^4	3.9×10^6
$\overline{H}(\mathbf{Q})$	1.2×10^3	5.7×10^3	7.2×10^4	4.0×10^6
$\underline{H}(\mathbf{Q})/\overline{H}(\mathbf{Q})$	0.95	0.96	0.98	0.99
$\underline{H}(\mathbf{P})$	1.1×10^3	0.6×10^4	0.7×10^5	3.9×10^6
$\overline{H}(\mathbf{P})$	4.8×10^3	2.1×10^4	1.8×10^5	7.2×10^6
$\underline{H}(\mathbf{P})/\overline{H}(\mathbf{P})$	0.23	0.26	0.40	0.55

requires about 1500 times as much information as that for the bacterium *E. coli*.

4. Final comments. We close our discussion with some brief remarks on the important question (2) raised in Section 1: how much information is gained by the hybridization and restriction digestion methods, respectively? It is not our intention to offer a thorough discussion of this topic here, as we hope to present something more complete in a future paper. Rather, our aim here is simply to point out that the situation is not quite as simple as the discussion in Lehrach et al. (1990), page 45, suggests.

Suppose that we collect data $D_1, D_2, \ldots, D_n$ on our clone library, for example, D_n could be the pattern of responses of our clones ($+$ or $-$) to the nth in a sequence of hybridization with short oligonucleotides. Each such

TABLE 2
Entropies of $\mathbf{W}, \mathbf{X}$ *and* $\mathbf{Y}$ *relative to E. coli*, $c = 5$

	Yeast	**Roundworm**	**Human**
$H(\mathbf{W})$	4.3	35	1350
$\overline{H}(\mathbf{X})$	4.5	37	1505
$\underline{H}(\mathbf{X})$	4.7	40	1700
$\overline{H}(\mathbf{Y})$	4.7	40	1690
$\underline{H}(\mathbf{Y})$	4.7	40	1700

data item has an entropy $H(D_n)$, indeed the full collection has an entropy $H(D_1, D_2, \ldots, D_n)$, but if our aim is constructing a clone map using these data, the relevant entropy is $H(\mathbf{X}|D_1, D_2, \ldots, D_n)$, the conditional entropy of the library configuration **X** given the data $D_1, D_2, \ldots, D_n$. The computation of this quantity is not at all straightforward, even if the data items $D_1, D_2, \ldots, D_n$ are mutually independent and identically distributed, given **X**, as might be the case with a sequence of hybridizations involving short oligonucleotides of the same length. In such a case $H(D_1, D_2, \ldots, D_n) = nH(D_1)$, but no such simplification occurs for $H(\mathbf{X}|D_1, \ldots, D_n)$, although it should be possible to determine the asymptotic behavior of this quantity as $n \to \infty$. In a future paper we hope to discuss this issue more fully.

APPENDIX A

Upper and lower bounds for H(Q) and H(Y). Let $\mathbf{u} = (1^{u_1}, 2^{u_2} \ldots)$ be a partition of the number N, and suppose that $\sum_1^N u_i = z$. We will use the notation $\mathbf{U}(\cdot)$ to denote the partition of N associated with the configuration in parentheses.

LEMMA A.1. *The number of configurations* **Q** *for which* $\mathbf{U}(\mathbf{Q}) = \mathbf{u}$ *is*

$$\frac{N!}{\Pi_{i=1}^N (i!)^{u_i} u_i!}. \tag{A.1}$$

PROOF. This is well known [see, e.g., Aigner (1979)].

LEMMA A.2. *The number of configurations* **Y** *for which* $\mathbf{U}(\mathbf{Y}) = \mathbf{u}$ *is*

$$\frac{N!}{\Pi_{i=1}^N u_i!} \frac{1}{2^{z-u_1}}. \tag{A.2}$$

PROOF. It is clear that the number we seek in this lemma is the number (A.1) multiplied by the number of directionless permutations of clones within islands. However, the latter is just

$$\prod_{i=2}^N \left(\tfrac{1}{2} i!\right)^{u_i}$$

and the result follows once we note that $\sum_{i=2}^N u_i = z - u_1$. □

LEMMA A.3. *The configurations* **Y** *with* $\mathbf{U}(\mathbf{Y}) = \mathbf{u}$ *are equally likely.*

PROOF. By symmetry.

EXAMPLE. It is easy to see that the configurations $\mathbf{y}_1 = \{(37)^*, (426)^*, (8), (51)^*\}$ and $\mathbf{y}_2 = \{(32)^*, (785)^*, (4), (16)^*\}$, for example, are equiprobable.

