Volume 138 Issue 2 June 2007 ISSN 0304-4076

JOURNAL OF Econometrics

INFORMATION AND ENTROPY ECONOMETRICS: A VOLUME IN HONOR OF ARNOLD ZELLNER

edited by
A. GOLAN AND Y. KITAMURA

Available online at www.sciencedirect.com
ScienceDirect

Journal of Econometrics 138 (2007) 405–429

JOURNAL OF Econometrics

www.elsevier.com/locate/jeconom

Information optimality and Bayesian modelling

Bertrand Clarke*

Department of Statistics, University of British Columbia, 6356 Agricultural Rd. Rm. 333, Vancouver, BC, Canada V6T 1Z2

Available online 14 June 2006

Abstract

The general approach of treating a statistical problem as one of information processing led to the Bayesian method of moments, reference priors, minimal information likelihoods, and stochastic complexity. These techniques rest on quantities that have physical interpretations from information theory. Current work includes: the role of prediction, the emergence of data dependent priors, the role of information measures in model selection, and the use of conditional mutual information to incorporate partial information.

JEL classification: Primary C-11; secondary C-19

Keywords: Entropy; Bayesian method of moments; Reference priors; Stochastic complexity; Data dependent priors

1. Overview

Conventional inference in statistics and econometrics roughly segregates into Bayes and Frequentist schools of thought. The likelihoodist school was a conceptual bridge between them. However, it had largely faded out by 1980, because, as a ‘‘minimalist’’ school, it could not provide useful interpretations for interval estimators. There was no prior, so they were not credibility sets. Notions of confidence did not apply either: the intervals, although defined in terms of the likelihood, did not have a probabilistic interpretation.

Arguably, the information theoretic approach is the likelihood school redux: information methods typically use a prior, but the prior has a physical interpretation.

*Corresponding author. Tel.: +1 604 822 4373; fax: +1 604 822 6960.
E-mail address: bertrand@stat.ubc.ca.

0304-4076/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2006.05.003

Interval estimators do not necessarily have an associated confidence, but they do inherit a confidence like interpretation from the concept of codelength or complexity. Overall, information methods fit more comfortably in a Bayesian paradigm than in a frequentist paradigm, but there are enough similarities and differences to please and annoy both sides. The text by Burnham and Anderson (1998) is a compendium of information theoretic techniques and applications from a model selection standpoint. The exposition of Bryant and Cordero-Brana (2000) is similar, but from a parametric standpoint.

The central idea is that one wants to choose the model, or the parameter indexing the model, that gives the shortest description length. A range of such values is obtained by choosing the models, or their parameters, that give a codelength within, say, ε of the minimal value. This is a variant on the maximum likelihood estimator, but the interpretation is different. Description length is commonly understood to mean the number of zeros and ones required to express a message, here a sequence of outcomes. Essentially, the optimal description length is given by the true likelihood, but that principle, formalized by the entropy, can be generalized to more elaborate settings.

The minimization is important because a small description length usually corresponds to more information; the outcomes it summarizes are more representative of the distribution. To do this summarization properly, the sequences of zeros and ones that form words must be elements of a fully specified language, or more formally, a codebook. Optimally, short codewords correspond to high probability events or frequently occurring messages and long codewords the opposite. Longer codelengths represent more information in an inferential sense because one has observed an event which is relatively less likely. In fact, codelengths characterize distributions.

This is particularly important for Bayesian models because the prior density can be regarded as a data source. The messages correspond to parameter values. The conditional density for the data given the parameter, or likelihood, represents the mechanism by which the source is transmitted. Decoding the message sent from the message received is the same as estimating a parameter.

To proceed, we list the main quantities that appear in information theoretic reasoning. The entropy of a probability distribution P with density p with respect to a dominating measure μ is

$$H(P) = \int p(x) \log \frac{1}{p(x)} \, \mathrm{d}\mu(x). \tag{1}$$

Expression (1) generalizes to give the relative entropy: let Q be another probability distribution with density q with respect to μ. We write

$$D(P\|Q) = \int p(x) \log \frac{p(x)}{q(x)} \, \mathrm{d}\mu(x). \tag{2}$$

The Shannon mutual information, SMI, between two random variables is the relative entropy between their joint distribution and the product of their marginal distributions. For X and Y with joint distribution P_{XY} and marginals P_X and P_Y with densities p_{XY}, p_X, p_Y, respectively (with respect to a bivariate dominating measure that we continue to denote by μ), the SMI is

$$I(X; Y) = D(P_{XY} \| P_X \times P_Y), \tag{3}$$

where $P_X \times P_Y$ is the joint probability formed from the product of marginals.

An extension of this treats all the distributions as conditional on another variable, say $Z = z$. The conditional Shannon mutual information, CSMI, is

$$I(X;Y|Z=z) = \int p(x,y|z)\log\frac{p(x,y|z)}{p(x|z)p(y|z)}\,\mathrm{d}\mu(x,y) \tag{4}$$

as a function of z, and on average

$$I(X;Y|Z) = \int I(X;Y|Z=z)p_Z(z)\,\mathrm{d}\mu(z). \tag{5}$$

The capacity is defined by taking a supremum over the possible marginal distributions for the first entry of the SMI. We write

$$\mathscr{C} = \sup_{p(\cdot)} I(X;Y), \tag{6}$$

where $p(\cdot)$ is a density for X and the conditional density $p(y|x)$ is fixed. One can perform this operation on a CSMI as well, but that is beyond our present scope.

A more complicated construction is required to define a quantity called the rate distortion function, RDF. The RDF is defined by taking an infimum over the conditional density for the second entry of the SMI given the first entry. The infimum is over a specific set of conditional distributions as defined by a distortion function. We fix the marginal for X and consider conditional distributions for Y which we regard as a representative of a region of X's. Thus, fix a distance d on vectors of outcomes $X = X^n = (X_1,\ldots,X_n)$ and choose a set $\hat{X}^n(j)$ for $j = 1,\ldots,M$ of vectors of outcomes of length n. The $\hat{X}^n(j)$'s will be the values of Y: given an outcome $X^n = x^n = (x_1,\ldots,x_n) = x$ we choose Y to be the $\hat{X}^n(j)$ where j achieves $\min_j d(X^n, \hat{X}^n(j))$.

We can do the discretization in many ways. The goal is to choose the smallest M—and the corresponding codewords $\hat{X}^n(j)$ for $j = 1,\ldots,M$—so that the distribution of the codewords achieve

$$R(D) = \inf_{\{p(\hat{x}|x):Ed(X,\hat{X})\leqslant D\}} I(X;\hat{X}), \tag{7}$$

where the minimization is over conditional densities $p(x|\hat{x})$ under which the average distortion is bounded by D, i.e., densities $p(x|\hat{x})$ for which $Ed(X,\hat{X})\leqslant D$. Here, $p(x)$ is fixed and the conditional varies—the opposite of the capacity. As before, we can replace the SMI with a CSMI. However, this is beyond our present scope.

In Section 2 we see that maximum entropy can be used as a 'worst case' principle for finding a parametric family. Zellner's Bayesian method of moments rests on this. Shannon's Source Coding Theorem shows that this corresponds to a code with the longest average codelengths. In Section 3, we see how inference relates to data transmission which rests on the SMI. Shannon's channel coding theorem gives the optimal rate for accumulating information. In Section 4, we accept that we cannot always transmit all the information we want. Shannon's rate distortion theorem characterizes an optimal likelihood that pre-processes the data to isolate the most important information.

In Section 5.1, we review some recent results in stochastic complexity; the results in Section 5.2 are new. In Section 6, we review some applications of the theory. In Section 7, our new considerations suggest extensions: network information theory may apply when data from several sources must be combined. This helps justify data-dependent priors (DDPs).

2. Shannon's source coding theorem

The entropy, (1), was originally intended for discrete variables. Let log be base 2 and let X be a random variable assuming one of three values, a, b, and c with probabilities conveniently chosen to be $\frac{1}{2}$, $\frac{1}{4}$, and $\frac{1}{4}$. Our task is to identify an outcome of X by the answers to a series of yes/no questions. The number of questions is the codelength and the answers yes or no, or one and zero, are the alphabet.

We might ask: is X equal a? Half the time the answer is yes and we stop. Half the time the answer is no and we go on to the second question: is X equal b? Given that we go on to question 2, half the time the answer is yes and half the time the answer is no. Overall, the expected number of questions is: $(1Q) \times \left(\frac{1}{2}\right) + (2Q's) \times \left(\frac{1}{2}\right) = \frac{3}{2}$. Unconditionally, one fourth of the time we ask the second question and get the answer yes while one fourth of the time we go on and get no. Thus, the entropy is $\left(\frac{1}{2}\right)\log 2 + \left(\frac{1}{4}\right)\log 4 + \left(\frac{1}{4}\right)\log 4 = \frac{3}{2}$. If we started with 'Is it c?' and then used 'Is it b?' we would get $\frac{7}{4} > \frac{3}{2}$.

2.1. Source coding

Formally define an (n, k, P_e) block code for a source $X_1, \ldots, X_n, \ldots$ to consist of an encoder $\phi : \mathscr{X}^n \to \{0,1\}^k$ and a decoder $\psi : \{0,1\}^k \to \mathscr{X}^n$ and error probability $P_e = P(\psi(\phi(X^n)) \neq X^n)$. The notation $\{0,1\}^k$ means all strings of zeros and ones of length k. A source is a distribution generating the data we want to encode. A block code means that we take the outcomes of the X_i's in strings, or 'blocks', of length n at a time. We do this repeatedly, implicitly assuming the sample is an integer multiple of n. The rate of such a code is defined to be $R_n = k/n$: we send k bits when we send n source symbols. We want the rate to be as low as possible, subject to ensuring we do not make too many errors. If we have a sequence of codes with block length n, the asymptotic rate of the sequence is $R = \lim_{n\to\infty}(k_n/n)$.

Shannon's first theorem, also called the source coding theorem, gives the operational significance of the entropy as the asymptotic lower bound for R.

Shannon's *first theorem*:

(A) For any $\delta > 0$ and $R > H$, $\exists$ rate R code with $P_e < \delta$.
(B) For any $R < H$ no sequence of rate R codes exists with $P_e \to 0$.

Note that blocks of outcomes of length n are represented by codewords of length k and are optimal in the limit. By contrast, Shannon codes have codewords of different lengths given by an explicit codelength function

$$\mathscr{L}(\phi(x)) = \left\lceil \log \frac{1}{P(x)} \right\rceil,$$

where ϕ denotes the code. Also, it is easy to see that any Shannon code has expected codelength within one bit of the entropy, and so is nearly optimal.

