Statistical Problem Classes and their Links to Information Theory

Bertrand Clarke* Jennifer Clarke† Chi Wai Yu ‡

August 6, 2012

Abstract

We begin by recalling the tripartite division of statistical problems into three classes, M-closed, M-complete, and M-open and then reviewing the key ideas of introductory Shannon theory. Focusing on the related but distinct goals of model selection and prediction, we argue that different techniques for these two goals are appropriate for the three different problem classes. For M-closed problems we give relative entropy justification that the Bayes information criterion (BIC) is appropriate for model selection and that the Bayes model average is information optimal for prediction. For M-complete problems, we discuss the principle of maximum entropy and a way to use the rate distortion function to bypass the inaccessibility of the true distribution. For prediction in the M-complete class, there is little work done on information based model averaging so we discuss the AIC and its properties and variants.

For the M-open class, we argue that essentially only predictive criteria are suitable. Thus, as an analog to model selection, we present the key ideas of prediction along a string under a codelength criterion and propose a general form of this criterion. Since little work appears to have been done on information methods for general prediction in the M-open class of problems, we mention the field of information theoretic learning in certain general function spaces.

Keywords: M-closed, M-complete, M-open, Bayesian, information theory, codelength, entropy, relative entropy, mutual information, rate distortion, model selection, prediction

1 Introduction

Arnold Zellner is noted for many contributions. Arguably, he is best noted for his work combining information theory and Bayesian statistics. More specifically, Zellner's perspective rested on what is now called Shannon theory. The core idea of Shannon theory is to regard messages as coming from a source distribution. That is, we assume messages are randomly generated and seek ways to express them compactly, transmit them efficiently, or compress them without too much loss. Many of the criteria that emanate from Shannon theory lead to techniques that provide good perfomance even in settings when the full information-theoretic model does not obviously hold.

Here we focus on the likelihood and the main point of this paper is to argue that the form of information-theoretic thinking appropriate to a statistical problem depends on the class of the problem and the perspective taken within that class. Specifically, we follow the tripartite division

*Departments of Medicine and of Epidemiology and Public Health, Center for Computational Science, University of Miami, Miami, FL, 33136 (email: bclarke2@med.miami.edu)

†Department of Epidemiology and Public Health, University of Miami, Miami, FL, 33136 (email: JClarke@med.miami.edu)

‡Department of Mathematics, Hong Kong University of Science and Technology

of statistical problems into three classes (M-closed, M-complete, and M-open), as advocated by Bernardo and Smith (1994), based on the nature and location of the true model relative to the chosen model class. Within each class we propose two perspectives: Identification and prediction. Thus, there are six basic categories and we argue that different techniques are appropriate to each of them. Indeed, we identify techniques appropriate to each of these classes but regard the paucity of techniques for M-open problems as a gap to be filled.

The notion of problem classes has also been implicitly recognized in economics for a long time, see Sawa (1978) who notes: 'It may be very likely that the true distribution is in fact too complicated to be represented by a simple mathematical function such as is given in ordinary textbooks.' Essentially, this means that identifying a *true* model, possible only in M-closed problems, is frequently impossible. How well we do otherwise will depend on a variety of factors leading us to distinguish between M-complete and M-open problems.

The M-closed problem class is optimistic: There is a true model which is not only uncoverable but can be regarded as actually on the discrete list of models we are considering. Clearly this is rarely going to be strictly true. So, in practice, it is enough to assume the list of models under consideration is rich enough that at least one of them is so close to the true model that the error due to model mis-specification is negligibly small compared with other sources of error.

By contrast, one might say the M-complete problem class is realistic. This problem class assumes there is a true model but that it is inaccessible. For instance, the true model may be so complicated that model identification given the data we actually have at hand is impossible. That is, we can know limited aspects of how a system responds but a fully detailed model for the response is impossible to identify (at least in the present). As an example, we might be able to infer the effect on GDP as a consequence of monetary policy if we knew enough about many of the other the variables defining the economic system. So, we can imagine that if we had complete macro-economic information – and enough of it – we would be able to construct a model for GDP. However, the chances of actually identifying such a model accurately given the usual circumstances of a researcher is remote. This means that we can use the concept of a true model but the role of the prior on models cannot reflect our belief that a given model is true. Thus, we compare different candidate wrong models in view of the conceptual existence of a true model. We deal with M-complete problems by assessing models we can formulate in terms of their tractability, ability to summarize data, provide reasonable fit, and make good predictions but do not get lured into believing our models are actually correct.

The M-open class is pessimistic. This problem class assumes that there is no true model. The best we can do is invoke models as actions to make predictions, not necessarily believing anything they may appear to say about the underlying mechanism. In this case we compare models (or predictors more generally) to each other and to the data predictively without reference to a true model. Thus, priors are merely weights to be adjusted, not tied to pre-experimental beliefs.

Examples from these three problem classes are easy to construct. Probably the price of an internationally traded commodity is an M-closed problem. For instance, one can write down a complete list of the factors affecting the price of oil and most of these can be measured, at least by proxy (e.g., political tensions mght be summarized by movements of military equipment). M-complete problems are, arguably, more typical. For instance, unemployment rates or interest rates summarize a great many effects many of which are not known but could in principle be known. That is, we can imagine there is a model to explain unemployment or interest even if we are unable to formulate one that actually works well in general. M-open problems commonly occur, but are

less familiar because they may involve the idea that people are not rational actors in the usual sense. For instance, modeling price stability is probably M-open: Possibly, the people affecting the price of a good have not yet made up their mind how they will react to a given price level under future circumstances. One might also argue that while the effect of monetary policy on GDP is M-complete, modeling the whole GDP as a function of all its components is M-open. En masse, people may not have decided if they intend to pay off debt, go on vacation, or invest in capital equipment and will not make such decisions on the basis of anything that can be modeled.

Within each problem class there are two perspectives. These are model identification and outcome prediction depending on whether the goal is to obtain a single mathematical representation to help understand the nature of the problem or merely to come up with a good predictor for future outcomes. In the M-closed case, model identification means model selection while prediction essentially means combining several models into a good predictor – which in the limit of large sample size converges to the true model. (It is well known that good predictors are usually the result of combining several mathematical representations; the classic reference for this is Clemen (1989), though there are more recent contributions such as Breiman (1994) for bagging and Raftery and Zheng (2003) for Bayes model averaging.) In the M-complete case, model identification as a whole does not make sense although some aspects of the true model might be inferred from a model we identify. From a predictive perspective, M-complete problems generally require model averaging bearing in mind that the models in the average do not get weighted according to their credibility, only according to their predictive success. In the M-open case where the concept of a true model is inapplicable the identification perspective means we seek a single unitary predictor while the predictive perspective means we are comfortable with a composite predictor, one comprised of subunits each making predictions that get combined to a single prediction.

It is intuitive that the M-closed class is less complex than the M-complete class which in turn is less complex than the M-open class. However, it is not clear how to represent these forms of complexity in codelength terms. Nevertheless, we expect that techniques appropriate for one complexity class of problems will work better for that class than for problems in other complexity classes. That is, we expect techniques that work well for M-closed problems not to work so well for M-complete or M-open problems. Likewise, we expect techniques that work well for M-complete problems should not work so well for M-closed or M-open problems and techniques for M-open problems should not work so well for M-closed or M-complete problems. We expect this to hold for both the model identification and prediction perspectives.

We comment that an important recent development in the information basis of prediction is Ebrahimi et al. (2010). These authors work with M-closed and M-complete problems from a model selection standpoint; see Theorems 1-4 for evaluations of mutual information between future outcomes and available random variables. Example 3 and Figs. 4 and 7 are particularly incisive for quantifying the information effect of Type II censoring and dependence. This contribution should be recalled in Sec. 3.1 and 4.1.

The structure of the rest of this paper is as follows. In Sec. 2, we review the basics of Shannon theory and their relationship with frequently occuring statistical ideas. Then, in Sec. 3 we turn to the M-closed problem class. From a model selection perspective, we provide an information-theoretic interpretation of the Bayes information criterion (BIC). From the model averaging perspective, we provide an information theoretic optimality for the Bayes model average [1](BMA). In Sec. 4 we focus on the M-complete problem class and show two ways one can propose parametric

[1]It is more conventional to call this the Bayesian model average, but dropping the 'sian' is shorter.

families to address M-complete problems. We also discuss the Akaike information criterion (AIC), and its variants, as good ways to predict in M-complete problems. In Sec. 5 on the M-open problem class, we review the extensive literature on 'prediction along a string' or which rests primarily on codelength. Then, from the identification perspective, we propose a reformulation of the usual problem to reflect the M-open class explicitly. From the predictive perspective, we briefly refer to current work using information theoretic concepts in reproducing kernel Hilbert spaces. In a final concluding section we review the implications of the overall view we have elaborated. Technical details for the results in Sec. 3 are relegated to two Appendices.

2 Shannon Theory

The Shannon theory view is that knowledge is the number of bits (zeros and ones) required to describe an event for which there is a probability. Loosely, a probability is a codebook and the probability of an event is a codeword. Obviously, if there are different problem classes, the properties of their probabilistic descriptions will be different and we might reasonably expect this to affect the way we represent them in terms of codebooks. Consequently, we will see that the three problem classes have different relationships with Shannon theory.

There are three fundamental quantities that occur in Shannon theory: The entropy $H(X)$ of a random variable X, the relative entropy $D(P\|Q)$ between two probabilities P and Q and the Shannon mutual information (SMI) $I(X;Y)$ between two random variables X and Y. We discuss their relationship to statistics in Subsec. 2.1 and then turn to formalities in Subsec. 2.2.

2.1 Statistical Concepts and Shannon Theory

Roughly, $H(X)$ is the average minimal codelength required to record the information in a random variable X. Maximum entropy as a principle therefore asks for the random variable with the longest minimal codelengths. When the minimal codelengths are long, the distribution tends to be spread out, and the entropy is high. When these codelengths are short, the distribution tends to be concentrated, and the entropy is low. One can therefore argue that the entropy is a better assessment of variability than the variance in some settings.

The entropy can also be maximized in some cases, usually subject to some constraints on X. Sometimes this gives convenient parametric families (see (3) below). The meaning is that we have found distributions that are as spread out i.e., uninformative, as possible given the constraints. If the 'maxent' distribution fits well then the remaining variability can be regarded as intrinsic to the data generator. If the maxent distribution does not fit well then we know that the constraints do not accurately describe the data generator. In either case, maxent distributions are often a good way to express assumptions about the data generator in terms of the likelihood.

The relative entropy $D(P\|Q)$ can be regarded as a measure of statistical distance between P and Q. It is not symmetric and does not satisfy the triangle inequality, however D does define a convex neighborhood base and therefore defines a mode of convergence. This mode is stronger than Hellinger (which is stronger than L^1), and weaker than χ^2, see Tsybakov (2009) p. 90. The relative entropy also satisfies a Pythagorean-like identity. So, D can be used as a loss function and for robustness purposes. Recently, Sancetta (2012) provided a thorough exposition of how predictive error under a variety of loss functions can be reduced to predictive error using relative entropy loss. Also, Soofi et al. (1995) proposes the information distinguishability $ID(P,Q) = 1 - e^{-D(P\|Q)}$ as

the 'right' calibration for for deciding when the difference between two models is too small to worry about.