COROLLARY A.1.

(i) $$H(\mathbf{Q} \mid \mathbf{U}) = \log_2 N! - L(\mathbf{U}) - M(\mathbf{U}),$$

(ii) $$H(\mathbf{Y} \mid \mathbf{U}) = \log_2 N! - L(\mathbf{U}) - Np(1-p),$$

where

$$L(\mathbf{U}) = \mathbb{E}\left\{\log_2 \prod_{i=1}^{N} U_i!\right\} \tag{A.3}$$

and

$$M(\mathbf{U}) = \mathbb{E}\left\{\log_2 \prod_{i=1}^{N} (i!)^{U_i}\right\}. \tag{A.4}$$

PROOF. These relations are consequences of Lemmas A.1 and A.2 and the equiprobable assertion of Lemma A.3.

We turn now to obtaining upper and lower bounds $\overline{L}(\mathbf{U})$, $\overline{M}(\mathbf{U})$ and $\underline{L}(\mathbf{U})$, $\underline{M}(\mathbf{U})$ of $L(\mathbf{U})$ and $M(\mathbf{U})$. In the calculations that follow, we use upper and lower bounds for factorials easily obtained from Stirling's formula [see, e.g., Feller (1968), page 52]

$$n^{n+1/2}e^{-n} \le n! \le n^{n+1/2}e^{-n}\sqrt{2\pi}e^{1/12}. \tag{A.5}$$

We also make use of the readily proved fact that the distribution of the sizes of islands is a truncated geometric with probability $p = e^{-c}$, where $c = NL/G$. More fully, the (ordered) sequence $F_1^N, F_2^N, \dots$ of island sizes consists of identically distributed random variables with common distribution $\mathrm{pr}(F^N = i) = pq^{i-1}/(1-q^N)$, $i = 1, 2, \dots, N$. Lander and Waterman (1988) give the proof for N large in which case F^N is approximated by a geometric. Taking the truncation into account gives more accurate results in our bounds when N is in the hundreds. It follows that $\mathbb{E}(Z - U_1) = Ne^{-c}(1 - e^{-c})$, since, for $i = 1, 2, \dots, N$,

$$\begin{aligned}\mathbb{E}U_i &= \mathbb{E}\sum_{j=1}^{Z} I_{\{F_j^N = i\}} = \mathbb{E}\{Z\}P(F^N = i)\\ &= np^2q^{i-1}/(1-q^N).\end{aligned}$$

[More precisely, $\mathbb{E}U_i \approx np^2q^{i-1}/(1-q^N)$, since Z is very weakly related to the sequence $\{I_{\{F_j^N = i\}}\}$ $j = 1, 2, \dots$. Equality holds if Z is independent of this sequence.] We note that the preceding approximations are not expected to work for very small N's, but we believe they do work when N is in the hundreds, say larger than 500.

LEMMA A.4.

$$\underline{L}(\mathbf{U})^+ \le L(\mathbf{U}) \le \overline{L}(\mathbf{U}),$$

where $x^+ = \max\{x, 0\}$,

$$\overline{L}(\mathbf{U}) = \mathbb{E}\{Z\}[\log_2 N - \log_2 e] + \tfrac{1}{2}\log_2 \mathbb{E}\{Z\} + \log_2\left(\sqrt{2\pi}e^{1/12}\right)$$

and

$$\underline{L}(\mathbf{U}) = \mathbb{E}\{Z\}[\log_2 N - (2c+1)\log_2 e + (e^c - 1)\log_2(1 - e^{-c})].$$

Proof. Since $\Sigma_1^N U_i = Z$, we must have

$$\binom{Z}{U_1 \quad U_2 \quad \cdots} \geq 1,$$

in which case

$$\begin{aligned}\mathbb{E}\{\log_2 \Pi_i U_i!\} &\leq \mathbb{E}\{\log_2 Z!\}\\ &\leq \mathbb{E}\{(Z + \tfrac{1}{2})\log_2 Z - (\log_2 e)Z + \log_2(\sqrt{2\pi} e^{1/12})\}.\end{aligned}$$

Now $Z \leq N$, and so the right-hand side of the preceding formula is

$$\leq \mathbb{E}\{Z\}\log_2 N + \tfrac{1}{2}\mathbb{E}\{\log_2 Z\} - (\log_2 e)\mathbb{E}\{Z\} + \log_2(\sqrt{2\pi} e^{1/12}),$$

which is just the expression $\overline{L}(\mathbf{U})$.