The form of the Shannon codelength invites you to use its negative as an exponent. This gives an equivalence of codes and probability mass functions: the Kraft–McMillan inequality states there is a uniquely decodable code with lengths $\mathscr{L}_1, \mathscr{L}_2, \ldots$ if and only if $\sum_{i=1}^{\infty} 2^{-\mathscr{L}_i} \leqslant 1$. Also, if a codelength function $\mathscr{L}(x)$ satisfies $\sum_x 2^{-\mathscr{L}(x)} \leqslant 1$ then $E\mathscr{L}(X) \geqslant H(X)$; Shannon codes satisfy $H(X) \leqslant E\mathscr{L}(X) \leqslant H(X) + 1$.

2.2. Maximizing entropy and the Bayesian method of moments

It is easy to see that for bounded discrete random variables the uniform has maximum entropy. Other 'maxent' results are derived by imposing constraints. For instance, the entropy of the $Normal(\mu, \sigma^2)$ density ϕ is $H = \left(\frac{1}{2}\right) \ln 2\pi e\sigma^2$ and the normal has maximum entropy for given mean and variance. More generally, the maximum of $H(p)$ over p's subject to $\sum_x g_i(x)p(x) = c_i$ for $i = 1, \ldots, m$ is achieved by

$$p^*(x) = c\mathrm{e}^{-\sum_{i=1}^{m} \lambda_i g_i(x)},$$

where the λ_i's are chosen so p^* will satisfy the constraints. This has been used in Soofi et al. (1995) and a discussion of the relationship between entropy and variance is in Soofi (1994). In effect, we impose constraints we believe are true to limit the range of possible average codelengths and then choose the worst case scenario.

Next, we review the Bayesian method of moments, BMOM, from Tobias and Zellner (1998), see also Zellner et al. (1997). The central intent of BMOM is to bypass prior specification by getting the posterior from a maxent argument. Operationally, BMOM equates the posterior expectation of a function of a parameter to its sample value and chooses the posterior to be the maximum entropy distribution subject to that constraint.

Here's the procedure: write

$$Y = X\beta + u,$$

where Y is an $n \times 1$ vector of observations assumed related to X, a given $n \times k$ matrix of rank k, and β is a $k \times 1$ vector of regression coefficients with fixed unknown values and u is an $n \times 1$ vector of realized error terms.

Let $D = (Y, X)$. There are two main assumptions. First, assume $X'E(u|D) = 0$ so that $\hat{\beta} = E(\beta|D) = (X'X)^{-1}X'Y$, i.e., the posterior mean of β is the least squares estimate. This gives $E(u|D) = \hat{u}$. Second, assume $Var(u|\sigma^2, D) = \sigma^2 X(X'X)^{-1}X'$ where σ^2 is a variance parameter. This gives $Var(\beta|\sigma^2, D) = \sigma^2(X'X)^{-1}$; further derivation gives $Var(\beta|D) = s^2(X'X)^{-1}$. Now, the posterior mean and variance of β are the usual estimates of parameters in large sample Bayes approaches.

If we maximize the entropy of the posterior density subject to the two moment constraints on β, $E(\beta|D) = (X'X)^{-1}X'Y$ and $Var(\beta|D) = s^2(X'X)^{-1}$, we get $\beta \sim Normal(\hat{\beta}, s^2(X'X)^{-1})$, because the normal is maxent under a second moment constraint. This is the same density for $(\beta|D)$ as one gets from the usual analysis. Thus, the usual analysis is optimal in a maximum entropy sense. Note that this is a *Bayesian* method of moments because the moment constraints are under the posterior distribution.

In maxent arguments, interest focuses on the choice of constraints. So, we change them by taking σ_o to be a fixed value. If we maximize the entropy of the posterior $(\beta|D)$ subject to $E(\beta|D) = (X'X)^{-1}X'Y$ and $Var(\beta|\sigma_o^2, D) = \sigma_o^2(X'X)^{-1}$ we get $\beta \sim Normal(\hat{\beta}, \sigma_o^2(X'X)^{-1})$, a special case of the last maximum entropy. Another example: set $E(\log \beta|D)$ and $E(1/\beta|D)$ equal to the values of the corresponding statistics. Maximizing $H(\beta|D)$ gives an inverted Gamma(λ_1, λ_2), with density

$$f^*(\beta) = \exp\left(-\left(\lambda_o + \lambda_1 \log \beta + \frac{\lambda_2}{\beta}\right)\right) \propto x^{-\lambda_1}\mathrm{e}^{-\lambda_2/\beta}.$$

The posterior density for $(\sigma^2|D)$, based on a diffuse prior and IID normal likelihood, is often an inverted Gamma. An interesting and unforeseen implication of this method is that, outside of certain limiting senses, the posteriors seem to correspond to the use of DDPs, a topic we take up in Section 7.

2.3. Relative entropy

For discrete random variables, the entropy is invariant under bijective transformations f of the sample space. That is, $H(f(X)) = H(X)$. However, this fails when X is continuous. This means that maxent arguments depend on the representation of the measure space. For instance, $H(aX) = H(X) + \log|a|$. Even worse, the entropy of a discretized continuous random variable does not converge to the entropy of its limit. That is, if $X_\delta = x_i \chi_{X \in S_i}$ where $x_i \in S_i$ and the S_i's are a partition of the range of X into bins of length δ, then $H(X_\delta) - \log 1/\delta$ converges to $H(X) \neq 0$ as $\delta \to 0$.

One way to get around this is to regard the entropy of a random variable as 'relative' to some other variable. Clearly, the relative entropy between two random variables is invariant under transformations of the measure space. In fact, the invariance of the discrete entropy follows from the invariance of the relative entropy if one uses the discrete uniform DU and DU_f on K values for X and $f(X)$: we have that $H(f(X)) = D(P_f \| DU) + \log K = D(P\|DU) + \log K = H(X)$. Optimizing an invariant quantity, like the mutual information, gives invariant optima as in reference analysis, see Section 3.2.

Although not a metric, the relative entropy is a distance on distributions. It is positive unless its arguments are equal; it defines a convex neighborhood base; it has a Pythagorean property see Csiszar (1975); and, locally, it behaves like squared error. As a distance, it is stronger than L^1 (which is a metric) but weaker than χ^2.

Information theoretically, it is the difference in average length between two Shannon codes. To see this, suppose P is the true distribution, taken to be discrete. Since P is unknown, we may be forced to use a code with codelengths $\mathscr{L}(x^n)$ derived from $Q \neq P$. The extra bits we end up using are the cost we pay for not knowing P. In a perfect world in which we knew P we would not have had to send them so we regard these extra bits as 'mistakes'. More formally, the redundancy, $R_n(P, \mathscr{L})$ is

$$\begin{aligned} R_n(P,\mathscr{L}) &= E_P(\mathscr{L}(X^n)) - H(P) = E_P \log \frac{1}{Q(X^n)} - E_P \log \frac{1}{P(X^n)} \\ &= D(P^n \| Q^n) = nD(P\|Q), \end{aligned}$$

the relative entropy. To find codes that get close to the entropy lower bound it is often easier to minimize the redundancy i.e., the relative entropy. Heuristically, maxent is 'min-rel-ent'.

3. Data transmission

The data transmission problem is to send a stream of data $X_1, \ldots, X_n, \ldots$ rapidly while ensuring the message received is decoded correctly. The problem is we may send 'X' but the receiver may incorrectly get 'X''. The model for this is called a channel.

3.1. Coding for channels

Suppose we have a list of messages $1, \ldots, M$ and we want to send one of them, w. First, w must be encoded to $X(w)$ which gets transmitted. The receiver gets Y. The receiver then decodes Y getting $\hat{w}$, hoping $\hat{w} = w$. The communication channel is modelled by a conditional density, $P(Y^n|X^n)$. That is, X^n is sent and Y^n is received. Clearly, for fixed X^n, the probability that the correct Y^n is received can be very high, but it is possible that other, incorrect, Y^n's get received. That is, for each X^n we imagine we get a distribution with a mode at $Y^n = Y^n(x^n)$ but tailing off away from Y^n. The probability of error is $P(\hat{W} \neq W | W\text{sent})$.

The simplest nontrivial example is the binary symmetric channel. A sender sends either 0 or 1 according to some distribution say q. If the sender sends 0, the receiver receives 0 correctly with probability $1 - p$ but receives 1, erroneously, with probability p. Likewise, if the sender sends 1, the receiver receives 0, erroneously, with probability p but receives 0 correctly with probability $1 - p$. This defines a conditional distribution $p(y|x)$ where $y = 0, 1$ is Bernoulli(p) given x, and $x = 0, 1$. Cover and Thomas (1991) show the capacity of the binary symmetric channel is $\mathscr{C} = 1 - h(p)$, where $h(p)$ is the entropy function of a Bernoulli(p). Other discrete and continuous channels also have closed form expressions.

Here we took messages one at a time. However, as with source coding, it will turn out that often we need to take ever longer blocks of messages to achieve optimal coding. So, we regard M not as the number of messages we might want to send but as the number of codewords we will use to represent n-fold outcomes X^n of X. The rate of such a code is $R = (\log M)/n$, in bits per transmission, or we can write $M = 2^{nR}$. Let λ_n be the supremal error (over $W = w$) for blocks of length n. We want a high rate of transmission with a low error probability, so we say a rate is achievable if and only if there is a sequence of $(2^{nR}, n)$ codes so that the error λ_n tends to zero as $n \to \infty$.

The SMI in (3) is the rate that would be achieved for a source $p(x)$. The source with the highest rate achieves the capacity, $\mathscr{C}$ defined in (6), of the channel to transmit information. This is formally given by Shannon's channel coding theorem, also called Shannon's second theorem. Like the source coding theorem there are two parts.

Shannon's second theorem:

(A) All rates below $\mathscr{C}$ are achievable. That is, for any $\varepsilon > 0$, and any proposed rate $R < \mathscr{C}$ there is a sequence of $(2^{nR}, n)$ codes with $\lambda_n \to 0$ as $n \to \infty$.

(B) No achievable rate is above $\mathscr{C}$. That is, any sequence of $(2^{nR}, n)$ codes with $\lambda_n \to 0$ must have rate $R \leqslant \mathscr{C}$.

3.2. Experimental design and reference analysis

The problem of data transmission across a channel is a statistical experiment if one regards the parameter value as the 'message' that some 'sender' wants to send us that our task is to decode. That is, we receive $X^n = x^n$ and try to decode the θ that it came from. This model led Lindley (1956) to propose the SMI and the CSMI as measures of the information in an experiment in a Bayesian context. He used them as design criteria for experiment selection and showed that certain classes of experiments were more informative than others.