A special form of the relative entropy is the SMI, $I(X;Y) = D(P_{X,Y} \| P_X \times P_Y)$. The SMI is the distance between a joint distribution and its product of marginals and so is a measure of dependence. When X is taken as an n-fold sample and Y is taken as a parameter θ, one can optimize to find the prior $w(\theta)$ that makes the posterior $w(\theta|x^n)$ as far from $w(\theta)$ as possible. This is the foundational concept behind reference priors – the prior that is least informative is the one that will be changed most on average upon receipt of the data. The literature on reference priors is vast and they have proved an extra-ordinarily useful principled way to choose priors.

It is well-known that information concepts characterize large deviation principles, Dembo and Zeitouni (1993) (see Chap. 6), error exponents in hypothesis tests see Cover and Thomas (1991) (Chap. 8), and recur frequently in modern model selection e.g., Shtarkov (1988), Barron and Cover (1991), Rissanen (1996) as well as in the classic Akaike information criterion (AIC) and Bayesian information criterion (BIC) which will be discussed below.

Taken together, it is seen that Shannon theory helps quantify many common statistical concepts. Consequently, it is worth providing a more formal review of the basics of Shannon theory, beginning with a simple example.

2.2 Source Coding Example

By regarding messages as probabilistically generated, Shannon theory ends up studying their codelengths. This is seen in the seminal work of Shannon (1948a), Shannon (1948b), as well as in Kullback and Leibler (1951) and Kullback (1959) (the earliest textbook). The classic texts include Ash (1965), Cover and Thomas (1991) but there are many others.

So, say we have a d letters in an alphabet generated from a source distribution and we want to represent them using only zeros and ones. Then, it is easy to see that we can use strings of zeros and ones with maximal length $\log_2 d$. Consider the 26 lower case letters of the English alphabet with six extra symbols, say ; , : , . , (,), and , (comma) giving $d = 32$. If we wanted to express these 32 'letters' in strings of 0's and 1's we can use strings of length $\log_2 32 = 4$. That is, the codelength of each symbol is four. If we receive a string of 0's and 1's of length $4n$ we can decode by recognizing each 4-tuple of digits as a letter in the augmented alphabet.

Now, one might suspect that assigning codewords of the same length to e – the most commonly occurring letter – and to : – which occurs rarely – is suboptimal. So, the length four coding for the augmented alphabet is likely suboptimal. Since brevity is accomplished by assigning shorter codelengths to more commonly occuring letters we could try assigning e the single digit 0. For ease of decoding we would want to start all other codewords for the other letters with 1. That way, once we found a 0 we would know we had an e and could move on to recognize the next letter.

To see how this form of expected codelength reasoning works suppose our alphabet consists of four 'letters' $\{Y, N, M, D\}$ for Yes, No, Maybe, and Don't Know and that the probabilities of being told one of the four letters in the alphabet are 1/2, 1/4, 1/8, 1/8. Then, we might seek codewords in $\{0,1\} \cup \{0,1\} \times \{0,1\} \cup \{0,1\} \times \{0,1\} \times \{0,1\}$ to express the letters compactly. So, write $w(Y)$ to mean the codeword – a finite string of zeroes and ones – assigned to the letter Y and write $\ell(w(Y))$ to mean its length. Define $w(N)$, $w(M)$ and $w(D)$ and $\ell(w(N)$, $\ell(w(M))$, and $\ell(w(D))$ similarly. One natural approach is to seek the coding that minimizes the expected codelength subject to the constraint that no codeword is the first part of any other codeword. These are called prefix codes and this means that once a codeword is recognized it can be decoded correctly. It can be checked

that one expected codelength minimizing prefix code assigns 0 to Y, 10 to N, 110 to M and 111 to D. In this case, if X is the 'source' i.e., represents the random selection of a letter from the alphabet, $E\ell(w(X)) = 1 * 1/2 + 2 * 1/4 + 3 * 1/8 + 3 * 1/8) = 7/4$. Note the codewords are not unique (110 and 111 can be interchanged) but the Kraft inequality ensures the code*lengths* are.

Unique codelengths are what allow us to go back and forth between the codeword and the probability the codeword occurs. Loosely, if x is a codeword, then $\log 1/P(X = x)$ is its codelength, a measure of the information it contains. Conversely any two distinct codewords x and x' with the same codelengths represent the same amount of information (though the actual information in x and x' is different). Thus, the main results of Shannon theory focus on how codelengths capture the concept of information.

2.3 Review of Shannon Theory

For discrete X, Shannon's First Theorem identifies the entropy of the source distribution as the minimal asymptotically achievable expected per-letter codelength for prefix codes. The entropy is

$$H(X) = \begin{cases} \sum_{j=1}^{J} p_j \log(1/p_j) & X \text{ discrete,} \\ \int p(x) \log(1/p(x)) dx & X \text{ continuous.} \end{cases} \tag{1}$$

In the discrete case, J is the number of codewords ($p_j = P(X = j)$) and in the continuous case dx really means p is a density with respect to a dominating measure μ suppressed in the notation.

The appearance of the logarithm, here assumed to have base 2, is essential to the codelength interpretation when X is discrete. Roughly, the logarithm of a probability of an event corresponds to the number of zeros and ones required to express the event as if it were a letter to be sent. This was seen in the case of coding the augmented alphabet where we noted strings of zeros and ones of length four were enough to code 32 letters. In fact, a Shannon code is any code for a discrete X with lengths given by $\ell(X) = \lceil \log 1/p(X) \rceil$. It can be proved that the Shannon codes have expected codelength within one bit of the optimal codelengths.

To see how (1) can be applied for discrete X, oberve that the entropy of X is $1/2 \log 2 + 1/4 \log 4 + 1/8 \log 8 + 1/8 \log 8 = 1/2 + 1/2 + 3/8 + 3/8 = 7/4$ which is the same as the expected codelength already found above. Indeed, equality holds whenever the probabilities of letters are of the form 2^{-k}; more generally the entropy is a lower bound. Shannon's First Theorem gives that the entropy lower bound is tight and can be achieved by an appropriate (Huffman) coding scheme.

The situation is quite different when X is continuous because the entropies of discretizations of a continuous random variable X do not converge to the entropy of X as the discretization gets finer. In fact, if X^δ is a discretized version of X for intervals of length δ, then

$$H(X^\delta) + \log \delta \to H(X). \tag{2}$$

For $\delta = 1/n$, $H(X^{1/n}) \approx H(X) - \log n$, i.e., the discrepancy grows as the discretization gets finer. This happens because it takes infinitely many bits to describe an arbitrary real number exactly.

Despite (2), the continuous entropy is an asessment of variability that can be invoked in Zellner's Bayesian method of moments and other maximum entropy procedures. Indeed, as noted in Subsec. 2.1 the maximum entropy (ME) principle can be used to give families of distributions effectively formalizing a sort of method of moments approach. Suppose we have a collection of constraints of the form $E_p T_k(X) = \lambda_k$ where the subscript p indicates the unknown model and T_k is a statistic

such as a moment, $k = 0, \ldots, K$. Then, writing $\lambda_0^K = (\lambda_0, \ldots, \lambda_K)$ it is not hard to prove, see Cover and Thomas (1991), that

$$p^*(x|\lambda_0^K) = \arg \max_{\{p: E_p T_k(X) = \lambda_k, k=0,\ldots,K\}} H(X) = C(\lambda_0^K) e^{\eta_0 + \sum_{k=1}^K \eta_k T_k(x)}, \tag{3}$$

where $\eta_k = \eta_k(\lambda_k)$ and C is a normalizing constant. Often (3) is credited to Kullback (1954). The expression (3) generalizes to the case that $H(X)$ is replaced by the relative entropy in which case it is often called the minimum discrimination information (MDI).

On the other hand, we may get around (2) by using the relative entropy. Consider two random variables X and Y with distributions P and Q with respect to the same dominating measure (which we ignore). The relative entropy between them is

$$D(P||Q) = \int p(x) \log \frac{p(x)}{q(x)} dx, \tag{4}$$

provided P is absolutely continuous with respect to Q. The relative entropy is a 'redundancy' because it represents the extra bits we would have to send (on average) if we used a suboptimal prefix code. If we have the (nearly optimal) Shannon codelengths $\ell(X) = \lceil \log 1/P(X) \rceil$ for a discretized form of X and we have another set of codelengths $\ell'(X)$ then we can write $P(X = x) \approx e^{-\ell(x)}$ and $Q(X) \approx e^{-\ell'(x)}$. This gives the expected excess bits as

$$E_P(\ell'(X) - \ell(X)) \approx D(P||Q) \geq 0. \tag{5}$$

So asymptotically minimizing the redundancy (specifically in 'block coding', see Cover and Thomas (1991)) effectively means choosing Q to be the Shannon codelengths from P.

Another form of the redundancy is

$$BR(w, Q) = \int w(\theta) D(P_\theta||Q) d\theta, \tag{6}$$

for a parametric family P_θ equipped with a prior w on θ. Expression (6) can be recognized as the Bayes risk of the action Q where D is the loss function. So, minimizing over Q gives the Bayes action for the decision theory problem of estimating P_θ. Parallel to (5) which identified the Shannon code with based on P as having minimal redundancy, Aitchison's Theorem, Aitchison (1975), identifies the Shannon code based the mixture $M(\cdot)$ of probabilities $P_\theta(\cdot)$ with respect to $w(\cdot)$ as having minimal $BR(w, Q)$ over Q in (6).

Formally, the Shannon mutual information (SMI) is

$$I(X; Y) = \int p(x, y) \log \frac{p(x, y)}{p(x)q(y)} dx. \tag{7}$$

Again, $P_{X,Y}$ must exist with respect to the same dominating measure as $P_X \times Q_Y$ and $P_{X,Y}$ must be absolutely continuous with respect to $P_X \times Q_Y$. Of particular importance is that without this absolute continuity one cannot use a Jensen's inequality argument on log to show non-negativity of the SMI. Indeed, Ebrahimi et al. (2008) p. 1229 notes the non-applicability of the SMI to singular distributions and that in the case of the Marshall-Olkin distribution one can get negative SMI's. On the other hand, both (4) and (7) behave well with discretization in the sense that, for instance, $I(X^\delta, Y^\delta) \to I(X; Y)$ as $\delta \to 0^+$.

Note that, mathematically, the SMI between $\Theta \sim w(\cdot)$ and $X \sim P_\theta$ is

$$I(\Theta; X) = BR(w, M_n) = E_M D(w(\cdot|x) \| w(\cdot)).$$

So, (7) can be interpreted as (i) a redundancy in source coding, (ii) as the minimal value of (6) and, (iii) as the distance between a posterior and a prior. Of great curiosity, the SMI has a fourth interpretation as a rate of transmission across an information theoretic-channel.

An information-theoretic channel is a conditional density $p(y|x)$ we intend to use repeatedly that gives the distribution of the y's received given the x's that were sent. Obviously we would like $p(y|x)$ to concentrate on the line $y = x$ but errors in transmission of x, i.e., noisy channels, usually make $p(\cdot|x)$ spread out around x. Shannon's Second Theorem identifies the capacity i.e., the maximal achievable average rate of transmission (in bits per usage), of the channel as

$$\mathcal{C} = \sup_{p(x)} I(X; Y). \tag{8}$$

The supremum is taken over possible source distributions because we are looking for a property of the channel; the optimization is also used to define reference priors giving them the interpretation as the (source) distribution for a parameter that permits fastest receipt of information from the likelihood (regarded as a channel).