For the lower bound $\underline{L}(\mathbf{U})$ we argue as follows:

$$\begin{aligned}L(\mathbf{U}) &= \sum_{i=1}^{N} \mathbb{E}\{\log_2 U_i!\}\\ &\geq \sum_{i=1}^{N} \mathbb{E}\{U_i \log_2 U_i - (\log_2 e)U_i\}\\ &= \sum_{i=1}^{N} \mathbb{E}\{U_i \log_2 U_i\} - (\log_2 e)\mathbb{E}\{Z\} \quad \text{since } \textstyle\sum U_i = Z\\ &\geq \sum_{i=1}^{N} \mathbb{E}\{U_i\}\log_2 \mathbb{E}\{U_i\} - (\log_2 e)\mathbb{E}\{Z\} \quad \text{since } x\log_2 x \text{ is convex.}\end{aligned}$$

Now $\mathbb{E}\{U_i\} = Np^2q^{i-1}/(1-q^N)$ where $p = e^{-c}$ and $q = 1 - p$ and so, continuing the preceding sequence of inequalities,

$$\begin{aligned}L(\mathbf{U}) &\geq \sum_{i=1}^{N} Np^2q^{i-1}\left[\log_2 \frac{Np^2}{1-q^N} + (i-1)\log_2 q\right] - (\log_2 e)\mathbb{E}\{Z\}\\ &= Np \log_2 \frac{Np^2}{q(1-q^N)} + Np(\log_2 q)c_N - (\log_2 e)\mathbb{E}\{Z\}\\ &= \mathbb{E}\{Z\}\log_2 \frac{Np^2}{q(1-q^N)} + (\log_2 q)c_N - (\log_2 e)\mathbb{E}\{Z\},\end{aligned}$$

which is seen to be $\underline{L}(\mathbf{U})$ once we recall that $\mathbb{E}\{Z\} = Ne^{-c}$ and $c_N = \mathbb{E}F^N$. This completes the proof of Lemma A.4. Note that the leading term in each case is $e^{-c}N\log_2 N$. Obviously, $U \geq 0$. Hence $L(\mathbf{U}) \geq L^+(\mathbf{U})$. □

In the following lemma a_N and b_N are moments $\mathbb{E}\{F^N \log_2 F^N\}$ and $\frac{1}{2}\mathbb{E}\{\log_2 F^N\}$, where F^N has a truncated geometric distribution with parameter $p = e^{-c}$, $c = NL/G$, and truncation at N.

LEMMA A.5.

$$\underline{M}(\mathbf{U}) \le M(\mathbf{U}) \le \overline{M}(\mathbf{U}),$$

where

$$\underline{M}(\mathbf{U}) = Ne^{-c}(a_N + b_N) - (\log_2 e)N$$

and

$$\overline{M}(\mathbf{U}) = Ne^{-c}(a_N + b_N) - (\log_2 e)N + \log_2(\sqrt{2\pi}e^{1/12}) \times \mathbb{E}\{Z\}.$$

PROOF. By definition,

$$M(\mathbf{U}) = \mathbb{E}\left\{\log_2 \prod_{i=1}^{N} (i!)^{U_i}\right\}$$
$$= \mathbb{E}\left\{\sum_i U_i \log_2 i!\right\}.$$

We first use the lower bound of (A.5), obtaining

$$M(\mathbf{U}) \ge \mathbb{E}\left\{\sum_{i=1}^{N} U_i(i + \tfrac{1}{2})\log_2 i - (\log_2 e)i\right\}$$
$$= \sum_{i=1}^{N} \big(i \log_2 i + \tfrac{1}{2}\log_2 i\big)\mathbb{E}\{U_i\} - (\log_2 e)N \quad \text{since } \sum_{i=1}^{N} iU_i = N.$$

Now $\mathbb{E}(U_i) = Np^2q^{i-1}/(1-q^N)$ as before.