Physically the SMI and CSMI are rates of transmission. This led Bernardo (1979, 1981) to propose reference analysis: the point is to compare subjective priors to the capacity achieving prior, especially in terms of the posteriors they generate, and to compare (optimal) actions under a subjectively chosen prior to (optimal) actions under the capacity achieving prior in game theoretic contexts. When doing this, the capacity achieving source is usually called the reference prior, leading to the reference action and reference posterior, see also Bernardo and Smith (1994).

Bernardo (1979) wrote $I(\Theta; X^n) = E_m D(w(\theta|X^n)\|w(\theta)) = H(\Theta) - H(\Theta|X^n)$. The left-hand side shows how much one expects prior beliefs w to change, on average, upon receipt of n data points. Rather than using the SMI to evaluate how much information there is in a prior, one can ask: What prior is it that changes most, on average, upon receipt of the data?

When there are no nuisance parameters, or other random quantities involved in the SMI, Bernardo identified Jeffreys prior, proportional to $\sqrt{|I(\theta)|}$ the root determinant of the Fisher information, as asymptotically optimizing density for θ. This is satisfying because Jeffreys prior emerges under a transformation invariance principle (it was on that basis that Jeffreys proposed it) and it emerges under a probability matching principle. The main downside is that Jeffreys prior is typically improper.

Establishing that Jeffreys really is the reference prior in the absence of nuisance parameters is done in a pair of articles by Clarke and Barron (1990, 1994). Berger and Bernardo (1989), Ghosh and Mukerjee (1992) gave the first treatment of the reference prior when nuisance parameters are present, the latter with a formal argument. They maximized CSMI's of the form $I(\Theta, X^n|\psi)$ treating ψ as the nuisance parameter. Sun and Berger (1998) and Clarke and Yuan (2003) gave more unifying treatments.

4. Data compression

In data compression, we permit the controlled loss of information by summarizing classes of outcomes by representatives. That is, given a representative we know which class of messages it summarizes but not which specific message led to it. This is the same idea as occurs in data aggregation. When two variables are summarized as one, for instance, the $\hat{X}$ in (7) represents the summary and the X represents the original data. Below, we will state an explicit form for the density $p(\hat{x}|x)$ of the optimal $\hat{X}$. Formally using this approach for data aggregation requires specification of the loss function d, which may be difficult in practice. However, the form of the density permits conceptualizing a variety of choices. We turn first to the information theoretic interpretation.

4.1. Distortion and rates

Let us fix a source P for X, and take the outcomes in blocks of length n, X^n. As in the definition of (7), let $\hat{X}^n(j)$ for $j = 1, \ldots, M$ be a set of representatives so that for any X^n we can assign $\phi(X^n) = \hat{X}^n(j_0)$ where j_0 achieves $\min_j d_n(X^n, \hat{X}^n(j))$, for some measure of distance $d = d_n$. Our main task is to choose the integer M, and the set $\hat{X}^n(j)$ for $j = 1, \ldots, M$ optimally, at least in the limit of large n.

For fixed D, d, and P, the RDF given in (7) is an infimum over the class of conditional densities for $\hat{X}^n$ given X^n in which the average distortion is bounded by D. If d is squared error loss then these densities have the property that $E_{X^n,\hat{X}^n}\sum_{i=1}^n (X_i - \hat{X}_i)^2 \leqslant D$. The goal of this search over conditional distributions is to find one that has minimal SMI. Regarding the data compression problem as a nonstandard channel coding problem is a merely one way to avoid transmitting the information that is least useful (in terms of d) so as to communicate the important messages better.

To see how Shannon's rate distortion theorem, also called Shannon's third theorem, expresses this formally, define the rate R to be $(\log M)/n$, the log of the number of representatives per outcome of X, taken n at a time. Thus, we set $M = 2^{nR}$. Given D, $X_1, \ldots, X_n, \ldots$ IID $P(\cdot)$, and $d(X, \hat{X})$ we have the following.

Shannon's third theorem:

(A) If $R > R(D)$ then $\forall \varepsilon > 0$ there is a code with rate R achieving average distortion less than or equal to $D + \varepsilon$.
(B) Every code with $R < R(D)$ has average distortion at least D.

When $X \sim$ *Bernoulli*(p) and the distortion, $d(X, \hat{X})$, is 0 for $X = \hat{X}$ and 1 otherwise the RDF is $R(D) \geqslant h(p) - h(D)$, see Cover and Thomas (1991). They also give the closed form expression for the RDF for the *Normal*$(0, \sigma^2)$. Outside a small number of standard tractable cases it is hard to evaluate $R(D)$ and to find the conditional density achieving the minimum. However, $R(D)$ and $p(y|x)$ can be found computationally by the Blahut–Arimoto algorithm. See Blahut (1972), Arimoto (1972), and Csiszar and Tusnady (1984).

4.2. Assuming minimal information in the likelihood

Despite the limited tractability of the RDF, Blahut (1972) gives expressions for the conditional distribution achieving the rate distortion function lower bound. He obtains

$$p^*(y|x) = \frac{m^*(y)\mathrm{e}^{-\lambda d(y,x)}}{\int m^*(x')\mathrm{e}^{-\lambda d(x',x)}\,\mathrm{d}x'},$$

where λ is a function of D and m^* is determined by

$$\int \frac{\mathrm{e}^{-\lambda d(y,x)} p(x)}{\int m^*(x')\mathrm{e}^{-\lambda d(x',x)}\,\mathrm{d}x'}\,\mathrm{d}x = 1.$$

Note that the marginal for X is $p(x) = \int p(x)p^*(y|x)\,\mathrm{d}x = m^*(y)$ and the optimal density is an exponential family that depends critically on the distortion function.

In this result, Blahut minimized a relative entropy which is similar to maximizing an entropy. This suggests $p^*(y|x)$ may have some useful noninformativity property. Here, we find a minimally informative *likelihood* by *minimizing* a relative entropy. Thus, to find a likelihood that makes relatively weak assumptions about the data we minimize $I(\theta; X^n)$ over conditional likelihoods for the data given the parameter, over the class of likelihoods used in the definition of the RDF.

In standard statistical notation we write

$$p^*(x|\theta) = \frac{m^*(x)\mathrm{e}^{-\lambda d(x,\theta)}}{\int m^*(y)\mathrm{e}^{-\lambda d(y,\theta)}\,\mathrm{d}y},$$

where $\lambda = \lambda(l)$ and $m^*(x)$ is determined by the same expression as m^* above. The minimization of $I(\Theta; X^n)$ is over a set S_l defined by

$$S_l = \left\{ p(x|\theta) | \int w(\theta)p(x|\theta)L(x,\theta)\,\mathrm{d}\theta\,\mathrm{d}x < l \right\}.$$

Clearly, S_l can be regarded as the set of conditional densities in which the Bayes risk of using X as an estimator for θ is bounded by l when the prior w is used. It is seen that l is an analog of the distortion D.

Analogous to reference priors, p^* is least informative in a sense that can be made precise, see Yuan and Clarke (1999a). So, it may be appropriate to use p^* for initial data analyses when little is known about the distribution. Yuan and Clarke (1999b) used this approach to generate several likelihoods of the form p^*, and therefore several posteriors, in an addiction research application. Because the posteriors gave similar inferences, robustness suggested the inferences are from the data rather than the prior or likelihood.

5. Stochastic complexity

Given a data set x^n and a class of distributions $\mathscr{F}$ the stochastic complexity of x^n with respect to $\mathscr{F}$ is the shortest codelength $\min_{F \in \mathscr{F}} L_F(x^n)$ for x^n where L_F is the codelength function for some fixed coding scheme. Now, $\hat{F} = \arg\min L_F(x^n)$ is an estimate for F_T, the true distribution.

5.1. Minimizing codelengths

We want to be sure that $\hat{F}$ will be a good estimate regardless of which member of $\mathscr{F}$ is true. This follows from an extension of Shannon's first theorem: Rissanen (1984) proves there is a best code defining a process p^* which asymptotically behaves like the data generating distribution. That is, in the parametric case,

$$\lim_{n\to\infty} \frac{E_\theta \log p_\theta(X^n)/p^*(X^n)}{(d/2)\log n} = 1$$

for almost all parameter values θ ($d = \dim(\theta)$) and the left-hand side cannot be made smaller no matter what p^* is used. A variety of coding schemes achieving the optimal asymptotic length

$$L(x^n) = -\log p(x^n|\hat{\theta}) + \left(\frac{d}{2}\right)\log n + \mathrm{o}(\log n),$$

where $\hat{\theta}$ is the maximum likelihood estimate, MLE, have been proposed and $L(x^n)$ is often used as the formula for stochastic complexity. Indeed, $L(x^n)$ gives $p^*(x^n) = n^{d/2}p_{\hat{\theta}}(x^n)$ times an error factor. Using this p^* in the last limit gives 1 as required. The central ideas of stochastic complexity can also be found in Barron (1985).

Another approach, due to Barron and Cover (1990) uses two stage coding. For simplicity, suppose $\mathscr{F}$ is countable and we have a source code for it. Given x^n, estimate the true density p_T by

$$\hat{p}_n = \arg\min_{q \in \mathscr{F}} \left[L(q) + \log \frac{1}{q(x^n)} \right].$$

Here, $L(q)$ is the codelength for q and the second term is the Shannon codelength for x^n under q. This can be represented as a Bayesian estimator: define the prior $w(q)$ on $\mathscr{F}$ by $L(q) = \log(1/w(q))$. Clearly, by the Kraft–McMillan inequality, $\sum_{q \in \mathscr{F}} 2^{-L(q)} = 1$ so the minimal complexity density estimator $\hat{p}_n$ is the posterior mode, $\max_q 2^{-L(q)} q(x^n)$, equivalent to the maximum likelihood estimator when $n \to \infty$. If the source code has a long codelength for the right model may be a problem when n is small, but this is analogous to a prior that puts little mass around the true value.

Barron and Cover (1990) show that $\hat{p}^n$ exists with probability 1, is consistent and is admissible (in the sense that $P_T(\hat{P}^n = P_T) \leqslant P_T(\hat{P}^* = P_T)$ never holds uniformly for any other P^*). The large sample behavior of the two stage codelength $L(q) + \log 1/q$ for various choices of $\mathscr{F}$ is given in Barron and Cover (1990). Rissanen (1984, 1996) considers stochastic complexity from a model selection standpoint. In place of two stage coding, he uses the information measure

$$I(x^n) = \min_{d,\theta} \left[\log \frac{1}{p_\theta(x^n)} + \frac{d}{2} \log n \right],$$

for a collection of parametric families $\{p_\theta\}$.