Expression (8) can be interpreted as saying there is an agent $p(\cdot)$ transmitting many independent copies X, as fast as possible in bits per transmission and that for each X the receiver receives a Y. Statistically, the message X is θ, the parameter value, and the channel is $p(x^n|\theta)$ where x_i is the version of θ received by receiver i, $i = 1, \ldots, n$ so that $Y = (X_1, \ldots, X_n) = X^n$. Usually, one assumes the received X_i's are independent and identically distributed (IID) given θ. Because the density is a product of the $p(x_i|\theta)$'s the n receivers pool their data $x^n = (x_1, \ldots, x_n)$ and use it to decode which θ was sent. If an agent were not transmitting at capacity, the rate of information receipt by the n receivers would be lower than $\mathcal{C}$. Using a reference prior implicitly assumes there is an 'agent' transmitting the data optimally using the likelihood as a channel.

The mutual information also arises in data compression. Data compression means that we are willing to sacrifice a bit of accuracy for the sake of big enough gains in some other sense, such as speed of transmission. Imagine a variable X that we want to compress. The way we do this in principle is to choose representatives $\hat{x}$ of X, effectively converting X to a random variable $\hat{X}$ taking finitely many values that we want to determine optimally. Now, $I(X; \hat{X})$ is the average rate of transmission if one imagines the channel $p(\hat{x}|x)$ in which X is sent but $\hat{X}$ is received. So, we want the channel that is slowest in the sense of fewest bits being received by a receiver on averagae per usage subject to a constraint that ensures some minimal number of bits per usage is being received on average. The condition that ensures some bits are transmitted at each usage of $P(\hat{x}|x)$ is $E(d(X, \hat{X})) \leq D$, where $d(\cdot, \cdot)$ is the distance or distortion between x and $\hat{x}$ bounded by $D \in \mathbb{R}$. Specifically, we limit attention to conditional densities i.e., channels, in the set

$$\mathcal{P}_D = \{p(\hat{x}|x) : \int p(x)p(\hat{x}|x)d(x, \hat{x}) \leq D\}. \tag{9}$$

These conditional densities have Bayes risk of $\hat{X}$ as an estimator for X bounded by D under the loss d. The rate distortion function (RDF) is

$$R(D) = \inf_{p(\hat{x}|x) \in \mathcal{P}_D} I(X, \hat{X}). \tag{10}$$

In effect, (10) optimizes over likelihoods, unlike decision theory where one optimizes over estimators. Shannon's Third Theorem is that the RDF – the smallest achievable transmission rate given D – is the same as the distortion rate function – the smallest achievable distortion for a given rate of transmission. The distribution acheiving the RDF lower bound does not have a specific name but has been studied Blahut (1972b) and can be approximated by the Blahut-Arimoto algorithm as explained in Sec. 4.1.

3 M-closed Class

We begin with the assumption that we have a model list $\mathcal{P} = \{p_k(\cdot|\theta_k) : k = 1, \ldots, K\}$ where $\theta_k \in \Omega_k \subset \mathbb{R}^{d_k}$. With some loss of generality, the Ω_k are assumed compact and that convergence of a sequence of parameter values in Euclidean norm is equivalent to the weak convergence of the distributions they index. Essentially, for each k, this makes the mapping $\theta_k \to P_{\theta_k}$ continuous so that $\mathcal{P}_k = \{p_k(\cdot|\theta_k) : \theta_k \in \Omega_k\}$ is the contnuous image of a compact set and hence compact itself. We assume the prior probabilities W_k on Ω_k have densities w_k with respect to Lebesgue measure and that the across-model prior probability W has a density $w(\cdot)$ with respect to counting measure.

While the proof of our main theorem in this section is greatly simplified if the $p_k(\cdot|\theta_k)$'s are disjoint, we prefer to state a slightly more general result. Let $\theta_{k_0} \in \Omega_{k_0}$ and $\theta_k \in \Omega_k$ and write

$$T_n(\theta_{k_0}, k) = \int_{\Omega_k} w(\theta_k) e^{-nD(P_{\theta_0} \| P_{\theta_k})} d\theta_k. \tag{11}$$

Given our compactness assumptions, if the $\mathcal{P}_k$ are mutually disjoint, $T(\theta_0, k)(n) = \mathcal{O}(e^{-\gamma n})$ for some $\gamma > 0$. However, (11) can be given asymptotically more generally. For instance, if the d_k are increasing, as would be the case if the parametric families were nested, and we have $d_{k_0} < d_k$ then $T(\theta_{k_0}, k)(n) = \mathcal{O}(n^{d_k/2})$. This follows by Laplace's method arguments, see de Bruijn (1958) (Chapter 4), Walker (1969), Clarke and Barron (1988); sufficient regularity conditions can be found in Clarke and Barron (1988) and are listed in Appendix A.

The within-model and across models marginals for the data are

$$m_k(x^n) = m(x^n|k) = \int w_k(\theta_k) p_k(x^n|\theta_k) d\theta_k \quad \text{and} \quad m(x^n) = \sum_{k=1}^{K} w(k) m(x^n|k), \tag{12}$$

where we use $x^n = (x_1, \ldots, x_n)$ to be the outcome of a sequence of random variables $X^n = (X_1, \ldots, X_n)$ assumed IID unless noted otherwise. We use subscripts and conditioning interchangably, e.g., $m_k(x^n) = m(x^n|k)$ and drop subscripts when no misunderstanding will result.

We begin by giving a codelength interpretation for the Bayes information Criterion (BIC) and then justify the BMA from a Shannon theory optimization.

3.1 Model Identification

The BIC arises from seeking the model that has the highest posterior probability, i.e., the mode of the posterior over models. This can be seen as taking the action that is optimal under a zero-one loss function, see Kass and Raftery (1995). A variant on this comes from examining Bayes factors, see Schwarz (1978), which are also optimal under zero-one loss. Whether one takes a decision

theoretic stance or a hypothesis testing stance, one is led to maximize

$$\log W(k|x^n) = \log m(x^n) + \log w(k) + \log \int p(x^n|\theta_k) w_k(\theta_k) d\theta_k,$$

in which the first two terms can be ignored. The first does not depend on k and in the second $w(k)$ is bounded above and away from zero so that $\log w(k)$ is negligible compared to the third term. If we use Laplace's method (or Lemma 3 in Appendix A) on the last term we get

$$\begin{aligned} \int p(x^n|\theta_k) w_k(\theta_k) d\theta_k &\approx p(x^n|\hat{\theta}_k) \int e^{-(n/2)(\theta_k - \hat{\theta}_k)^T I_k(\hat{\theta}_k)(\theta_k - \hat{\theta}_k)} w_k(\theta_k) d\theta_k \\ &\approx p(x^n|\hat{\theta}_k) \left(\frac{2\pi}{n}\right)^{d_k/2} \left(\det I_k(\hat{\theta}_k\right)^{-1/2}, \end{aligned} \tag{13}$$

where $\hat{\theta}_k$ is the MLE under model k and $I_k(\cdot)$ is the Fisher information of model k. Using (13) in $-2 \log W(k|x^n)$ and simplifying leads to

$$BIC(k) = BIC(k, x^n) = \log p(x^n|\hat{\theta}_k) - \frac{d_k}{2} \log n$$

as the BIC value for model p_k. Choosing the model with largest $BIC(k)$ is consistent under mild conditions, but see Berger et al. (2003) for an even better approximation.

Here, we provide a justification for the BIC based on Shannon theory. This is more reasonable since the zero-one loss treats two wrong models as equally wrong even when one is much better than the other. So, recall that Aitchison's theorem shows that the Bayes code i.e., the code based on $m(\cdot)$, is optimal because the objective function is the redundancy (see 4, 5, and 6) integrated over all parameters and hence called the Bayes redundancy; this terminology is not standard but is frequently used, see Yu (1996), Catoni (2012). To extend Aitchison's Theorem to the multi-model case, we begin with a proposition comparing mixture densities from different models.

Proposition 1: Fix $\theta_0 \in \Omega_{k_0}$ as the true value of the parameter and suppose we are comparing model k to model k_0, where $d_k > d_{k_0}$. Then, if $T_n(\theta_0, k) = o(n^{-d_k/2})$, we get

$$\frac{m_k(x^n)}{m_{k_0}(x^n)} = \mathcal{O}_P(n^{-(d_k - d_{k_0})}) \to 0,$$

in P_{k_0} probability, under the regularity conditions A1–A7 in Appendix A.

The proof is given in Appendix B.

Unfortuntely, the proposition as stated only handles the case that the biggest model is true. Writing $P_{\theta_k}^n$ for the n-fold product of P_{θ_k}s, Aitchison's Theorem can be stated as

$$\arg \inf_Q \sum_{k=1}^K \int w(\theta, k) D(P_{\theta_k}^n || Q_n) d\theta_k = m(\cdot).$$

Note that here, $m(\cdot)$ has an argument of the form x^n. Now suppose we have K models in increasing size and the K-th, the largest, is true. Then, we have

$$\begin{aligned} \log \frac{1}{m_K(x^n)} &= \log \frac{1}{w(K) m_K(x^n) + \sum_{k=1}^{K-1} w(k) m_k(x^n)} \\ &= \log \frac{1}{m_K(x^n)} + \log \frac{1}{w(K) + \sum_{k=1}^{K-1} w(k) m_k(x^n)/m_{k_0}(x^n)}. \end{aligned}$$

A Laplace's method argument gives that when model K is true and the true parameter value is θ_K

$$\begin{aligned}\log\frac{1}{m_K(x^n)} &= -\log p(x^n|\hat{\theta}_K) + \frac{d_K}{2}\log n - \frac{1}{2}\log(2\pi)^{d_K}|I(\hat{\theta}_k)| + \log\frac{1}{w_K(\hat{\theta}_K)} + o_P(1)\\ &= -BIC(K) + \hat{C} + o_P(1)\end{aligned}$$

in P_{θ_0}-probability as $n \to \infty$ and $\hat{C}$ converges to a constant. So, if we choose a uniform partition of the space of x^n-values with side length Δ then as $D \to 0$ the partition becomes finer and finer. Write the discretized form of $m_K(x^n)$ and $m_{K,\Delta}(x^n)$. Then, the Shannon codelengths satisfy

$$\log\lceil\frac{1}{m_{K,\Delta}(x^n)}\rceil \approx \log\lceil\frac{1}{m_K(x^n)}\rceil$$

which can be approximated as in Lemma 3 in Appendix A or Clarke and Barron (1988). Thus, when the largest model K is true and $\theta_K \in \Omega_K$ is the true value of the parameter, maximizing the BIC is equivalent to choosing the model that assigns the shortest Bayes codelengths to the data for a fine enough discretization. Since the largest model is rarely true, we have the following.

Theorem 1: Assume the regularity conditions A1–A7 in Appendix A. Also, assume there is a K' so that if $k_0 < K'$

$$T_n(\theta_0, k) = \mathcal{O}(n^{-(K'-d_{k_0})/2}), \tag{14}$$

and that for $k_0 \geq K'$ is a $\gamma > 0$ so that for $k \geq K'$

$$T_n(\theta_0, k) = \mathcal{O}(e^{-\gamma n}). \tag{15}$$

Then, using the BIC for model selection in a finite M-closed problem is equivalent to choosing the model in $\mathcal{P}$ assigning the smallest Shannon codelength to the string formed by the data under a suitably fine discretization in P_{k_0}-probability, in the limit of large n.

The proof is in Appendix B.