To complete this, we need to recall that

$$\mathbb{E}\{F^N \log_2 F^N\} = \sum_{i=1}^{N} i(\log_2 i)\frac{pq^{i-1}}{1-q^N}$$

and

$$\frac{1}{2}\mathbb{E}\{\log_2 F\} = \frac{1}{2}\sum_{i=1}^{N} (\log_2 i)\frac{pq^{i-1}}{1-q^N}.$$

As mentioned in the statement of the lemma, these will be denoted by a_N and b_N, respectively, giving

$$M(\mathbf{U}) \ge Ne^{-c}(a_N + b_N) - (\log_2 e)N = \underline{M}(\mathbf{U}).$$

Turning now to the upper bound, the same reasoning leads to

$$M(\mathbf{U}) \le Ne^{-c}(a_N + b_N) - (\log_2 e)N + \big(\log_2(\sqrt{2\pi}e^{1/12})\big)\mathbb{E}\{Z\},$$

where we have used the fact that $\sum_i U_i = Z$. However, the right-hand side of the preceding formula is just $\overline{M}(\mathbf{U})$ and we are finished. □

LEMMA A.6.

$$0 \le H(\mathbf{U}|Z) \le \mathbb{E}\{Z\}H(F^N).$$

PROOF.

$$H(\mathbf{U}|Z) = \sum_k \mathrm{pr}(Z = k)H(\mathbf{U}|Z = k)$$

and

$$H(\mathbf{U}|Z=k) \le H\big(F_1^N,\dots,F_k^N\big) \le kH(F^N),$$

since **U** is a function of $Z = k$ identically distributed random variables with the same truncated geometric distribution as F^N and its (conditional) entropy is bounded from above by the entropy of $F_1^N, F_2^N, \dots, F_k^N$ when they are independent. The lemma now follows by substituting this second equation in the previous one. □

Corollary A.2.

$$0 \le H(\mathbf{U}) \le Ne^{-c}H(F^N) + \log_2 N.$$

Proof. The relation is an immediate consequence of the lemma, once we recall that $Z \le N$ and $\mathbb{E}Z = Ne^{-c}$. □

APPENDIX B

An upper bound for H(X). In his thesis Newberg (1993) obtained recurrence relations and asymptotic expressions for the total number $C(N)$ of interleavings involving any number of islands which can be formed from N equal-sized randomly located cloned fragments. His asymptotic expression is given in the following result.

Proposition B.1.

$$C(N) \sim \frac{e^{3/8}\sqrt{2}}{8N}\left(\frac{4N}{e}\right)^N \quad \textit{as } N \to \infty.$$

Corollary B.1.

$$\begin{aligned} H(X) &\le \log_2 C(N) \\ &= N\log_2 N + N\log_2\left(\frac{4}{e}\right) - \log_2 N \\ &\quad + \frac{3}{8}\log_2(e) - \frac{5}{2} + o(1) \quad \textit{as } N \to \infty. \end{aligned}$$

APPENDIX C

Proofs of Results A and B.

Proof of Result A. Note that, for $S = \mathbf{Y}$ or $\mathbf{Q}$,

$$H(S) = H(S \mid \mathbf{U}) + H(\mathbf{U}).$$

The bounds for $S = \mathbf{Y}$ follow from Corollaries A.1(ii) and A.2 and Lemma A.4. The bounds for $S = \mathbf{Q}$ follow from Corollaries A.1(i) and A.2 and Lemma A.5. The bounds for $S = \mathbf{X}$ follow from Corollary B.1 and the fact that

$\mathbf{X}$ is a function of $\mathbf{Y}$ and the lower bound on $H(\mathbf{Y})$. We dropped the constant term in the upper bound for $\mathbf{X}$ in Corollary B.1 since it makes only negligible difference. Finally, the bounds for $S = \mathbf{P}$ follow from the facts that

$$H(\mathbf{P}) \geq H(\mathbf{Q}) \geq \underline{H}(\mathbf{Q})$$

and

$$H(\mathbf{P}) \leq H(\mathbf{X}) \leq \overline{H}(\mathbf{X})$$

(because $\mathbf{Q}$ is a function of $\mathbf{P}$ and $\mathbf{P}$ is a function of $\mathbf{X}$). □