Now suppose we have relatively little data. Then, the lower order terms, implicit in the $o(\log n)$ may matter. Indeed, Clarke and Barron (1990) show that

$$D(P_\theta \| M_n) = \frac{d}{2} \log \frac{n}{2\pi e} + \log \frac{|I(\theta)|^{1/2}}{w(\theta)} + o(1)$$

forcing $\log 1/m(x^n)$ to be large on average when $|I(\theta)|$ is small. For small n, this overpenalizes the best models. To get around this, Rissanen (1996, Eq. (6)), derives the maximum likelihood, or ML, codelength,

$$L^*(x^n) = \log \frac{1}{\hat{p}(x^n|\hat{\theta})} + \frac{d}{2} \log \frac{n}{2\pi} + \log \int |I(\theta)|^{1/2} \mathrm{d}\theta + o(1), \tag{8}$$

as the stochastic complexity of x^n relative to a class of processes. The density $\hat{p}(x^n|\hat{\theta})$ is the normalization in the ML code defined by

$$\hat{p}(x^n|\hat{\theta}) = \frac{p(x^n|\hat{\theta}(x^n))}{\int p(x^n|\hat{\theta}(x^n))\, \mathrm{d}x^n}.$$

This does not usually have an easily computed codelength function and the denominator must be finite for the code to exist.

5.2. Regret and prediction

In the last subsection, we looked at average properties. Now, we adopt a pointwise criterion. Specifically, Shtarkov (1988), Rissanen (1996), Cesa-Bianchi and Lugosi (2000) and Xie and Barron (2000) discuss the predictive setting in which the distribution achieves

$$\inf_q \sup_{x \in \chi^n} \sup_\theta \log \frac{p(x|\theta)}{q(x)}, \tag{9}$$

the minimal regret under a log scoring rule, is $q_{\mathrm{opt}}(x^n) = \hat{p}(x^n|\hat{\theta}) = p(x^n|\hat{\theta}) / \int p(x^n|\hat{\theta})\, \mathrm{d}x^n$, the ML code (8). Observe that $\hat{p}(x^n|\hat{\theta})$ does not give a stochastic process because marginalizing it at stage n over x_n does not give the optimum for stage $n-1$. Our contribution is that the Bayes code defined by m is a stochastic process that is asymptotically equivalent.

5.2.1. The discrete case

Let w denote a prior density on the parameter θ, write $x = x^n$ and consider

$$\inf_{q} \sup_{x\in\chi^n} \sup_{\theta} \log \frac{w(\theta)p(x|\theta)}{q(x)}. \tag{10}$$

It is seen that (10) is not a Bayes risk; it is a predictive criterion, independent of the model used to generate the data. Analogous to (9), (10) is optimized by the ML code density, or the Shtarkov solution, $q_w(x) = w(\hat{\theta})p(x|\hat{\theta})/\int w(\hat{\theta})p(x|\hat{\theta})\,\mathrm{d}x$, where $\hat{\theta}$ is now the mode of $w(\theta)p(x|\theta)$. Clearly, the Shtarkov solution is not the Bayesian mixture $m(x) = \int w(\theta)p(x|\theta)\,\mathrm{d}\theta$. Although, $m(x)$ is not optimal for any n, it is asymptotically equivalent to the optimal Shtarkov solution.

We begin by assuming that w is a discrete prior density with respect to counting measure, with probability denoted W, and that θ assumes values, $\theta_1,\ldots,\theta_k\ldots$. This is the setting of Xie and Barron (2000). Extending their work, we want to give conditions to ensure the mixture $m(\cdot)$ is the unique stochastic process achieving

$$\inf_{q} \lim_{n\to\infty} \sup_{x\in\chi^n} \sup_{\theta} \log \frac{w(\theta)p(x|\theta)}{q(x)}. \tag{11}$$

Here, χ^n denotes the n-fold sample space and q ranges over the collection of limits of densities on χ^n. We show that if (11) is achieved by some $q^*(\cdot)$, and a partitioning statistic (defined below) exists, then $q^*(\cdot)$ and $m(\cdot)$ have the same asymptotic performance as the optimal Shtarkov solutions give.

Suppose $q^*(x)$ denotes a limit of densities, with probabilities Q_n^* on χ^n, that optimizes (11). Then, since the set over which the infimum is taken includes m, (11) is upper bounded by putting m in for q. Indeed, we see that $1/(1+\sum_{j\neq i} w(\theta_j)p(x|\theta_j)/w(\theta_i)p(x|\theta_i))<1$ so (11) is bounded from above by zero, and, for all x and all θ, $w(\theta)p(x|\theta)\leqslant q^*(x)$. A lower bound on (11) takes more work; our result is the following.

Proposition 1. *Suppose there is a partitioning statistic i.e., a statistic whose asymptotic distributions give θ concentrate on disjoint sets in χ^∞. Then, asymptotically, the mixture $m(x^n) = \sum_{i=1}^{\infty} w(\theta_i)p(x|\theta_i)$ is the unique stochastic process achieving* (11).

Proof. By hypothesis, there is a statistic that discriminates asymptotically among parameter values in the sense that there is a sequence δ_n and a real number and $\varepsilon>0$ so that for each n the sets

$$A_{i,n} = \{x^n : |\theta_i - \bar{x}| \leqslant \varepsilon\},$$

are disjoint, the union is the whole space χ^n, and $\forall i \exists N_i \forall n \geqslant N_i \ P_{\theta_i}(A_{i,n}) \geqslant 1-\delta_n$, with $P_{\theta_i}(\bigcup_{j\neq i} A_{i,n}) \leqslant \delta_n$. For convenience, we have denoted the statistic by $\bar{X}$ and taken θ, the parameter indexing the asymptotic distributions, to be the candidate population means.

Consider $i=1$ and choose $A \subset A_{1,n}$. We have the lower bound

$$Q^*(A) \geqslant w(\theta_1)P_{\theta_1}(A\cap A_{1,n}) = w(\theta_1)P_{\theta_1}(A|A_{1,n})P_{\theta_1}(A_{1,n}) \geqslant w(\theta_1)P_{\theta_1}(A|A_{1,n})(1-\delta_n). \tag{12}$$

To derive an analogous upper bound for Q^*, let $A \subset A_{1,n}$ and note $Q^*(A) = 1 - Q^*(A^c)$. Now, we lower bound $Q^*(Q^c)$: we have

$$\begin{aligned} Q^*(A^c) &= \sum_i Q^*(A^c \cap A_{i,n}) \geqslant \sum_i W(\theta_i) P_{\theta_i}(A^c \cap A_{i,n}) \\ &= W(\theta_i : i \neq 1) + W(\theta_1) P_{\theta_i}(A^c \cap A_{1,n}). \end{aligned}$$

This gives

$$Q^*(A) \leqslant 1 - W(\theta_i : i \neq 1) - W(\theta_1) P_{\theta_i}(A^c \cap A_{1,n}) = W(\theta_1)(1 - P_{\theta_i}(A^c \cap A_{1,n})).$$

By the disjointness of the union, we have

$$P_{\theta_i}(A^c \cap A_{1,n}) = 1 - P_{\theta_1}(A \cup A_1^c) = 1 - P_{\theta_1}(A) - P_{\theta_1}(A_1^c) \geqslant 1 - P_{\theta_1}(A) - \delta_n,$$

because $A_1^c = \bigcup_{j \neq i} A_{j,n}$ is disjoint. These last two inequalities give

$$\begin{aligned} Q^*(A) &\leqslant W(\theta_1)(P_{\theta_1}(A) + \delta_n) \leqslant W(\theta_1) P_{\theta_1}(A|A_{1,n}) P_{\theta_1}(A_{1,n}) + \delta_n \\ &\leqslant W(\theta_1) P_{\theta_1}(A|A_{1,n}) + \delta_n. \end{aligned} \tag{13}$$

Putting (12) and (13) together we have for any $A \subset A_{1,n}$ that

$$W(\theta_1) P_{\theta_1}(A|A_{1,n}) + \delta_n \geqslant Q^*(A) \geqslant W(\theta_1) P_{\theta_1}(A|A_{1,n}) - \delta_n.$$

Now, consider any set $A \subset \chi^n$ and write $Q^*(A) = \sum_i Q^*(A \cap A_{i,n})$, $M(A) = \sum_i M(A \cap A_{i,n})$. For any given $\varepsilon > 0$, we can choose a large but finite collection of θ's and an N so that for $n \geqslant N$ $|M(A) - Q^*(A)|$ is less than ε. Thus, Q^* and M converge in probability. □

5.2.2. *Continuous θ: a Laplace's method argument*

Here we give Xie and Barron (2000) style results for mixtures for the continuous coding context where Rissanen (1996) used the Shtarkov solution. In particular, it will be seen that the constant C_m found in Xie and Barron (2000) for discrete alphabets is zero, in the continuous case. So, suppose the θ in (10) has a continuous distribution. For convenience, we rewrite (10) as

$$\inf_q \sup_{n \in \mathbb{N}} \sup_{x \in X^n} \sup_\theta \log \frac{p(x|\theta)}{q(x)} \frac{w(\hat\theta)}{|\hat I(\tilde\theta(x))|^{1/2}} \tag{14}$$

in which q is a stochastic process, $\tilde\theta$ denotes the mode of the posterior and $\hat I$ indicates the empirical Fisher information. (Thus, $\tilde\theta = \hat\theta$ when the prior density is unity.) Thus, (10) has been modified by introducing a double supremum in place of the limit in (11) and a function of x giving an empirical form of the Jeffreys prior.