Satisfying the hypothesis (15) can be difficult because if there is any overlap between the densities indexed by Ω_{k_1} and Ω_{k_2} for some $k_1 < k_2$ a key product in the proof (see 32) does not go to zero in probabilitiy. In nested cases, one way to satisfy (15) is by careful prior selection so that W_{k_1} assigns very little of its mass to Ω_{k_1}. Even though this might require large n to get useful results, one can let $\theta_0 \in \Omega_{k_2}$, fix $\xi > 0$ and write

$$N_\xi(\theta_0, k_1) = \{\theta_{k_1} \in \Omega_{k_1} : D(P_{\theta_0}\|P_{\theta_{k_1}}) > \xi\}.$$

Now, assume we have a sequence of priors $W_{k_1,n}(\cdot)$ with $W_{k_1,n}(N_\xi^c(\theta_0, \theta_{k_1})) \leq e^{-\alpha n}$ for some $\alpha > 0$. Then, since $D(P_{\theta_0}\|P_{\theta_{k_2}})$ is lower semi-continuous and hence bounded on compact sets,

$$\begin{aligned}T_n(\theta_0, \theta_{k_1}) &\leq \int_{N_\xi} w(\theta_k)e^{-nD(P\theta_0\|P_{\theta_{k_1}})}d\theta_{k_1} + \int_{N_\xi^c} w(\theta_{k_1})e^{-nD(P\theta_0\|P_{\theta_{k_1}})}d\theta_{k_1}\\ &\leq W_{k_1,n}(N_\xi)e^{-\epsilon n} + CW_{k_1,n}(N_\xi^c)n^{k_2/2}\\ &\leq e^{-\epsilon n} + e^{-\alpha' n} = \mathcal{O}(e^{-n\max(\alpha',\epsilon)}),\end{aligned} \tag{16}$$

where $\alpha' \in (0, \alpha)$, $n^{k_2/2}$ comes from a Laplace's method argument, and $C > 0$ is a constant.

An alternative to careful prior selection is to force nested families to be disjoint. One can perturb the smaller families a little so that they are formally disjoint from the larger families but very little model mis-specification has been introduced. One way to do this in some generality is to observe that many common parametric families such as the normal, binomial, Gamma, Poisson, Dirichlet, multinomial and so forth can be written as $f_\eta(x) = e^{\eta \cdot x - \psi(\eta)} f_0(x)$ i.e., as an exponential tilt, see Blahut (1987) Chap. 4, of a 'carrier' $f_0 \geq 0$ where η is the natural parameter and ψ is the normalizer. Clearly, different f_0's yield different parametric families. The simplest example is to write the $N(\mu, 1)$ density as $f(x) = e^{\mu x - \mu^2/2}\phi(x)$ where ϕ is the density of a $N(0, 1)$. This is a subfamily of the $N(\nu, \sigma^2)$ family but changing ϕ to the density of a t_k random variable for some moderate value of k gives a family that is no longer a subfamily of $N(\mu, \sigma^2)$ even though it is very close to the $N(\mu, 1)$ family.

Likewise, the Exponential(β) distribution can be regarded as a Gamma(1, β) distribution which is nested in the two parameter Gamma(α, β) distribution. One can perturb the carrier of the Gamma(1, β) distribution to make it no longer Gamma(1, β) and hence no longer a subfamily of the Gamma(α, β), but still very similar to the $\alpha = 1$ subfamily.

3.2 Prediction Perspective

Our goal here is to predict the next outcome X_{n+1} given n outcomes x^n. One standard predictor is the Bayes model average, see Hoeting et al. (1999), the summary of the models and their posterior weights. Denote the predictive distribution by $m(x_{n+1}|x^n)$ and the joint posterior i.e., for k and θ_k, by $w(k, \theta_k|x^n)$ with marginal posteriors analogously denoted. Then, in contrast to the standard L^2 justification of the BMA, we can condition on x^n, and state Aitchison's theorem as follows.

Proposition 2: Let Q be a distribution for X_{n+1} having density with respect to the same dominating measure as the p_k's. Then,

$$\arg \inf_Q \sum_{k=1}^K \int w(k, \theta_k|x^n) D(P_{\theta_k}||Q) d\theta_k = m(\cdot|x^n). \tag{17}$$

Moreover, the predictive density $m(x_{n+1}|x^n)$ occuring in (17) is the BMA for the next outcome and the point predictor under squared error loss is its condtional expectation given $X^n = x^n$.

Proof: The result follows immediately from writing

$$m(x_{n+1}|x^n) = \sum_{k=1}^K \int p(x_{n+1}|k, \theta_k) w(\theta_k|k, x^n) w(k|x^n) d\theta_k = \sum_{k=1}^K m(x_{n+1}|x^n, k) w(k|x^n) \tag{18}$$

and applying Aitchison's theorem with the posterior in the role of the prior. □

Thus, in the M-closed setting, model selection identifties a model while averaging generally does not. On the other hand, predictions from model averages tend to be better than predictions from selected models, especially for small samples. Moreover, because M_n is Bayes optimal under relative entropy and $D(P_\theta^n \| M_n) = \sum_{i=0}^{n-1} E_\theta D(P_\theta \| M(\cdot|X^i))$, (17) means the BMA is sequentially the minimum relative entropy density predictor and estimator.

4 M-Complete Class

Information theoretic ideas have been explored in M-complete problems from a variety of perspectives. In terms of model identification there have been two basic approaches although both involve

finding parametric families with useful information theoretic properties. One is to find an ME or MDI model and assess its sensitivity to perturbations of its inputs hoping to be satisfied that the stability is reasonable. Another approach invokes the RDF to find a generic collection of models that lead to very similar inferences. In both cases, the argument is that if the inferences are robust to the modeling strategy they are more likely to be valid. On the other hand, from a prediction perspective, the AIC and its offshoots have had a greater impact than robustness.

4.1 Model Identification

Recall that (3) identifies a parametric family say Ω_λ consisting of elements $p^*(x|\lambda_0^K)$ based on maximum entropy given a collection of statistics T_k for $k = 1, \ldots, K$ with constraints $ET_k(X) = \lambda_k$. It is natural to ask what choices for the T_k's are reasonable and whether any of the resulting p^*'s actually fit the data generating mechanism. This comes down to asking (i) if the parameter estimates are reasonable, see Soofi et al. (1995) Sec. 4, (ii) if the constraints match those of the data generator, see Ebrahimi et al. (2008), and (iii) fundementally how these ensure (or fail to ensure) that $D(P\|P_\lambda^*)$ is small, see Mazzuchi et al. (2008). Here, we use P with density p to mean the actual data generator which we hope is an element of Ω_λ and P_λ^* to mean the probability associated with the density $p^*(x|\lambda_0^K)$.

Observe that if we model the data generator P by $p^*(x|\lambda_0^K) \in \Omega_\lambda$ as in (3) then we can write

$$\begin{aligned} D(P\|P_\lambda^*) &= \int p(x) \log \frac{p(x)}{p^*(x|\lambda)} dx = -H(P) + \int p(x) \log \frac{1}{p^*(x|\lambda)} dx \\ &= -H(P) - \int p(x) \left[\log C(\lambda_0^K) + \eta_0 + \sum_{k=1}^K \eta_k T_k(x) \right] dx \\ &= -H(P) - \left[\log C(\lambda_0^K) + \eta_0 + E_p \eta_k T_k(X) \right] \end{aligned} \tag{19}$$

So, it is seen that

$$\exists\ \lambda \quad P = P_\lambda \iff D(P\|P_\lambda^*) = H(P_\lambda^*) - H(P_\lambda) \tag{20}$$

because both sides are equivalent to $E_p \eta_k T_k(X) = E_{p_\lambda} \eta_k T_k(X)$ see Soofi et al. (1995) and Ebrahimi et al. (2008). One implication of (20) is that tests based on the entropy difference on the right in (20) must assume $P \in \Omega_\lambda$ to be valid, see Mazzuchi et al. (2008) p. 431.

Given this, there are two cases. First, if the constraints are reasonable so that the assumption that the data generator is in the parametric family (3) is tenable, we want to know (20) is small. To check this, we must estimate the two terms on the right in (20). Usually, the second is estimated nonparametrically. For instance, one can take the entropy of the empirical distribution function (EDF) $\hat{F}_n$; other estimators are possible as well. From $\hat{F}_n$, we can find $\hat{\lambda}$, an estimate of λ from the constraints, i.e., for all k, set $E_{\hat{F}_n} T_k(X) = \hat{\lambda}_k$ and write the probability associated with $\hat{F}_n$ as $P_{\hat{\lambda}}$. Using this $\hat{\lambda}$, we can define the set $\Omega_{\hat{\lambda}}$ and obtain $P_{\hat{\lambda}}^* = \arg\max_{P \in \Omega_{\hat{\lambda}}} H(P)$, giving the second last term in (20) and therefore a value for $D(P_{\hat{\lambda}}\|P_{\hat{\lambda}}^*)$. This procdure is necessary because $\hat{F}_n$ is a proxy for P that is needed for both terms on the right in (19) and to obtain (20).

If $D(P_{\hat{\lambda}}\|P_{\hat{\lambda}}^*)$ is small, we can assume that the constraints have produced a useful parametric family that has a density in it that we can plausibly use for prediction or other inferences and our estimate of the parameter indentifying that density is good. If the difference is not small then we know that our choice of $\hat{\lambda}$ is poor because we have assumed $P \in \Omega_\lambda$ for some λ.

Second, if the constraints are not reasonable so that $P \notin \Omega_{\hat{\lambda}}$ for any λ then in (20), the estimate $\widehat{H(P_\lambda)} = H(\hat{F}_n)$ of $H(P_\lambda)$ should be far from the estimate of $H(P^*_\lambda)$ leading us to conclude that Ω_λ was poorly chosen and forcing us to rechoose the constraints.

Since P is unknown, the only confusion that can arise occurs when our estimate of $D(P\|P_\lambda)$ is large: We cannot tell whether $P \in \Omega_\lambda$ for some λ and we have chosen a poor $\hat{\lambda}$ (i.e., the nonparametric estimator is not well smoothed by the parametric family) or whether $P \notin \Omega_\lambda$ so Ω_λ itself is poorly chosen. On the other hand, we should be able to distinguish these two cases by verifying that our estimate $\hat{\lambda}$ is good or testing the sensitivity of inferences to different sets of constraints as suggested in Soofi et al. (1995).

Note that this is a method that is to be used for the M-complete case because it de facto assumes a data generator that is approximable and whose distance from a servicable approximation can be assessed. It does not assume we will ever uncover the true data generator, just that we can understand some of its properties and refine it until our approximation is sufficiently good.

Going one step further into the realm of hard-to-get-at models, if the true model is too hard to approximate reliably then we might seek a distribution that can act as a surrogate. Recall the RDF, treat the Θ as a parameter whose information is being compressed into the data, and find the parametric family that achieves the RDF lower bound. So, write (9) and (10) as

$$\mathcal{P}_D = \left\{ p(x|\theta) : \int\int p(x|\theta)w(\theta)d(x,\theta)dxd\theta \le D \right\} \quad \text{and} \quad R(D) = \inf_{p\in\mathcal{P}_D} I(\Theta, X). \tag{21}$$

Note that $I(\Theta, X)$ is Lindley's measure of the information in X about Θ, see Lindley (1956). It is shown in Blahut (1972b) that the RDF lower bound for given D is achieved by

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

where $m^*_\lambda(x)$ is determined from

$$\int w(\theta)\frac{e^{-\lambda d(x,\theta)}}{\int m^*_\lambda(y)e^{-\lambda d(y,\theta)}dy} = 1$$

for x's that have $m^*_\lambda(x) > 0$, where $\lambda = \lambda(D) \ge 0$ is a transformation (usually decreasing) of D from the Bayes risk scale to a factor in the exponent. Here, m^*_λ is the marginal, $m^*_\lambda(x) = \int p^*_\lambda(x|\theta)w(\theta)d\theta$; see Cover and Thomas (1991) Chap. 13.7 for a derivation. See Yuan and Clarke (1999a) for its inferential properties in terms of the resulting posteriors and see Yuan and Clarke (1999b) for an example of how to use these densities in data analysis.