PROOF OF RESULT B. (i) and (ii) follow directly from the finite-sample bounds on $H(\mathbf{X})$, $H(\mathbf{Y})$ and $H(\mathbf{P})$, and the exact expressions for $H(\mathbf{W})$ and $H(\mathbf{V})$, and so does

$$H(\mathbf{P}) \geq (1 - e^{-c}) N \log_2 N(1 + o(1)).$$

Because $\mathbf{Q}$ is a function of $\mathbf{P}$,

$$H(\mathbf{P}) = H(\mathbf{Q}) + H(\mathbf{P} \mid \mathbf{Q}).$$

For any given configuration $\mathbf{Q}$, let $\mathbf{U} = \mathbf{U}(\mathbf{Q})$. Then, for any island of i clones, $\mathbf{P}$ can only take $2^{i(i+1)/2}$ possible values. It follows that

$$\begin{aligned} H(\mathbf{P} \mid \mathbf{Q}) &\leq \mathbb{E}\log_2\Big(\prod_i 2^{U_i \times i(i+1)/2}\Big) \\ &\leq \sum_i \mathbb{E}U_i(i^2 + i)/2 \\ &= \sum_i Np^2 q^{i-1}(i^2 + i)/2 \\ &= Ne^{-c} \sum_i pq^{i-1}(i^2 + i)/2 \\ &= Ne^{-c}\big(\mathbb{E}\{F_N^2\} + \mathbb{E}\{F_N\}\big)/2(1 - q^N) \\ &= O(N) \quad \text{as } N \to \infty. \end{aligned}$$

Hence

$$H(\mathbf{P}) \leq H(\mathbf{Q}) + O(N) = (1 - e^{-c}) N \log_2 N(1 + o(1)). \qquad \square$$

Acknowledgment. The authors would like to thank Dr. Phil Green for his very valuable comments and insights, which led to a greatly improved revision of their initial paper.

REFERENCES

AIGNER, M. (1979). *Combinatorial Theory.* Springer, New York.

ALIZADEH, F., KARP, R. M., NEWBERG, L. A. and WEISSER, D. C. (1993). Physical mapping of chromosomes: a combinatorial problem in molecular biology. In *Proceedings of the Fourth Annual ACM–SIAM Symposium on Discrete Algorithms*, Austin, TX. ACM, New York.

BRANSCOMB, E., SLEZAK, T., PAE, R., GALAS, D., CARRANO, A. V. and WATERMAN, M. (1990). Optimizing restriction fragment fingerprinting methods for ordering large genomic libraries. *Genomics* **8** 351–366.

INFORMATION AND CLONE MAPPING

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

Cox, D. R., Burmeister, M., Price, E. R. and Myers, R. M. (1990). Radiation hybrid mapping: a somatic cell genetic method for constructing high resolution maps of mammalian chromosomes. *Science* **250** 245–250.

Craig, A. G., Nizetic, D., Hoheisel, J. C., Zehetner, G. and Lehrach, H. (1990). Ordering of cosmic clones covering the Herpes simplex virus type-I (HSV-I) genome—a test case for fingerprinting by hybridization. *Nucleic Acids Res.* **218** 2653–2660.

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

Fu, Y.-X., Timberlake, W. E. and Arnold, J. (1992). On the design of genome mapping experiments using short synthetic oligonucleotides. *Biometrics* **48** 337–359.

Green, E. D. and Green, P. (1991). Sequence-tagged sites (STS) content mapping of human chromosomes: theoretical considerations and early experiences. *PCR Methods and Applications* **1** 77–90.

Lander, E. S. and Waterman, M. S. (1988). Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2** 231–239.

Lehrach, H., Drmanac, R., Hoheisel, J., Larin, Z., Lennon, G., Monaco, A. P., Nizetic, D., Zehetner, G. and Poustka, A. (1990). Hybridization fingerprinting in genome mapping and sequencing. In *Genome Analysis. Genetic and Physical Mapping* (K. E. Davies and S. M. Tilghman, eds.) **1** 39–81. Cold Spring Harbor Laboratory Press.

Nelson, D. O. and Speed, T. P. (1994). Statistical issues in constructing high resolution physical maps. *Statist. Sci.* **9** 334–354.

Newberg, L. A. (1993). Finding, evaluating and counting DNA physical maps. Ph.D. dissertation, Dept. Electrical Engineering and Computer Science, Univ. California, Berkeley.

Olson, M. V., Dutchik, J. E., Graham, M. Y., Broudeur, G. M., Helms, C., Frank, M., MacCollin, M., Sheinman, R. and Frank, T. (1986). Random-clone strategy for genomic restriction mapping in yeast. *Proc. Nat. Acad. Sci. U.S.A.* **83** 7826–7830.

Rényi, A. (1984). *A Diary on Information Theory*. Wiley, New York.

Trask, B. J. (1991). Fluorescence in situ hybridization—applications in cytogenetics and gene mapping. *Trends in Genetics* **7** 149–154.

Department of Statistics
University of California
367 Evans Hall #3860
Berkeley, California 94720-3860
E-mail: binyu@stat.berkeley.edu