Extending the reasoning in Xie and Barron (1997) we can bound (14) as follows. We have

$$\sup_{n \in \mathbb{N}} \sup_{x \in X^n} \log \frac{p(x|\tilde\theta)}{m(x)} \frac{w(\tilde\theta)}{|\hat I(\tilde\theta)|^{1/2}} \geqslant \inf_q \sup_{n \in \mathbb{N}} \sup_{x \in X^n} \log \frac{p(x|\tilde\theta)}{q(x)} \frac{w(\tilde\theta)}{|\hat I(\tilde\theta)|^{1/2}} \tag{15}$$

$$\geqslant \sup_n \inf_q \sup_{x \in X^n} \log \frac{p(x|\tilde\theta)}{q(x)} \frac{w(\tilde\theta)}{|\hat I(\tilde\theta)|^{1/2}} \tag{16}$$

$$\geqslant \sup_n \sup_{\tilde{q}} \inf_q E_{\tilde{q}} \log \frac{p(x|\tilde{\theta})}{q(x)} \frac{w(\tilde{\theta})}{|\hat{I}(\tilde{\theta})|^{1/2}} \tag{17}$$

$$\geqslant \sup_n \sup_{\tilde{q}} E_{\tilde{q}} \log \frac{p(x|\tilde{\theta})}{\tilde{q}(x)} \frac{w(\tilde{\theta})}{|\hat{I}(\tilde{\theta})|^{1/2}} \tag{18}$$

$$\geqslant \sup_n E_m \log \frac{p(x|\tilde{\theta})}{m(x)} \frac{w(\tilde{\theta})}{|\hat{I}(\tilde{\theta})|^{1/2}} \tag{19}$$

$$\begin{aligned} = \sup_n \Bigg(& E_m \log \frac{w(\tilde{\theta})}{|\hat{I}(\tilde{\theta})|^{1/2}} \\ & + \int w(\theta) \int p(x^n|\theta) \log \frac{p(x^n|\tilde{\theta})}{p(x^n|\theta)} \, \mathrm{d}x^n \mathrm{d}\theta \\ & + \int w(\theta) \int p(x^n|\theta) \log \frac{p(x^n|\theta)}{m(x^n)} \, \mathrm{d}x^n \mathrm{d}\theta \Bigg). \end{aligned} \tag{20}$$

In this chain of inequalities, (15) follows by upper bounding the infimum by the particular choice of m. The next, (16), follows by using 'minimax $\geqslant$ maximin' on the outer operations. The lower bound (17) follows by using the fact that averages are less than suprema on the inner two operations. Expression (18) follows because the information inequality $D(\tilde{q}\|q) \geqslant 0$ can be rearranged to show that (17) is minimized when $q = \tilde{q}$. Expression (19) follows because we can lower bound a supremum by choosing any element over which the supremum is taken, in this case m. The last lower bound, (20), follows from multiplying and dividing by $p(x|\theta)$ in the last logarithm and separating the terms.

We will show that the left-hand side of (15) and expression (20) have the same asymptotic form thereby giving an expression for (14), when the $p(\cdot|\theta)$'s are IID. The lower bound is easier; we begin with it.

There are three terms in (20). Since we are lower bounding, it is enough to take a limit over n. The first term is $E_w E_{p(\cdot|\theta)} \log w(\tilde{\theta})/|\hat{I}(\tilde{\theta})|^{1/2}$. The inner expectation has limit $\log(w(\theta)/I(\theta)) + \mathrm{o}(1)$, so the limit of the term is $\int w(\theta) \log(w(\theta)/|I(\theta)|^{1/2})\, \mathrm{d}\theta + \mathrm{o}(1)$.

To deal with the second term, Taylor expand $\log w(\theta)p(x|\theta)$ at $\tilde{\theta}$. The inner integral tends to half the expectation of a χ^2 random variable for each θ, that is $d/2 = (d/2)\log e$. Under reasonable hypotheses this convergence is o(1) uniformly for θ in compact sets.

The third term is the Shannon mutual information $I(\Theta; X^n)$, between Θ with density w, and X^n with conditional density $p(x|\theta)$. Asymptotically $I(\Theta; X^n)$ converges to $(d/2)\log(n/2\pi e) - \int w(\theta)\log(w(\theta)/|I(\theta)|^{1/2}) + \mathrm{o}(1)$, see Clarke and Barron (1994).

Adding the three asymptotic forms for the three terms gives $(d/2)\log(n/2\pi) + \mathrm{o}(1)$, as $n \to \infty$ as a lower bound.

Expression (15) admits the same expression as an upper bound by a Laplace's method argument. For simplicity, we look only at the one-dimensional case. The generalization to any finite-dimensional θ is a matter of notation.

Parallel to Conditions (i) and (v) in Rissanen (1996, pp. 41–42), we restrict our attention to the following sets. Let Ω denote the parameter space and for $\eta > 0$, let

$$A_\eta = \left\{ \theta | I(\theta) < \frac{1}{\eta}, I(\theta) > \eta \right\}.$$

Next, A_η helps define the set

$$B = B_{n,\eta} = \{x^n | \tilde{\theta} \in \Omega - A_\eta^c\}$$

in the sample space on which estimates $\tilde{\theta}$ correspond to finite Fisher information. For given $\delta, \varepsilon > 0$, we also use A_η to define the set

$$C = C_{n,\eta,\delta,\varepsilon} = \{x^n | \forall \theta \in \Omega - A_\eta^c, \text{ and } |\theta - \theta'| < \delta \Rightarrow |\hat{I}(\theta) - \hat{I}(\theta')| < \varepsilon\}$$

in the sample space, on which the empirical Fisher information $\hat{I}(\cdot)$ has a small enough modulus of continuity.

When $B \cap C$ is the whole sample space and we ignore the supremum over n, expression (15) is

$$\sup_{x^n \in B \cap C} \log \frac{p(x^n|\tilde{\theta}) w(\tilde{\theta})}{m(x^n) |\hat{I}(\tilde{\theta})|^{1/2}} \leqslant - \inf_{x^n \in B \cap C} \log \int_{B(\tilde{\theta},\delta)} w(\theta) \frac{p(x^n|\theta)}{p(x^n|\tilde{\theta})} \, \mathrm{d}\theta \frac{|\hat{I}(\tilde{\theta})|^{1/2}}{w(\tilde{\theta})}$$

$$\leqslant - \inf_{x^n \in B \cap C} \log \int_{B(\tilde{\theta},\delta)} \frac{w(\theta)}{w(\tilde{\theta})} |\hat{I}(\tilde{\theta})|^{1/2} \mathrm{e}^{-(n/2)(\theta - \tilde{\theta})^2 \hat{I}(\theta^*)} \, \mathrm{d}\theta, \tag{21}$$

where θ^*, between θ and $\tilde{\theta}$, arises from the Taylor expansion of $\log p(x^n|\theta)$ at $\tilde{\theta}$ and the inequality follows from restricting the domain of integration over θ in m. The continuity of $w(\theta)$ gives

$$|\theta - \tilde{\theta}| < \delta \quad \Rightarrow \quad w(\theta) \geqslant w(\tilde{\theta})(1 - \varepsilon),$$

which we use on $B(\tilde{\theta}, \delta)$ in (21), an open ball centered at $\tilde{\theta}$ with radius δ. This implies that for δ small enough we also have

$$\hat{I}(\theta^*) \leqslant (1 + \eta) \hat{I}(\tilde{\theta})$$

for $x^n \in B \cap C$, for use in (21), because $\hat{I}$ is bounded away from zero.

Now, using $w(\theta)/w(\tilde{\theta}) \geqslant (1 - \varepsilon)$, the bounds on the empirical Fisher information, and multiplying and dividing by normalizing constants we get that (21) is bounded above by

$$- \inf_{x^n \in B \cap C} \log \frac{\tilde{w}(\tilde{\theta})(1 - \varepsilon)}{w(\tilde{\theta})} \sqrt{\frac{2\pi}{n(1 + \eta)}} \int_{B(\tilde{\theta},\delta)} |\hat{I}(\tilde{\theta})|^{1/2} \sqrt{\frac{n(1 + \eta)}{2\pi}} \mathrm{e}^{-(n/2)(1+\eta)(\theta - \tilde{\theta})^2 \hat{I}(\tilde{\theta})} \, \mathrm{d}\theta. \tag{22}$$

Laplace's method shows that normal integrals such as that in (20) are within $\mathrm{e}^{-\alpha n}$ of 1, for some $\alpha > 0$. So, for large n we can increase the domain of integration to the whole real line. This gives the new upper bound

$$\sup_{x^n \in B \cap C} \log \sqrt{\frac{n}{2\pi}} \frac{1 - \varepsilon}{\sqrt{(1 + \eta)}} + K \mathrm{e}^{-(\alpha/2)n}, \tag{23}$$

for some $K > 0$. Clearly, if we replace the supremum over $n \in \mathbb{N}$ with a supremum over $n \geqslant N$ and let $N \to \infty$ slowly while ε, δ and η go to zero, we get $\left(\frac{1}{2}\right) \log(n/2\pi) + \mathrm{o}(1)$ as the asymptotic expression for the upper bound, the same as the lower bound.

We state this, informally, as the following.

Theorem 1. *Under a collection of regularity conditions, the mixture distribution defines a stochastic process that is an asymptotic equalizer rule under logarithmic loss achieving the same limit as the Shtarkov solutions. That is: the optimization in* (15) *is asymptotically* *achieved by m and Shtarkov's solutions so that both achieve*

$$\inf_q \sup_{n\in\mathbb{N}} \sup_{x\in X^n} \log \frac{p(x|\tilde{\theta})}{q(x)} \frac{w(\tilde{\theta})}{|\hat{I}(\tilde{\theta})|^{1/2}}$$

as $n \to \infty$ *and* $(d/2)\log(n/2\pi) + \mathrm{o}(1)$ *is the limiting form.*

Here, in the continuous case, we did not show m and Shtarkov's solution converge to each other as in the discrete case. We only have that m bounds the performance of Shtarkov's optimum asymptotically and so asymptotically is a solution. Below, we verify that m and the Shtarkov solutions are close as densities.

5.2.3. *Equivalences*

Consider a general criterion of the form

$$\inf_q \sup_{x\in\chi^n} \sup_\theta \log \frac{p(x|\theta)}{q(x)} \frac{w(\theta)}{J(\theta)} \frac{b(x)}{\lambda(x)}, \tag{24}$$

where J denotes Jeffreys prior. It is seen that (24) reduces to Shtarkov's original criterion by choosing $w = J$ and $b = \lambda$. The quantity examined in the last section replaced w to cancel J, set $b = 1$ and replaced λ with an empirical Fisher information. Here we argue: (A) values of expressions like (24) are close to each other, (B) solutions to optimizations like (24) are close to each other, and (C) those solutions are also close to m.

We consider 3 versions of (24)—one with no b/λ, one with no w/J, and one that uses the empirical Fisher information as in the last subsection. To see the effect of the prior, we use a measure of peakedness

$$\phi_\delta(x) = W\{\mathscr{E}_\delta\}, \tag{25}$$

where W is a probability function on Ω, as proposed by Dawid. In (25), the set $\mathscr{E}$ is

$$\mathscr{E} = \mathscr{E}_\delta(x) = \left\{\theta : \log p(x|\theta) \geqslant \log p(x|\hat{\theta}) - \frac{\delta^2}{2}\right\} \tag{26}$$

for $\delta > 0$ and Taylor expansion gives

$$\mathscr{E}_\delta(x) \doteq \{\theta : |\theta - \hat{\theta}| \leqslant (n\hat{I}(\hat{\theta}))^{-1/2}\}. \tag{27}$$

This measures how peaked $p(x^n|\theta)$ is at $\hat{\theta}$ because as the peakedness goes up the range of θ's near $\hat{\theta}$ that give (24) a value close to its maximum narrows. That is, for fixed δ, if there is a sharp peak (high Fisher information near $\hat{\theta}$) then $\phi(x)$ will be relatively small; if the peak is spread out (lower Fisher information near $\hat{\theta}$) then the set $\mathscr{E}$ will be relatively larger making $\phi(x)$ larger.