The main technique for obtaining $p^*_\lambda(x|\theta)$ is the Blahut- Arimoto algorithm, see Blahut (1972a,b) and Arimoto (1972); see Csiszar and Tusnady (1984) for its convergence properties. This algorithm is a well-known member of the larger class of alternating minimization algorithms (sometimes called alternating projection algorithms) that have been successful in solving optimization problems in two variables, here x and θ since λ is given. (However, in general neither x nor θ need be unidimensional.) Indeed, the popular EM algorithm can be regarded as a special case of alternating minimization, see Gunawardana and Byrne (2005) who provide a re-interpretation of the EM-algorithm in information-theoretic terms.

The form of the algorithm that we implemented is described in Cover and Thomas (1991), see Chapter 13.8 for the conversion of the RDF into a double minimization problem making the

alternating minimization algorithm class applicable. Fix d, λ, w, and an initial m_0 to form

$$p_{1,\lambda}(x|\theta) = \frac{m_0(x)e^{-\lambda d(x,\theta)}}{\int m_0(y)e^{-\lambda d(y,\theta)}}$$

Marginalizing gives

$$m_1(x) = \int p_{1,\lambda}(x|\theta)w(\theta)d\theta.$$

Next, replace m_0 by m_1 to form $p_{2,\lambda}(x|\theta)$. Now one obtains m_2 from $p_{2,\lambda}(x|\theta)$ by mixing out θ. Thus, one generates a sequence $p_{j,\lambda}(x|\theta)$ for a given λ, x, and θ. Csiszar (1974) showed that $p_{j,\lambda}(x|\theta)$ converges to $p^*_\lambda(x|\theta)$ as $j \to \infty$, independently of m_0. We assess convergence of $p_{j,\lambda}(x|\theta)$ to its limit $p^*_\lambda(x|\theta)$ in supremum norm, terminating when $\sup_{x,\theta} |p_{j,\lambda}(x|\theta) - p_{j-1,\lambda}(x|\theta)| \leq \epsilon$.

To see what $p^*_\lambda(x|\theta)$ looks like for unidimensional θ, we used code in which $\theta \in [-4, 4]$ and $x \in [-8, 8]$ assume values on an evenly spaced grid, both with 200 points, see `www.ccs.miami.edu/\~bclarke/infopapers/PCLIT`. Then, we chose $\lambda = 1.5, 3$, used two priors $N(0, 1)$ and t_2, and three loss functions, squared error, absolute error, and linex loss. The linex loss was studied in Zellner (1986) and is asymmetric; we set $b = 1$ and $a = 2.5$. A representative set of the 12 plots are shown in Fig. 1. The typical shape of a minimially informative likelihood (MIL) $p^*_\lambda(x|\theta)$ is flattish central portion which rises sharply as θ gets close to ± 4. The central portion is flatter for heavier tailed priors i.e., more dispersed input signal. In addition, the 'U' shape of the surface becomes relatively stronger for λ small, i.e., large ℓ, permitting more distortion. In addition, the number of 'bumps' increases as we move from linex, to squared error, to absolute error, i.e., as $d(\cdot, \cdot)$ loses convexity. The effective range of x shifts as θ varies. In all cases, $p^*_\lambda(x|\theta)$ as a function of θ for fixed x and λ is unimodal while for fixed λ and θ it can be multimodal in x.

Since these parametric families are data-compression optimal in their various settings, we can use them to form posteriors and assess the stability of the inferences. So, consider the data on 2010 constant price percent change in GDP for the 17 Eurozone countries available at `imf.org`. These data are very difficult to model, but one can imagine they admit a model, however complex, that would be impossible to approximate well and thus would be in the M-complete problem class. Using these 17 data points, we can form a posterior for each of the 12 parametric families. Four of these are shown in Fig. 2. These posteriors are representative of the 12 that we obtained and are broadly in agreement over values of λ and choices of d and prior. All have roughly the same location; the differences are in the dispersion and these are (arguably) not very large. The seeming bimodality in the second row of panels in Fig. 2 is a computational artifact; requiring a smaller ϵ for uniform approximation (in x and θ of $p_{j,\lambda}(x|\theta)$) would smooth these out at the cost of more computing time unless a more sophisticated algorithm were used.

4.2 Prediction Perspective

The AIC is one of the most interesting information criteria because it has been examined in so many ways. Originally, given in Akaike (1974), here we present a clarified derivation. The goal is to find the distribution Q, from a collection of distributions, that is as close as possible to an unknown true distribution, P. That is one seeks to minimize

$$D(P\|Q) = -H(P) - \int p(x) \log q(x) dx \tag{22}$$

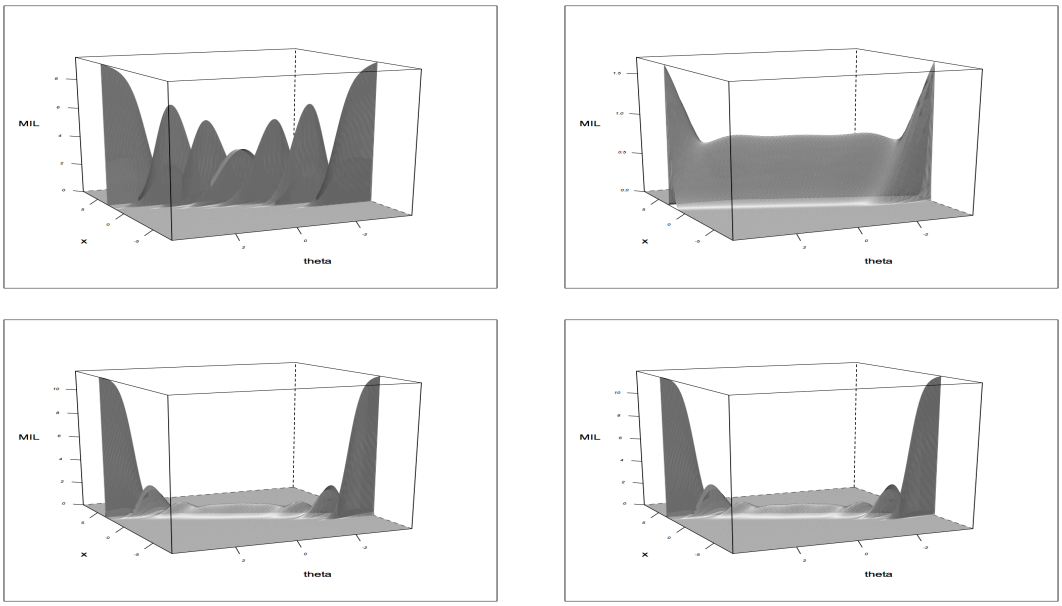

Figure 1: The four panels show examples of p^*_λ's. Upper left: Absolute error, t_2 prior, $\lambda = 1.5$. Upper right and lower left: Squared error with normal and t_2 priors and $\lambda = 1.5, 3$ respectively. Lower right: Linex loss with t_2 and $\lambda = 3$.

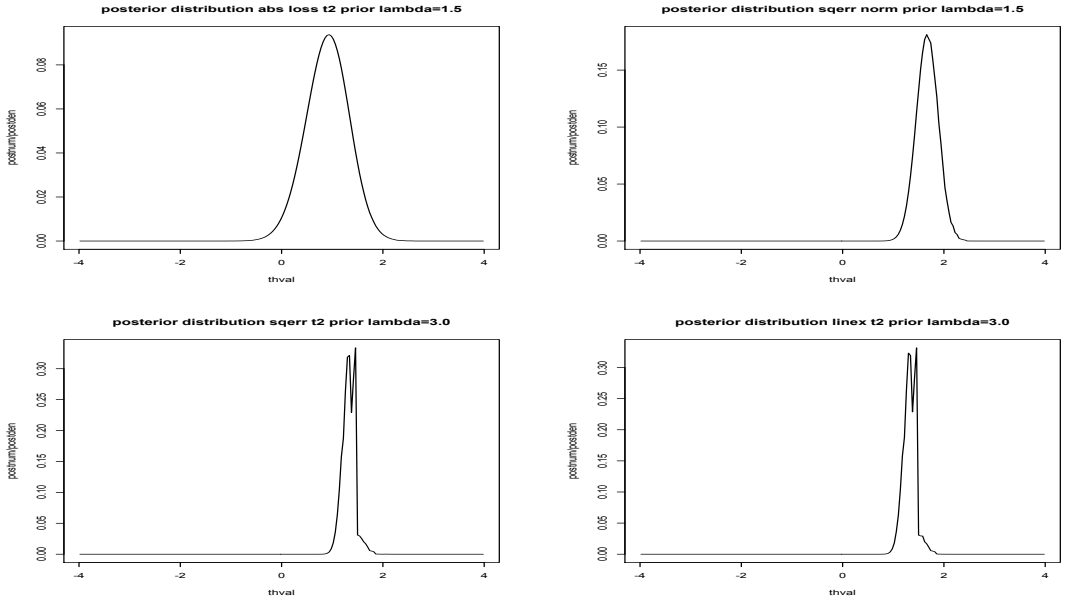

Figure 2: The four panels show the posteriors corresponding to the parametric families in 1.

over Q. For practical purposes, it is enough to maximize the integral in the second term over q. The problem is that this cannot be done because P is unkown. To make the problem easier, let us assume Q comes from a parametrized family of densities $q(\cdot|\theta_k)$ for a range of k and write

$$\Delta(\theta_k) = \int p(y) \log q(y|\theta_k) dy = E_Y \log q(Y|\theta_k),$$

where E_Y is the expectation over Y using P. Now we can consider $\Delta(\hat{\theta}_k)$ where $\hat{\theta}_k = \hat{\theta}_k(x) = \arg\min q(x|\theta_k)$ is the MLE from $q(\cdot|\theta_k)$ given x. Now, even though P is unknown so we cannot use $\Delta(\hat{\theta}_q)$ directly, we can approximate it by $\log q(x|\hat{\theta}_k)$ which is known. This would lead us to choose the family q_k with the highest maximized likelihood – a reasonable proposition except that adding more parameters will typically increase the maximized likelihood. Another problem is that $\log q(x|\hat{\theta}_k)$ as an approximation for $\Delta(\hat{\theta}_k)$ is biased. So, we can improve the approximation by using

$$\mathsf{Bias} = E_X \left[\Delta(\hat{\theta}_k) - \log q(X|\hat{\theta}_k) \right]$$

as a correction to $\log q(x|\hat{\theta}_k)$, where E_X means E is taken over X using P. To get a useful form for Bias, write it as

$$E_X \left[\Delta(\hat{\theta}_k) - \log q(X|\theta_k) \right] + E_X \log \frac{q(X|\theta_k)}{q(X|\hat{\theta}_k)}, \tag{23}$$

for some choice of θ_k. The first term in (23) is

$$E_X E_Y \left[\log q(Y|\hat{\theta}_k) - \log q(Y|\theta_k) \right] \tag{24}$$

because $E_X \log q(X|\theta_k) = E_Y \log q(Y|\theta_k)$. Now, setting $p(\cdot) = q(\cdot|\theta_k)$ for some fixed value of θ_k, (24) becomes

$$-E_X D(Q(\cdot|\theta_k) \| Q(\cdot|\hat{\theta}_k)). \tag{25}$$

So, if $Y = (Y_1, \ldots, Y_n)$ where the Y_i's are IID, (25) becomes

$$-E_X n D(\theta_k \| \hat{\theta}_k) = -\frac{1}{2} E_X n (\theta_k - \hat{\theta}_k)^2 I(\tilde{\theta}_k) \tag{26}$$

by Taylor expanding the relative entropy $D(\theta_k \| \hat{\theta}_k)$ between $q(\cdot|\theta_k)$ and $q(\cdot|\hat{\theta}_k)$ at $\hat{\theta}_k$. Here, $I(\theta)$ is the Fisher information of $q(\cdot|\theta_k)$ evaluated at $\tilde{\theta}_k$, a point between θ_k and $\hat{\theta}_k$. By the choice of Y we also get $X = (X_1, \ldots, X_n)$ IID so by consistency of the MLE we get that $\tilde{\theta}_k \to \theta_k$ in probability and hence $I(\tilde{\theta}_k) \to I(\theta_k)$. Making this substitution lets us invoke the asymptotic normality of the MLE $\hat{\theta}$ and so recognize the argument of the expectation in (26) as a χ^2 with $\dim(\theta_k)$ degrees of freedom, asymptotically as $n \to \infty$. So, we can approximate the first term in (23) as $-\dim(\theta_k)/2$. Likewise, the second term in (23) can be recognized asymptotically as $-\dim(\theta_k)/2$ by assuming Wilks' theorem in L^1. Taken together, $\mathsf{Bias} \approx -\dim(\theta_k)$ and therefore maximizing $\log q(x^n|\hat{\theta}(x^n)) - \dim \theta_k$ is approximately equivalent to minimizing (22) when P is of the form $q(\cdot|\theta_k)$. So, we have a model selection principle, ostensibly for the M-closed setting.

Let us compare the sampling distributions of the AIC and BIC for selecting among a large but finite collection of linear models based on a common set of explanatory variables. First, the

sampling distribution of BIC will be much tighter than that of the AIC. This means the BIC is more efficient. Second, the sampling distribution of the BIC is located more-or-less at the data generator (assuming it is in the family of models) while the AIC is frequently located away from the data generator. That is, the AIC is inconsistent, see Shibata (1983), because the probability of choosing too large a model is not small.

So, if the AIC is not very good for model selection....what is it good for? One answer is prediction. Shibata (1981) showed that maximizing the AIC in the limit of increasingly many variables provides a way to achieve the optimal sum of squared errors in prediction in the limit of increasingly many models and sample size. Moreover, Shibata (1980) shows that the AIC has an efficiency property in the same limiting sense. So, although intended for model selection in the M-closed case, ironically, AIC has its optimality properties in the *predictive* M-complete case. Most recently, this same sort of property – consistency, asymptotic in the size of the model space – has been demonstrated for a variety of methods including the AIC, see Yanagihara et al. (2012).

Suppose we apply a criterion such as AIC to a model list and are in the imprudent case that one of the variables physically necessary for describing the given phenomenon only occurs in one of the models, has a meaningful but not overwhelming effect, and we were unlucky enough to get a data set that did not let us include it. Then, as the sample size grows, even if the model list grows and we get a model that appears consistent, it will be only be consistent in the sense that the variable we have incorrectly omitted has been itself modeled by other variables – even though the omitted variable is physically necessary to describing the response.

This property of the AIC – searching for models that are useful – is used in van Erven et al. (2012). They advocate using the AIC initially and then at some point switching to the BIC. Thus, the AIC gets you in the right neighborhood of the true model and the BIC zeros in on the data generator. Essentially, the AIC is being used to convert an M-complete problem into an M-closed problem that BIC can solve well.

Another property that the AIC has that makes it compelling in an M-complete setting is that optimality in the limit of large models is also seen in the oracle inequality of Yang (2005) who shows that for smooth classes the squared error risk of using AIC converges at the minimax optimal rate. This result extends Akaike (1978) who showed minimax optimality in a normal case.

Note that despite BMA and other model averaging methods having been applied successfully in M-closed problems there is little literature on the use of averaging methods for prediction in the M-complete case. An exception might be called Akaike model averaging using the AIC weights rather than the BIC weights to form an average of models of varying complexity. This has been suggested, but to date neither a theoretical nor compuational comparison seems to have been done. See Burnham and Anderson (1998) Chap. 4.2 for a discussion and statistical references. Interestingly, areas of application that confront large degrees of model uncertainty i.e., M-complete problems, have already started using Akaike model averaging to good effect see Johnson and Omland (2004) and Symonds and Moussalli (2010).

To complete this discussion of the AIC, we remark that Akaike (1981) proposes a Bayesian version of his criterion. The basic criterion seems to be to add the 'prior predictive log-likelihood' to the 'incremental log-likelihood', respectively and then maximixe

$$\begin{aligned} AIC_B(k) &= \int p_k(x^n|\theta_k)p_k(y^n|\theta_k)\log m_k(y^n|x^n)w_k(\theta_k)dx^n dy^n d\theta_k \\ &+ \left[\int m_k(x^n)\log m_k(x^n)dx^n - \log m_k(x^n)\right]. \end{aligned}$$

This makes use of the data twice as in the earlier derivation of the AIC.

Finally, to complete this section we discuss another information-type criterion called the deviance information criterion or DIC first introduced by Spiegelhalter et al. (2002) but subsequently developed by Berg et al. (2004) and Celeux, et al. (2006) for an EM/missing data context. The central idea is to define a general concept of deviance

$$D(\theta) = -2\log p_k(x^n|\theta_k) + 2\log h(x^n)$$

where h is a known e.g., the constant function one or $m(x^n)$. One can regard

$$p_D = E_{\Theta|x^n} D(\Theta) - D(E_{\Theta|x^n}\Theta) = \overline{D(\theta)} - D(\bar{\theta}),$$

the posterior mean deviance minus the deviance at the posterior mean (with mild abuse of notation), as an effective dimension or more accurately as a complexity measure. Now, by analogy with the AIC or BIC, the DIC is a combination of fit and complexity:

$$\begin{aligned} DIC &= \bar{D(\theta)} + p_D = D(\bar{\theta}) + 2p_D = 2\bar{D(\theta)} - D(\bar{\theta}) \\ &= -4E_{\Theta|x^n}\log p_k(x^n|\theta_k) + 2\log p_k(x^n|\bar{\theta}). \end{aligned} \tag{27}$$

Some simple manipulations using the results in the Appendix A give us an asymptotic approximation to (27) and provide slightly more interpretability. In the simplest case, (27) is

$$\begin{aligned} DIC &\approx -4\int w(\theta|x^n)\log w(\theta|x^n) - 4\int w(\theta|x^n)\log\frac{m(x^n)}{w(\theta)} + 2\log p(x^n|\hat{\theta}) \\ &= 4H(\Theta|X^n = x^n) - 4\int w(\theta|x^n)\log\frac{1}{w(\theta)}d\theta - 4\log m(x^n) + 2\log p(x^n|\hat{\theta}) \\ &\approx 4H(\Theta|X^n = x^n) + 4\log w(\theta) - 4\log m(x^n) + 2\log p(x^n|\hat{\theta}) \\ &= 4H(\Theta|X^n = x^n) + 2\log\frac{w(\theta)p(x^n|\hat{\theta})}{m(x^n)} + 2\log w(\theta) - 2\log m(x^n) \end{aligned} \tag{28}$$

where the first approximation is using the MLE rather than the posterior mean and the second uses the concentration of the posterior $w(\theta|x^n)$ at θ.

The first term is the conditional entropy of the posterior which is asymptotically normal, i.e., $w(\theta|x^n) \sim N(\hat{\theta}, (nI(\theta))^{-1}$. This means

$$H(\Theta|x^n) \approx \frac{1}{2}\log[(2\pi e)^{d_k}|nI(\theta)|]. \tag{29}$$

Moreover, as noted in Lemma 3 of Appendix A,

$$\log\frac{p(X^n|\hat{\theta}_{k_0})w_k(\hat{\theta}_{k_0})}{m_{k_0}(X^n)} = \frac{1}{2}\log(2\pi)^{d_k}|nI(\theta)| + o_P(1). \tag{30}$$

Using (29) and (30) in (28) provides a lot of simplification, focussing on the sum of the conditional entropy $H(\Theta|x^n)$ – an average shortest codelength criterion – and $-\log 1/(w(\theta)m(x^n))$ which can be regarded as a sort of Shannon codelength under the product of marginals distribution for (Θ, X^n). Otherwise put, minimizing the DIC over models is like asking the sum of posterior codelengths for θ and the joint codelengths for (Θ, X^n) not to be any bigger than necessary.

5 M-Open

The M-open class of problems is more complex than either the M-closed or M-complete classes. So, we expect tecniques that will perform well for M-open problems will be more complex than those required for M-closed or M-complete problems. As before, we consider the identification and prediction perspectives. However, the nature of M-open problems precludes model identification. So, here, we take identification to mean the use of a unitary predictor i.e., a predictor not comprised of subunits. By contrast, the prediction perspective means we seek a composite predictor, one that is comprised of subunits that, hopefully, are intelligible. While heuristic, this distinction provides a dividing line between the techniques of the two subsections below.

5.1 Predictor Identification

The task is to identify a predictor for a sequence $x_1, \ldots, x_n, ...$ without using knowledge of the data generator. The agent issuing the predictions is called the Forecaster, F and the forecasts are denoted p_i for x_i at time i; it is more typical for forecasts to be probabilistic than numerical. We assume F has access to 'experts' E_θ who give predictions at time i based on $q(\cdot|\theta)$ where θ is an index of the various experts. The experts issue probability forecasts that F may use. Overall, this can be treated as a sequential game between F and nature, N. Round 1 begins with each expert indicating a prediction for x_1. Then, F takes these predictions and forms another, hopefully good, prediction. Finally N reveals x_1 so that F and N settle up. Round 2 is the same excecpt that F, the E_θ's, and N retain knowledge of previous predictions and outcomes; there is no restriction on how N can choose x_2. There are many variations on this basic sequential game and an important compendium of material on this is Cesa-Bianchi and Lugosi (2006). However, key results go back to Shtarkov (1988), Haussler and Barron (1992), and Cesa-Bianchi et al. (1997), among others.

Here, we focus on probability predictions and assume that the loss is assesed by codelength. So, instead of evaluating the diffrence between x_i and p_i directly we phrase our cost in terms of the difference between the codelengths of two words. The basic version of this is the following.

Suppose at stage i expert E_θ issues a distribution $q_\theta(\cdot|x^{i-1})$ and F issues a prediction $p(\cdot|x^{i-1})$. Once N reveals x_i, F's loss is $\ell(p(\cdot|x^{i-1}), x_i) = \log 1/p(x_i|x^{i-1})$ and E_θ's loss is $\ell(q_\theta(\cdot|x^{i-1}), x_i) = \log 1/q_\theta(x_i|x^{i-1})$. Writing $p^n = p(x^n)$ and $q_\theta^n = q(x^n|\theta)$, F's cumulative loss over n rounds is $\ell(p^n, x^n) = \log 1/p(x^n)$ and E_θ's is $\ell(q_\theta^n, x^n) = \log 1/q(x^n|\theta)$.