The three versions of (24) that we consider are

$$\inf_q \lim \sup_{x\in\chi^n} \sup_\theta \log \frac{p(x|\theta)}{q(x)} \frac{w(\theta)}{J(\theta)}, \tag{28}$$

$$\inf_q \lim \sup_x \sup_\theta \log \frac{p(x|\theta)\phi_\delta(x)}{q(x)}, \tag{29}$$

and the online version that we suggest is most important

$$\inf_q \lim \sup_x \log \frac{p(x|\tilde{\theta})w(\tilde{\theta})}{\hat{I}(\tilde{\theta})^{1/2}q(x)}. \tag{30}$$

It is seen that the main difference between (28) and (29) is that in the first, the supremum over θ gives the mode of the posterior while in the second one gives the MLE. The posterior mode, $\tilde{\theta}$, and the MLE, $\hat{\theta}$, are very close: $\sqrt{n}(\hat{\theta}-\tilde{\theta}) \to 0$ as $n \to \infty$ so they are closer to each other than either is to a true value. Moreover, (29) and (30) are close: if δ is small and W has density $w(\theta)$, we have $\phi_\delta(x) \doteq 2\sqrt{\delta}w(\hat{\theta})(n\hat{I}(\hat{\theta}))^{-1/2}$. Using this approximation in (29) gives (30), when $\delta = n/4$. More formally, (A) and (B) are accomplished by the following.

Proposition 2. *The solution to* (28) *is*

$$q_a(x) = \frac{p(x|\tilde{\theta})w(\tilde{\theta})/J(\tilde{\theta})}{\int p(x|\tilde{\theta})w(\tilde{\theta})/J(\tilde{\theta})\,\mathrm{d}x},$$

the solution to (29) *is*

$$q_b(x) = \frac{p(x|\hat{\theta})\phi_\delta(x)}{\int p(x|\hat{\theta})\phi_\delta(x)\,\mathrm{d}x},$$

and the solution to (30) *is*

$$q_c(x) = \frac{p(x|\tilde{\theta})w(\tilde{\theta})/\hat{I}(\tilde{\theta})^{1/2}}{\int p(x|\tilde{\theta})w(\tilde{\theta})/\hat{I}(\tilde{\theta})^{1/2}\,\mathrm{d}x}.$$

For $\delta = n/4$, we have

$$\log \frac{p(x|\tilde{\theta})w(\tilde{\theta})}{\hat{I}(\tilde{\theta})^{1/2}q(x)} \approx \log \frac{p(x|\hat{\theta})W(\mathscr{E}_\delta(x))}{q(x)} \approx \log \frac{p(x|\tilde{\theta})w(\tilde{\theta})}{I(\tilde{\theta})^{1/2}q(x)}, \tag{31}$$

as $n \to \infty$ in P_θ-probability in the sense that the absolute value of the difference of the logarithms of any two is $\mathrm{O}_{\mathrm{P}_\theta}(1)$. Moreover, for any $i,j = a,b,c$,

$$\left|\log \frac{q_i(x)}{q_j(x)}\right| = \mathrm{O}_{\mathrm{P}_\theta}(1) \tag{32}$$

as $n \to \infty$.

Proof. If the solutions claimed exist, they are optimal because the log density ratio from any other density gives a larger value, See Xie and Barron (2000), or Shtarkov (1988). The normalizing constant in the solution to (29) is

$$\int_\chi \int_{\mathscr{E}_\delta} p_{\hat{\theta}}(x) 1_{\theta \in \mathscr{E}_\delta(x)}\,\mathrm{d}W(\theta)\,\mathrm{d}x.$$

For $\theta \in \mathscr{E}_\delta(x)$, $p(x|\hat{\theta})$ is bounded by $\mathrm{e}^{\delta^2/2}p(x|\theta)$ which gives the upper bound $\mathrm{e}^{\delta^2/2}W(\mathscr{E}_\delta)$, which is finite.

The three criteria are close by noting that the Taylor expansion to evaluate the integral defining W gives $\phi_\delta(x)$ is approximately $2\sqrt{\delta}w(\hat{\theta})(n\hat{I}(\hat{\theta}))^{1/2}$. For $\delta = n/4$, $\phi_\delta(x) = w(\hat{\theta})/\sqrt{\hat{I}(\hat{\theta})^{1/2}}$. It is seen that the empirical forms of the Fisher information converge to their limits $I(\theta)$ under P_θ with the fixed sample size stopping time. Now, Wilks' theorem ensures that the differences in the three criteria are bounded in probability when one examines the difference between the second and third expression.

To see the solutions q_i are close in the same sense, consider $\log q_a/q_c$. It is

$$\log\frac{p(x|\tilde\theta)w(\tilde\theta)/J(\tilde\theta)}{p(x|\tilde\theta)w(\tilde\theta)/\hat I(\tilde\theta)^{1/2}}+\log\frac{\int p(x|\tilde\theta)w(\tilde\theta)/\hat I(\tilde\theta)^{1/2}\,\mathrm{d}x}{\int p(x|\tilde\theta)w(\tilde\theta)/J(\tilde\theta)\,\mathrm{d}x}. \tag{33}$$

Since empirical Fisher informations converge in P_θ probability the first term tends to zero. The second term also goes to zero as the approximation improves, provided the integrals exist, see the asymptotic form in Theorem 1 in Rissanen (1996). The other cases are similar. □

Cesa-Bianchi and Lugosi (2000) give a version of (32) using a supremal difference of logarithms of distributions as a metric.

Next, we turn to (C), the final point of this section, to see that the Shtarkov solutions and m are close. Recall from Clarke and Barron (1989, Propositon 4.1) that under various regularity conditions, the mixture and $p(x|\hat\theta)$ are asymptotically close. A minor extension shows the same holds for $p(x|\tilde\theta)$. This gives that

$$\left|\log\frac{w(\tilde\theta)p(x|\tilde\theta)}{(n/2\pi)^{(d/2)}|\hat I(\tilde\theta)|^{1/2}}-\log m(x)\right|\to 0 \tag{34}$$

in P_θ probability as $n\to\infty$. Moreover, modifying the proof of Theorem 1 in Rissanen (1996, expressions (17) and (19))to use the posterior mode in place of the MLE, ensures

$$\log\int_{\tilde\theta(x)\in A_\eta} w(\tilde\theta)p(x|\tilde\theta)\,\mathrm{d}x=\frac{d}{2}\log\frac{n}{2\pi}+\log\int_{A_\eta}\sqrt{|I(\theta)|}w(\theta)\,\mathrm{d}\theta+\mathrm{o}(1), \tag{35}$$

parallel to the reasoning in (33). Clearly, (34) controls how far apart two mixtures can be, and (35) helps control how far apart two Shtarkov solutions can be.

Now, we can use (34) and (35) to control how far apart a Shtarkov solution and a mixture solution can be. Multiply and divide in the logarithm in (34) to get that the absolute value of

$$\begin{aligned}&\log\frac{w(\tilde\theta)p(x|\tilde\theta)}{\int_{\tilde\theta(x)\in A_\eta}w(\tilde\theta)p(x|\hat\theta)\,\mathrm{d}x}+\log\frac{\int_{\tilde\theta(x)\in A_\eta}w(\tilde\theta)p(x|\tilde\theta)\,\mathrm{d}x}{(n/2\pi)^{(d/2)}\int|I(\theta)|^{1/2}w(\theta)\,\mathrm{d}\theta}\\&\quad+\log\frac{\int w(\theta)|I(\theta)|^{1/2}\,\mathrm{d}\theta}{|\hat I(\tilde\theta)|^{1/2}}-\log m(x)\end{aligned} \tag{36}$$

goes to zero in P_θ probability. The first and last terms in (36) will give what we want if we control the middle two terms. By (35), the second term in (36) goes to zero. Under P_θ, the third term tends to $\log(\int w(\theta)|I(\theta)|^{1/2}\,\mathrm{d}\theta/|I(\theta)|^{1/2})$, a constant. Provided this constant has a finite integral over θ, with respect to $w(\cdot)$, we continue to get convergence in probability to a constant under m.

Thus, we have shown the following:

Proposition 3. *Under regularity conditions fromClarke and Barron (1989), and Rissanen (1996) we have that the difference between the Shtarkov and mixture solutions is* *asymptotically bounded, in the sense that, as $n \to \infty$,*

$$\left| \log \frac{w(\tilde{\theta})p(x|\tilde{\theta})}{\int_{\tilde{\theta}(x)\in A_\eta} w(\tilde{\theta})p(x|\hat{\theta})\,\mathrm{d}x} - \log m(x) \right| = O_M(1), \mathrm{O}_{\mathrm{P}_\theta}(1). \tag{37}$$

6. Using the theory

In this section, we outline how to use codelength properties for gambling, model selection, and density estimation. In applications, the stochastic complexity is called the minimum description length or MDL because, when a code is fixed, the codelengths are the number of bits required to specify the data one has using the models available. This is, it is the length of the word, in the given code, that gives a full description of the data. Various websites in the references can be consulted for more details and current references that are discussed here.

6.1. Applications

The most accessible treatment of MDL techniques in elementary problems is Bryant and Cordero-Brana (2000, Sections 3 and 4). In the multinomial context, they redo standard testing and estimation problems such as the equality of proportions and means (complete with tables for critical values) with simple numerical examples for a variety of data sets. Their approach uses the two stage coding ideas presented in Section 5.1.

Among MDL criteria, the ML codelength given in (8), asymptotically equivalent to $\log 1/m(x^n)$ and to the two stage code approach, is optimal. The central quantities in most MDL techniques are density ratios of the form $q_{n,\mathrm{opt}}(X^n)/p(X^n|\hat{\theta})$ arising from

$$\min_{q_n} \max_{\theta, X^n} \log \frac{q_n(X^n)}{p(X^n|\theta)} = \min_{q_n} \max_{X^n} \log \frac{q_n(X^n)}{p(X^n|\hat{\theta})}. \tag{38}$$

These arise in gambling (on horses or stock markets), prediction (under log-score or with AR processes), density estimation, and model selection in regression (linear and generalized).

Following Xie and Barron (2000), consider betting on n races run by m horses where the odds for horse x to win are $O_k(x_k|x_1,\ldots,x_{k-1})$ to 1 in race k. We place our bets in proportion to $q_n(x_k|x_1,\ldots,x_{k-1})$. If $X^n = (X_1,\ldots,X_n)$ are the indices of the winning horses, then at time n we would have wealth

$$S(X^n, q_n) = \prod_{k=1}^{n} q_n(X_k|X_1,\ldots,X_{k-1})O_k(x_k|x_1,\ldots,x_{k-1}) \equiv q_n(X^n)\mathrm{O}(X^n). \tag{39}$$

If the horse races were random and the probabilities for winning were $\theta = (\theta_1,\ldots,\theta_m)$ for the horses, then we would choose $q_n(i) = \theta_i$ if they were known. Regardless of the randomness of the races, a fixed betting strategy θ would lead to wealth

$$S(X^n, p^n_\theta) = \prod_{k=1}^{n} p(X_k|X_1,\ldots,X_{k-1})O_k(x_k|x_1,\ldots,x_{k-1}) \equiv p(X^n|\theta)\mathrm{O}(X^n). \tag{40}$$

The ratio of current wealth to ideal wealth is (39) over (40) which is of the form (38) which was discussed in the last section. Treating the horses as stocks and the odds as price ratios means we can get an expression for the minimax log wealth ratio in this case also.

In the time series context, Hansen and Yu (2001) consider Gaussian ARMA(p, q) models for X_t, with $\mathrm{N}(0, \sigma^2)$ errors and no unit roots. Thus, X_t is a stationary, second order Gaussian process. If we want to make predictions for time $t+1$, $t+2$ and so forth conditional on the first t outcomes then X_{t+1} is $\mathrm{N}(\hat{x}_t(\beta), \sigma^2 r_t(\beta))$ for some functions $\hat{x}_t$ and r_t which can be determined in terms of the parameters in the original process that we have grouped together into a vector β. Now, the predictive density for time $t+1$ is

$$q_t(x_{t+1}|\beta) = \frac{\mathrm{e}^{(x_{t+1}-\hat{x}_{t+1})^2/2r_t\sigma^2}}{2\pi r_t \sigma^2} \tag{41}$$

and the likelihood function is the product of these denoted $q(\beta)$. The two stage MDL takes the form of the Bayes information criterion

$$\log 1/q(\hat{\beta}) + (p+q+1)/2 \log n, \tag{42}$$

where $\hat{\beta}$ is the MLE, which can be minimized over (p, q). Note that the predictive MDL is

$$PMDL(p, q) = \sum_{t=1}^{n} \log \frac{1}{q_t(x_{t+1}|\hat{\beta}_t)}. \tag{43}$$

In principle, (43) is amenable to a minimax analysis similar to (38).

Rissanen et al. (1992) studied a histogram estimator for densities. One version, from Barron et al. (1998), sets $p(x|v, m) = mv_{i(x)}$ in which $v = (v_1, \ldots, v_m)$ is a vector of probabilities for m equally spaced bins in $[0, 1]$ and $i(x)$ is the index of the bin containing x. Using the Jeffreys noninformative prior in the ML-code density for x^n gives an L^* which we write as a model

$$\log \frac{1}{\hat{p}(x^n|m)} = \log \frac{1}{\hat{p}(x^n|\hat{v}, m)} + \frac{m-1}{2} \log \frac{n}{2\pi} + \log \frac{\pi^{m/2}}{\Gamma(m/2)} + \mathrm{o}(1), \tag{44}$$

since m defines the model class. In (6.9), $\hat{v} = \hat{v}(x^n)$ and $\hat{v}_i(x^n) = c_j(x^n)/n$ where $c_j(x^n)$ is the number of x_i's in bin j. Expression (44) can be used to select an m directly. Alternatively, the mixture of $\hat{p}(x^n|m)$'s over m

$$\hat{p}(x_{n+1}|x^n) = \sum_{m=1}^{m(n)} w_n(m) \hat{p}(y|x^n, m) \tag{45}$$

in which the weights are

$$w_n(m) = \frac{\hat{p}(x^n|m)}{\sum_{m=1}^{m(n)} \hat{p}(x^n|m)}$$

has some remarkable features when $m(n) \propto n^{1/3}$.

For variable selection in among nested models in a linear regression context, Barron et al. (1998) derive the ML codelength as in (8). Let $Y_t = X_t\beta + \varepsilon_t$ in which $(X_t = X_{1,t}, \ldots, X_{k,t})$ and ε_t is $\mathrm{N}(0, \sigma^2)$ we have models of the form

$$p(y^n|X_n, \beta, \tau) = \frac{1}{(2\pi\tau)^{n/2}} \mathrm{e}^{-(1/2\tau)(y^n - X_k\beta)'\Sigma_n(y^n - X_n\beta)},$$

where $X_k = X_{n,k}$ is the design matrix using k of the explanatory variables and $\tau = \sigma^2$. Clearly, $\beta = \beta_k$ but we drop the subscript. The ML codelength to be minimized over k is

$$\log\frac{1}{\hat{p}(y^n|X_k)} = \frac{n}{2}\log(2\pi e\hat{\tau}) + \frac{k+1}{2}\log\frac{n}{2\pi} + \log V + \log\int_{\tau_o}^{\infty}|I(\beta,\tau)|^{1/2}\,\mathrm{d}\tau - \frac{1}{12(n-k)} - \frac{k}{2n} \tag{46}$$

plus a negligible error term. In (46), τ_o is chosen so the ML code normalization factor is finite, $\hat{\tau}$ is the usual estimator of σ^2, $V = \mathrm{Vol}(\{\beta' X_k' X_k \beta \leqslant R\})$ where $R \geqslant \hat{\beta} X_k' X_k \hat{\beta}$. See Bryant and Cordero-Brana (2000) for numerical examples in this context and Hansen and Yu (2002) for an extension to generalized linear models.

7. Future directions

Here, we point out areas for further information theoretic development. One of these is DDPs.

7.1. Network information theory

The theory in Sections 1–3 generalizes to channels that involve several senders or receivers with various communication linkages and decoding approaches. This leads to CSMIs. Among a huge variety of possibilities, we only consider the multiple access channel because it is well-developed information theoretically and may have more immediate relevance to statistical problems.

A multiple access channel is one in which two or more senders send information to a common receiver. Thus, we imagine X_1 and X_2 are sent simultaneously across a channel and Y is received. The conditional distribution defining the channel is $p(y|x_1, x_2)$, and generalizes to three or more senders. Bypassing formalities, see Cover and Thomas (1991, Chapter 14.3), the capacity region for this channel is defined by CSMIs: it is the closure of the convex hull of all (R_1, R_2) satisfying $R_1 < I(X_1; Y|X_2)$, $R_2 < I(X_2; Y|X_1)$, and $R_1 + R_2 < I(X_1X_2; Y)$. The multiple access channel is only one way to combine sources of information. Alternatively, one can regard data as coming from a channel which is not fully known.

CSMIs such as define capacity regions, have been optimized in Ghosh and Mukerjee (1992), Berger and Bernardo (1989), and Sun and Berger (1998) who regarded the conditioning as a nuisance parameter. More generally, one can optimize CSMIs of the form $I(\phi, T_n|\Psi, S_n)$ where Ψ is a nuisance parameter, S_n is a nuisance statistic and we are really interested in estimating Φ with T_n. If $S_n = s_n$ or $\Psi = \psi$ is held fixed, then as in the earlier cases, one expects discrete priors that converge to continuous Jeffreys type priors. More generally, partial information reference priors, see Clarke and Yuan (2003), are usually ratios of variances. These include DDPs that arise from physically modelling the data transmission.

In the model selection setting, the optimal ML code version of MDL is only optimal under the SMI. If one changes the optimality criterion to a CSMI, either data dependent or not, alternative optimal coding schemes can emerge. Indeed, it is tempting to propose choosing model lists by the RDF, or the conditional RDF, and then selecting among them by a conditional MDL. At this time, such methods have not been investigated apart from Xie and Barron (2000, Section 9).

7.2. *Data-dependent priors*

Bayesians often eschew data dependence in their priors on the grounds that it is incoherent. However, gambling optimality differs from information theoretic optimality which may be more relevant. Moreover, in practice, Bayesians often choose their priors after seeing the data, automatically becoming incoherent. Indeed, there are contexts such as mixtures where Wasserman (2000) has established a sense in which DDPs are inferentially optimal.

The key reference for coherency ruling out DDPs is Freedman and Purves (1969). Here, our point is to argue that the implications of the Freedman–Purves Theorem are narrower than is commonly appreciated and to suggest remedies. Indeed, aside from specific mathematical properties, DDPs depend more on the data and so are less subjective.

7.2.1. *Coherency and data dependence in priors*

First, we restate the central result from Freedman and Purves (1969). Let a_λ be the set of all payoff functions for odds λ. Write the odds as $\lambda = \lambda(x, A)$ where $x \in \mathscr{X}$ and $A \subset \Theta$, assuming $\mathscr{X}$ and Θ are finite sets. Let $C(\Theta)$ be the set of all continuous real functions on $\Theta = \{\theta_1, \ldots, \theta_k\}$ for some integer k.

The Freedman–Purves Theorem is that if $\lambda(x, A)$ is not of the form of posterior odds then $a_\lambda = C(\Theta)$, i.e., the (many) payoff functions that mean the bookie will lose money that are in $C(\Theta)$ are possible, i.e., are in a_λ. That is, λ not posterior odds implies one can bankrupt the bookie. The strategy of proof is to show that $a_\lambda \neq C(\Theta)$ implies there is a unique prior π which generates λ.

This theorem must be examined to see what exactly it implies about betting strategies. First, note it is typically the case that $a_\lambda = C(\Theta)$, regardless of the form of the odds. Here's why: consider k bets for singleton sets $A_1 = \{\theta_1\}, \ldots, A_k = \{\theta_k\}$. The element of a_λ represented by an A_i is a continuous function on Θ and so takes k real values. The form of such a function is given by Eq. (4) in Freedman and Purves (1969). Let us write the first of these, for $A_1 = \{\theta_1\}$ as a vector of k entries, one for each value of θ. We have

$$v_1 = [1 - 1/(1 + \lambda(x, \theta_1))p(x|\theta_1), [-1/(1 + \lambda(x, \theta_1))]p(x|\theta_2), \ldots,\\ [-1/(1 + \lambda(x, \theta_1))]p(x|\theta_k)].$$

Suppose $v_2, \ldots, v_k$ have been defined similarly. Then, we can define new vectors such as

$$v_1' = (1 + \lambda(x, \theta_1))v_1 - (1 + \lambda(x, \theta_2))v_2$$

which has 0's in the last $k - 2$ positions and first two entries are $1 + \lambda(x, \theta_1)p(x|\theta_1)p(x|\theta_1)$ and $1 + \lambda(x, \theta_2)p(x|\theta_2)$. We can do the same for the unprimed vectors. Now, the typical case for the values of the likelihood $p(x|\theta)$ is that there are k linearly independent vectors with

$$\dim(a_\lambda) = k = \dim(C(\Theta))$$

which forces $a_\lambda = C(\Theta)$. Thus, by allowing all k bets θ_i we usually get the conclusion of the Freedman–Purves Theorem, regardless of the form of λ.