A standard goal for F is to predict as well as the best expert by minimizing the 'regret' – the extra bits one has to send due to not knowing which expert is the best. The (cumulative) regret of not following E_θ is

$$\ell(p^n, x^n) - \ell(q_\theta^n, x^n) = \log \frac{q(x^n|\theta)}{p(x^n)}.$$

This can be maximized over θ and then minimized over x^n. So, it is reasonable to predict using the p^n with the smallest minimax regret, namely

$$p^* = \arg\inf_p \left[\sup_{x^n} \sup_\theta \log \frac{q(x^n|\theta)}{p(x^n)} \right], \tag{31}$$

when it exists. A theorem due to Shtarkov (1988), see also Cesa-Bianchi and Lugosi (2006) (Chap. 9.4) gives that

$$p^*(x^n) = \frac{p(x^n|\hat{\theta})}{\int p(x^n|\hat{\theta}) dx^n},$$

where $\hat{\theta}$ is the MLE. This is analogous to Proposition 2. This solution has been used in Rissanen (1996) and the Bayesian version examined in Clarke (2007).

There are two steps to reformulating this sequential game so the solution will be more applicable in the M-open case. First, assume there is side information available for each round. For instance, for each i, each expert E_θ may be of the form $q_\theta(x_i|x^{i-1}, z_i)$ where the extra information z_i as available to the experts and to F. Often z_i is understood to be a selection of the outcomes in $(x_1, \ldots, x_{i-1})$ indicating which ones are relevant at round i. Then, the experts reduce to $p_\theta(x_i|z_i)$ and a treatment analogous to (31) is in Cesa-Bianchi and Lugosi (2006), Chap. 9.9 and 9.10.

However, seeking a best overall expert contradicts the M-open feature that the data generator is not structured enough for the 'best expert' to be a useful concept. So, a further generalization of the structure is needed to permit the experts, i.e. predictors, to use the information more flexibly whether it is past data in the x_i's or other data in the z_i's. Let $S_{i-1} = (S_{i-1,1}, \ldots, S_{i-1,m_{i-1}})$ where $S_{i-1,j}$ depends only on the x's and z's up to time $i-1$. That is, S_{i-1} is a string of length m_{i-1} of statistics for use at time i. Let $S^{i-1} = (S_1, \ldots, S_{i-1})$ be the growing collection of statistics. An expert is now of the form $q_\theta(x_i|S^{i-1})$ and F's choice at stage i is still denoted $p(x_i)$. Now, at stage i, F might adopt a Bayesian view and weight experts θ by a prior $w(\theta)$ and choose $p(x_i)$ to make

$$\log \frac{w(\theta) q_\theta(x_i|S^{n-1})}{p(x_i)}$$

small over θ. So, F's Bayesian cumulative regret can be written

$$\log \frac{\prod_{i=1}^n w(\theta) q_{\theta_i}(x_i|S^{n-1})}{\prod_{i=1}^n p(x_i)},$$

where $\prod_{i=1}^n p(x_i)$ is merely convenient notation for F's optimal choice, suppressing dependence on θ's, x_i's and z_i's which may be present. It is natural to seek

$$\arg \inf_{p(x_1),\ldots,p(x_n)} \left[\sup_{x^n} \sup_{\theta_1,\ldots,\theta_n} \sup_{S^n} \log \frac{\prod_{i=1}^n w(\theta) q_{\theta_i}(x_i|S^{n-1})}{\prod_{i=1}^n p(x_i)} \right],$$

and this amounts to finding the best sequence of experts to follow admitting that the particular way each expert uses information is also part of the optimization. Thus a solution must identify a way to select the S_i's and θ_i's. Examples of this have been found but there is little general theory.

5.2 Prediction Perspective

This class of problems is so difficult even to conceive clearly that developments of prediction methods remain in their infancy. However, there are signs of growth. The concept of information-theoretic learning in reproducing kernel Hilbert spaces (RKHS) i.e., based on relevance vector machines that are optimal in RKHS and have information theoretic properties, is gaining some attention see Principe (2010). Loosely, the idea is that RKHS's come up frequently in signal processing, a key topic in information theory even if only indirectly based on Shannon theory. So the role of the reproducing kernel and the implications of the structure it imposes has an effect on the storage, compression, and transmission of information. The Representer Theorem, see Scholkopf and Smola (2002) Chap. 4, is a key result in RKHS's and gives an optimal predictor in an M-open setting. Choosing the kernel information theoretically or using some kind of information

driven model average of Representer Theorem solutions may provide good performance for problems in the M-open class, although how these individual learners combine for improved prediction or decoding under information theoretic principles remains unclear. However, the body of work, already extensive, is growing.

6 Conclusions

Here we have partitioned the class of statistical problems into three subclasses, M-closed, M-complete, and M-open depending on the complexity of the modeling task. Within each class we have defined two perspectives, the identification perspective and the prediction perspective. Thus, we have examined six problem classes. We have argued that the techniques appropriate for each of these six classes are different and identified what we think are possible if not good techniques for each. The take-home message here is an instance of the 'no-free-lunch' theorem, Wolpert (1997): Good performance of a technique on one class of problems will be balanced by poor performance of the technique on another class of problems so that generic features of each given problem must be considered for optimal technique selection.

Acknowledgements: The authors are grateful to Prof. E. Soofi whose comments and suggestions greatly improved this paper.

Appendix A: Regularity Conditions and Background Results

Here we briefly record the regularity conditions mentioned in Theorem 1 and the implications that we use from them. Note that the first three are essentially the Wald (1949) hypotheses for consistency of the MLE. The remaining assumptions are to control the behavior of the likelihood.

A1 For each x, as $\|x\| \to \infty$, $p(x|\theta) \to 0$.

A2 For some large $r > 0$,

$$E\left(\log \sup_{\theta' : \|\theta' - \theta_0\| > r} \frac{P(X|\theta')}{p(X|\theta_0)}\right)^2 < \infty.$$

A3 For each θ and $\delta > 0$ small enough, the function

$$\rho(x|\theta, \delta) = \sup_{\theta' : \|\theta' - \theta\| < \delta} p(x|\theta')$$

satisfies

$$E\left(\log \frac{p(X|\theta_0)}{\rho(X|\theta, \delta)}\right) < \infty.$$

A4 For each x, $p(x|\theta)$ is twice continuously differentiable with respect to θ.

A5 The Fisher information matrix $I(\theta_0)$ exists and is positive definite.

A6 The prior density w_k with respect to Lebesgue measure on $\mathbb{R}$ is continuous and $w(\theta_0) > 0$.

A7 For some $\xi > 0$ and every $j, k = 1, \ldots, d$

$$E \sup_{\|\theta - \theta_0\| < \xi} \left| \frac{\partial^2}{\partial \theta_j \partial \theta_k} \log p(X|\theta) \right|^2 < \infty.$$

Now, it is not hard to prove the following.

Lemma 1: Under assumptions A1 through A6, the event

$$\frac{P_{k,\hat{\theta}}(X^n)}{m_k(X^n)} \leq a(1-e^{-a'n})\left(\frac{n}{2\pi}\right)^{d/2}(\det \ln \hat{I}(\hat{\theta}))^{1/2}/w(\hat{\theta})$$

has probability going to 1, under P_{k,θ_0}, as $n \to \infty$. Here, $\hat{\theta}$ is the maximum likelihood estimator under $P_{k,\theta}$, $\hat{I}(\cdot)$ is the empirical Fisher information and $a, a' > 0$ are constants. If A7 is added, then the empircal Fisher information $\hat{I}$ can be replaced by its limit $I(\theta_0)$, again with P_{k,θ_k} probability going to one as $n \to \infty$.

Proof: This is substantially due to Walker (1969), see also Clarke and Barron (1988), Appendix A) For the last statement, take the supremum over a small set around $\hat{\theta}_0$ inside the expectation, invoke the consistency from A1–A3 and apply A7. □

Lemma 2: With P_{θ_k} probability going to one as $n \to \infty$, A1–A6 give the inequality

$$\frac{P_{k,\theta_0}(X^n)}{P_{k,\hat{\theta}_0}(X^n)} \leq e^{-n\hat{I}(\tilde{\theta})(\hat{\theta}_0-\theta_0)^2}$$

Under A7 we also get

$$\frac{P_{k,\theta_0}(X^n)}{P_{k,\hat{\theta}_0}(X^n)} \leq e^{-n(1+a)I(\theta_0)(\hat{\theta}_0-\theta_0)^2}$$

for some $a > 0$, where $\tilde{\theta} \in B(\theta_0, \hat{\theta}_0)$, in P_{θ_0} probability.

Proof: Assumptions A1–A3 guarantee that $\hat{\theta}_0$ is consistent for θ_0 so it is enough to focus on the set $\{\|\theta_0 - \hat{\theta}_0\| \leq \eta\}$ for some pre-assigned η. On such as set, a second order Taylor expansion of $\log p(x|\theta)$ at $\hat{\theta}_0$ gives the result, with A7 used to replace the empirical Fisher information with the correct Fisher information and the $\tilde{\theta}$ by θ_0. □

Lemma 3: Let A1–A6 be satisfied. Then, for fixed k, as $n \to \infty$,

$$\left|\log \frac{p(X^n|\hat{\theta}_k)w_k(\hat{\theta}_k)}{m_k(X^n)} - \frac{d}{2}\log\frac{n}{2\pi} - \frac{1}{2}\log\det \hat{I}(\tilde{\theta})\right| \to \infty.$$

If, in addition, A7 is satsified then

$$\left|\log \frac{p(X^n|\hat{\theta}_k)w_k(\theta_k)}{m_k(X^n)} - \frac{d}{2}\log\frac{n}{2\pi} - \frac{1}{2}\log\det I(\theta)\right| \to \infty.$$

Proof: This follows from applying Laplace's method to $m(x^n)$. □

Appendix B: Proofs of Main Results in Section 3

Proof of Proposition 1: Write

$$\frac{m_k(x^n)}{m_{k_0}(x^n)} = \frac{m_k(x^n)}{p_{\theta_0}(x^n|k_0)}\frac{p_{\theta_0}(x^n|k_0)}{m_{k_0}(x^n)}. \tag{32}$$

The first factor in (32) is

$$\frac{m_k(x^n)}{p_{\theta_0}(x^n|k_0)} = \int w_k(\theta_k)e^{-n((1/n)\ln\frac{p(x^n|\theta_0)}{p(x^n|\theta_k)})}d\theta_k = \int w_k(\theta_k)e^{-n\hat{D}(P_{\theta_0}||P_{\theta_k})}d\theta_k \tag{33}$$

where $\hat{D}(P_{\theta_0}||P_{\theta_k})$ is the empirical relative entropy. Let us ignore the integration over θ_k in (33) write for each fixed $\theta_k \in \Omega_k$

$$G(\theta_k) = \{\hat{D}(P_{\theta_0}\|P_{\theta_k}) - D(P_{\theta_0}\|P_{\theta_k}) < (1/2) \min_{\theta_k \in \Omega_k} D(P_{\theta_0}\|P_{\theta_k}) \tag{34}$$

so that

$$G^c(\theta_k) = \{\hat{D}(P_{\theta_0}\|P_{\theta_k}) > D(P_{\theta_0}\|P_{\theta_k}) + (1/2) \min_{\theta_k \in \Omega_k} D(P_{\theta_0}\|P_{\theta_k}). \tag{35}$$

Now, we can upper bound (33) by

$$\begin{aligned}
\int \quad & w_k(\theta_k)\mathbb{I}_{G_{\theta_k}} e^{-(n/2)\min_{\theta_k\in\Omega_k} D(P_{\theta_0}||P_{\theta_k})} e^{-nD(P_{\theta_0}||P_{\theta_k})} d\theta_k \\
+ \quad & \int w_k(\theta_k)\mathbb{I}_{G^c_{\theta_k}} e^{-\hat{D}(P_{\theta_0}||P_{\theta_k})} d\theta_k \\
\leq \quad & e^{-(n/2)\min_{\theta_k\in\Omega_k} D(P_{\theta_0}||P_{\theta_k})} + \int w_k(\theta_k) e^{-(n/2)\min_{\theta_k\in\Omega_k} D(P_{\theta_0}||P_{\theta_k}) + D(P_{\theta_0}\|P_{\theta_k})} \\
\leq \quad & 2e^{-(n/2)\min_{\theta_k\in\Omega_k} D(P_{\theta_0}||P_{\theta_k})} \int w_k(\theta_k) e^{-nD(P_{\theta_0}|||P_{\theta_k})} d\theta_k \\
\leq \quad & 2T_n(\theta_0, k) = \mathcal{O}(n^{d_k/2}).
\end{aligned} \tag{36}$$

This follows because $e^{-nD(P_{\theta_0}||P_{\theta_k})}$ is bounded by one since (i) the relative entropy is lower-semicontinuous in its second argument for fixed values of its first argument and (ii) the union of compact sets Ω_k over k is compact in the topology of setwise convergence.