Thus, we see that the converse to the Freedman–Purves Theorem—λ is the posterior odds implies Dutch book i.e., a sequence of bets that are sure to bankrupt the bookie eventually, cannot be made—is unproved and, in principle, may fail. Oversimplifying, we have that not Bayes implies incoherent and so coherent implies Bayes, but it is not clear that Bayes implies coherent or that incoherent implies not Bayes.

Since each bookie corresponds to a set of payoff functions we can distinguish two cases of bookies. In case 1, which is typical, the set of payoff functions is the set of all functions. In this case, Dutch book exists, we get incoherence, but the Freedman–Purves Theorem does not formally say anything about whether the Bookie is Bayesian.

In case 2, which is not typical, the set of payoff functions is not the set of all functions. In this case, the Freedman–Purves Theorem applies. Its content is that the bookie is Bayesian, i.e., for any observation x and set of parameters A, he posts odds $\lambda(x, A) = W(A^c|x)/W(A|x)$ to 1 against the occurrence of A when x has occurred. Now, Dutch book cannot be made, and we get coherence. The proof is that any system of bets has expected payout zero, under the prior and this follows from the discussion in Freedman and Purves (1969, p. 1179). It is seen that the bookies against whom Dutch book cannot be made are topologically unusual in the set of all mathematically possible bookies and the domain of application of the Freedman–Purves Theorem is to a relatively narrow class of Bookies.

7.2.2. Information boundedness

The central feature of the Bayesian paradigm is that inference is made from the posterior distribution for the parameter given the data. This is possible as long as the data will ultimately determine the inferences via the likelihood, not the prior—or else the prior becomes the de facto likelihood. We propose a definition that will exploit the gap between the finite sample Freedman–Purves Theorem and a notion of asymptotic coherency in which data dependence in the prior drops out fast enough.

More formally, write $w_x(\theta)$ to indicate a generic DDP and $p(x|\theta)$ to be a generic conditional density. For IID data x^n, $p(x^n|\theta)$ concentrates at the true value θ_T, provided the data really came from $p(\cdot|\theta_T)$. What happens to $w_{x^n}(\theta)$? If it concentrates at a different value, say θ^* then the posterior will put mass at 2 points—θ^* and θ_T—a form of inconsistency if $\theta^* \neq \theta_T$. However, if $\theta^* = \theta_T$, then the prior is a likelihood and no longer represents a separate source of information.

Intuitively, the Fisher information of a DDP prior should be smaller than the Fisher information from the likelihood, and should go to zero, asymptotically. That is if $(1/n)I_{n,p}(\theta)$ is the Fisher information per data point for a data set of size n for the prior and $(1/n)I_{n,l}(\theta)$ is the Fisher information per data point for a data set of size n for the likelihood we want not just $I_{n,p}(\theta) \leqslant I_{n,}(\theta)$ but also $(1/n)I_{n,p}(\theta) \rightarrow \infty$. Consequently, we propose an alternative, weaker criterion to finite sample coherency that permits DDPs.

Definition. A DDP $w_x(\theta)$ is information bounded if and only if $\exists N$ so that for any given x of any sample size,

$$\arg \min_{x^m} \int w_x(\theta) \log \frac{w_x(\theta)}{w(\theta|x^m)} \, \mathrm{d}\theta$$

is achieved for some x^m with $m < N$.

Note that the main quantity is a relative entropy. Thus, information boundedness means that the minimal redundancy of coding with respect to $w(\theta|x^m)$ when $w_x(\theta)$ is the source achieved for bounded sample size. The content of the definition is that no matter how much data is used in the DDP it never exhibits a level of concentration higher than could be achieved with N data points. Equivalently, the posterior variance of the prior is bounded away from zero.

If the prior depends on the data through a finite-dimensional statistic that converges to a fixed value in probability, information boundedness holds. That is, if $w_{x^n}(\theta) = w_{T_n}(\theta)$ and $T_n \to \tau$ then $w_{x^n}(\theta) \to w_\tau(\theta)$. So, if $w_\tau(\theta)$ and the $w_{T_n}(\theta)$'s are well defined, the DDP is information bounded. This may imply asymptotic coherency.

Clearly, information-bounded DDPs do not change the main asymptotic results of consistency and asymptotic normality because their Fisher information goes to zero. Moreover, it is clear that Eq. (3) in Freedman and Purves (1969), representing coherent Bayesian odds, will hold in the limit of large n implying that some classes of DDPs will be asymptotically coherent. The DDPs identified in empirical Bayes techniques, Reid et al. (2002), Wasserman (2000), and Clarke and Yuan (2003) are information bounded.

Acknowledgments

This paper was written while the author was a sabbatical visitor at the Institute of Statistics and Decision Sciences, Duke University, and the Statistics and Applied Mathematical Sciences Institute.

References

Arimoto, S., 1972. An algorithm for computing the capacity of arbitrary discrete memoryless channels. Transactions on Information Theory 18, 14–20.

Barron, A.R., 1985, Logically smooth density estimation. Ph.D. Dissertation, Stanford University, Stanford, CA.

Barron, A.R., Cover, T., 1990. Minimum complexity density estimation. Transactions on Information Theory 37, 1034–1054.

Barron, A.R., Rissanen, J., Yu, B., 1998. The minimum description length principle in coding and modelling. Transactions on Information Theory 44, 2743–2760.

Berger, J.O., Bernardo, J.M., 1989. Estimating a product of means: Bayesian analysis with reference priors. Journal of the American Statistical Association 84, 200–207.

Bernardo, J.M., 1979. Reference posterior distributions for Bayesian inference. Journal of the Royal Statistical Society, Series B 41, 113–147.

Bernardo, J.M., 1981. Reference decisions, Istituto Nazionale di Alta Matematica. Symposia Mathematica 25, 85–94.

Bernardo, J.M., Smith, A.F.M., 1994. Bayesian Theory. Wiley, Chichester, UK.

Blahut, R.E., 1972. Computation of channel capacity and rate distortion functions. Transactions on Information Theory 18, 460–473.

Bryant, P.G., Cordero-Brana, O.I., 2000. Model selection using the MDL principle. American Statistician 54, 257–268.

Burnham, K.P., Anderson, D.R., 1998. Model Selection and Inference: A Practical Information-theoretic Approach. Springer, New York.

Cesa-Bianchi, N., Lugosi, G., 2000. Worst Case Bound for the Logarithmic Loss of Predictors. Kluwer Academic Publishers, Netherlands.

Clarke, B., Barron, A.R., 1989. Information theoretic asymptotics of Bayes methods. University of Illinois, Department of Statistics Technical Report # 26.

Clarke, B., Barron, A.R., 1990. Information theoretic asymptotics of Bayes methods. Transactions on Information Theory 36, 453–471.

Clarke, B., Barron, A.R., 1994. Jeffreys prior is asymptotically least favorable under entropy risk. Journal of Statistical Planning and Inference 41, 37–61.

Clarke, B., Yuan, A., 2003. Partial information reference priors: derivation and interpretations. Journal of Statistical Planning and Inference 123, 313–345.

Cover, T.M., Thomas, J.A., 1991. Elements of Information Theory. Wiley, New York.

Csiszar, I., 1975. I-divergence geometry of probability distributions and minimization problems. Annals of Probability 3, 146–158.

Csiszar, I., Tusnady, G., 1984. Information geometry and alternating minimization procedures. Statistics and Decisions: Supplement Issue 1, 205–237.

Freedman, D., Purves, R.A., 1969. Bayes method for bookies. Annals of Mathematics and Statistics 40, 1177–1186.

Ghosh, J.K., Mukerjee, R., 1992. Noninformative priors. In: Berger, J.O., Bernardo, J.M., Dawid, A.P., Smith, A.F.M. (Eds.), Bayesian Statistics, vol. 4. Oxford University Press, Oxford, pp. 195–203.

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

Hansen, M., Yu, B., 2002. Minimum description length model selection criteria for generalized linear models. In: Goldstein, D. (Ed.), Science and Statistics: Festschrift for Terry Speed. Lecture Notes Monograph Series, vol. 40. Institute of Mathematical Statistics, Hayward, pp. 145–164.

Lindley, D., 1956. On a measure of information provided by an experiment. Annals in Mathematics and Statistics 27, 986–1005.

References to the properties and applications of MDL can be found at ⟨http://www.csse.monash.edu.au/~dld/MML.html⟩ and ⟨http://www.mdl-research.org/⟩. Other websites such as ⟨http://www.bayesian.org⟩ have references for other topics covered here.

Reid, N., Mukerjee, R., Fraser, D., 2002. Some aspects of matching priors. In: Moore, M., Froda, S., Léger, C. (Eds.), Mathematical Statistics and Applications: Festschrift for C. VanEeden Lecture Notes Monograph Series, vol. 42. Institute of Mathematical Statistics, Hayward, pp. 31–44.

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

Rissanen, J., 1996. Fisher information and stochastic complexity. Transactions on Information Theory 42, 40–47.

Rissanen, J., Speed, T., Yu, B., 1992. Density estimation by stochastic complexity. Transactions on Information Theory 38, 315–323.

Shtarkov, Y., 1988. Universal sequential coding of single messages. Transactions from Problems in Information Transmission 23, 3–17.

Soofi, E.S., 1994. Capturing the intangible concept of information. Journal of the American Statistical Association 89, 1243–1254.

Soofi, E.S., Ebrahimi, N., Habibullah, M., 1995. Information distinguishability with application to analysis of failure data. Journal of the American Statistical Association 90, 657–668.

Sun, D., Berger, J.O., 1998. Reference priors with partial information. Biometrika 85, 55–71.

Tobias, J., Zellner, A., 1998. Further results on BMOM analysis of the multiple regression model. See: ⟨http://www-gsb.uchicago.edu/fac/arnold.zellner/more/CURRENT-PAPERS/paper.ps⟩.

Wasserman, L.A., 2000. Asymptotic inference for mixture models using data-dependent priors. Journal of the Royal Statistical Society Series B 61, 159–180.

Xie, Q., Barron, A.R., 1997. Minimax redundancy for the class of memoryless sources. Transactions on Information Theory 43, 646–657.

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

Yuan, A., Clarke, B., 1999a. An information criterion for likelihood selection. Transactions on Information Theory 45, 562–571.

Yuan, A., Clarke, B., 1999b. A minimally informative likelihood for decision analysis and robustness. Canadian Journal Statistics 27, 649–665.

Zellner, A., Tobias, J., Ryu, H., 1997. BMOM Analysis of parametric and semi-parametric regression models. See: ⟨http://www-gsb.uchicago.edu/fac/arnold.zellner/more/CURRENT-PAPERS/paper2.ps⟩.