For the second factor in (32), A1–A7 in Appendix B and Lemma 2 ensure that

$$\frac{p_{\theta_0}(x^n|k_0)}{m_{k_0}(x^n)} = \frac{p(x^n|\theta_0)}{p(x^n|\hat{\theta}_0)} \frac{p(x^n|\hat{\theta}_0)}{m(x^n|k_0)} \leq \Big(\frac{n}{2\pi}\Big)^{d_{k_0}/2} \sqrt{\det[I(\theta_0)]} e^{-n(\hat{\theta}-\theta_0)^T I_{k_0}(\theta_0)(\hat{\theta}_0-\theta_0)} = \mathcal{O}(n^{d_{k_0}/2}), \tag{37}$$

with P_{θ_0}-probability going to one, where $\hat{\theta}$ is the MLE in the k_0-th parametric family and $I_{k_0}(\cdot)$ is the Fisher information of p_{k_0}.

Now, using (36) and (37) in (32), the rate from the index of disjointness rate wipes out the rate from the k_0 model so the proposition follows. □.

Proof of Theorem 1: For countably many k, writing $P^n_{\theta_k}$ for the n-fold product of P_{θ_k}'s, Aitchison's Theorem can be stated as

$$\arg \inf_Q \sum_{k=1}^K \int w(\theta, k) D(P^n_{\theta_k}||Q_n) d\theta_k = m(\cdot).$$

Now, suppose model k_0 is true. Then we have

$$\begin{aligned}
\log \frac{1}{m(x^n)} &= \log \frac{1}{w(k_0)m_{k_0}(x^n) + \sum_{k\in\{1,\ldots,K\}\setminus\{k_0\}} w(k) m_k(x^n)} \\
&= \log \frac{1}{m_{k_0}(x^n)} + \log \frac{1}{w(k_0) + \sum_{k\in\{1,\ldots,K\}\setminus\{k_0\}} w(k) m_k(x^n)/m_{k_0}(x^n)}.
\end{aligned} \tag{38}$$

If $k_0 < K'$ ofr some $K' \leq K$, then Proposition 1 suffices. Otherwise, we must again see that the second term in (38) is negligible.

For $k \geq K'$ we use a version of (33) that follows from (36) and our hypothesis on $T_n(\theta_0, k)$: With P_{θ_0}-probability going to one

$$\frac{m_k(x^n)}{p_{\theta_0}(x^n|k_0)} \leq e^{-\gamma n}. \tag{39}$$

Using this with (37) we get an extension of Proposition 1.

As before, when model k_0 is true and the true parameter value of θ_{k_0} with MLE $\hat{\theta}_{k_0}$

$$\begin{aligned} \log \frac{1}{m_{k_0}(x^n)} &= -\log p(x^n|\hat{\theta}_{k_0}) + \frac{d_{k_0}}{2}\log n - \frac{1}{2}\log(2\pi)^{d_{k_0}}|I(\hat{\theta}_{k_0})| + \log \frac{1}{w_{k_0}(\hat{\theta}_{k_0})} + o_P(1) \\ &= -BIC(k_0) + \hat{C} + o_P(1) \end{aligned} \tag{40}$$

in $P_{\theta_{k_0}}$-probability as $n \to \infty$ and $\hat{C}$ converges to a constant. So, if we choose a uniform partition of the space of x^n-values with side length Δ then as $D \to 0$ the partition becomes finer and finer. Write the discretized form of $m_{k_0}(x^n)$ and $m_{k_0,\Delta}(x^n)$. Then, the Shannon codelengths satisfy

$$\log\lceil \frac{1}{m_{k_0,\Delta}(x^n)} \rceil \approx \log\lceil \frac{1}{m_{k_0}(x^n)} \rceil$$

which can be approximated using (40). □

References

Aitchison, J.(1975) Goodness of prediction fit. *Biometrika* **62**, 547-554.

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

Akaike, H.(1978) A Bayesian analysis of the minimum AIC procedure. *Ann. Inst. Stat. Math.*, **30**, 9-14.

Akaike, H.(1981) Likelihood of a model and information criteria. *J. Econ.*, **16**, 3-14.

Arimoto, S.(1972) An algorithm for computing the capacity of arbitrary discrete memoryless channels *IEEE Trans. Inform. Theory* **18**, 14-20.

Ash, R.(1965) *Information Theory* Dover, NY.

Barron, A. and Cover, T.(1991) Minimum complexity density estimation. *IEEE Trans. Inform. Theory*, **37**, 1034-1054.

Berg, A., Meyer, R., and Yu, J.(2004) Deviance information criterion for comparing stochastic volatility models JBES, **22**, 107-120.

Berger, J., Ghosh, J., and Mukhopadhyay, N.(2003) Approximation and consistency of Bayes factors as model dimension grows. *J. Stat. Planning Inference*, **112**, 241-258.

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

Blahut, R.(1972a) Computing of channel capacity and rate-distortion functions *IEEE Trans. Inform. Theory* **18**, 460-473.

Blahut, R.(1972b) *An hypothesis testing approach to information theory* PhD thesis, Cornell University, Ithaca, NY.

Blahut, R.(1987) *Principles and Practice of Information Theory* Addison-Wesley, Reading MA.

Breiman, L.(1994) Bagging predictors *Tech. Rep. 421, Stat. Dpt. UC Berkeley*

Burnham, K. and Anderson, D.(1998) *Model Selection and Inference* Springer, NY.

Catoni, O.(2012) `www.math.ens.fr/cours-apprentissage/Catoni/learning04.pdf`

Celeux, G., Forbes, F., Robert, C. and Titterington, D.(2002) Deviance information criterion for missing data models. *Bayes Analysis*, **1**, 651-674.

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

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

Cesa-Bianchi, N., Helmbold, and Panizza, S.(1997) On Bayes methods for on-line Boolean prediction *NeuroCOLT Technical Report Series* **NC-TR-97-010**.

Clarke, B.(2007) Information optimality and Bayesian models. *J. Econometrics*, **138**, 405429.

Clarke, B. and Barron, A.(1988) Information-theoretic asymptotics of Bayes methods. *Technical Report #26*, Department of Statistics, University of Illinois.

Clemen, R.(1989) Combining forecasts: A review and annotated bibliography. *Int. J. For.*, **5**, 559-583.

Csiszar, I.(1974) On the computation of rate distortion functions. *IEEE Trans. Inform. Theory*, **20**, 122-124.

Csiszar, I. and Tusnady, G.(1984) Information geometry and alternating minimization procedures. *Stat. and Dec.*, 205-237.

de Bruijn, N.(1958) *Asymptotic Methods in Analysis* Dover Publications, NY.

Dembo, A. and Zeitouni, O.(1993) *Large Deviation Techniques and Applications* Springer, NY.

Ebrahimi, N., Soofi, E., and Soyer, R.(2008) Multivariate maximum entropy identification, transformation, and dependence. *J. Mult. Anal.*, **99**, 1217-1231.

Ebrahimi, N., Soofi, E. and Soyer, R.(2010) On the sample information about the parameter and prediction *Stat. Sci.* **25**, 348-367.

van Erven, T., Grunwald, P. and de Rooij, S.(2012) Catching up faster by switching sooner: A predictive approach to adaptive estimation with an application to the AIC-BIC dilemma, with discussion. *J. Roy. Stat. Soc. Ser. B*, **74**, 361-417.

Gunawardana, A. and Byrne, W.(2005) Convergence theorems for alternating minimization procedures. *J. Mach. Learn. Res.*, **6**, 2049-2073.

Haussler, D. and Barron, A.(1992) How well do Bayes methods work for on-line prediction of $\{0, 1\}$ values? *Tech. Report, Computer and Information Sciences, U. Cal. Santa Cruz.*, UCSC-CRL-92-37.

Hoeting, J., Madigan, D. , Raftery, A. and Volinsky, C..(1999) Bayesian model averaging *Stat. Sci.* **14**, 382-401.

Johnson, J. and Omland, K.(2004) Model selection in ecology and evolution. *Trends in Ecology and Evolution*, **19**, 101-108.

Kass, R. and Raftery, A.(1995) Bayes factors *J. Amer. Stat. Assoc.* **90**, 773-795.

Kullback, S. and Leibler, R.(1951). On information and sufficiency. *Ann. Math. Stat.*, **22**, 79-86.

Kullback, S.(1954) Certain inequalities in information theory and the Carmer-Rao inequality. *Ann. Math. Stat.*, **25**, 745-751.

Kullback, S.(1959) *Information Theory and Statistics* Wiley and Sons, NY.

Lindley, D.(1956) On a measure of the information provided by an experiment. *Ann. Math. Stat.*, **27**, 986-1005.

Mazzuchi, T., Soofi, E., and Soyer, R.(2008) Bayes estimate and inference for entropy and information index of fit. *Econ. Rev.*, **27**, 428-456.

Principe, J.(2010) *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*, Springer, NY.

Raftery, A. and Zheng, Y.(2003) Performance of Bayes model averaging *J. Amer. Stat. Assoc.* **98**, 931-938.

Rissanen, J.(1996) Fisher information and stochastic conplexity *IEEE Trans. Inform. Theory* **42**, 40-48.

Sancetta, A.(2012) Universality of Bayesian predictions. *Bayes Anal.*, **7**, 1-36.

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

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

Schwarz, G.(1978) Estimating the dimension of a model *Ann. Stat.* **6**, 461-464.

Shannon, C.(1948a) A mathematical theory of communication. *Bell Sys. Tech. J.*, **27**, 379-423.

Shannon, C.(1948b) A mathematical theory of communication. *Bell Sys. Tech. J.*, **27**, 623-656.

Shibata, R.(1980) Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Ann. Stat.*, **8**, 147-164.

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

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

Shtarkov, Y(1988) Universal sequential coding of single messages *Trans. Problems Information Transmission.* **23**, 3-17.

Soofi, E., Ebrahimi, N. and Habibullah, M.(1995) Information distinguishability with applicatin to the analysis of failure data. *J. Amer. Stat. Assoc.*, **90**, 657-668.

Spiegelhalter, D., Best, N.., Carlin, B., and van der Linde, A.(2002) Bayesian measures of complexity and fit. *J. Roy. Stat. Soc. Ser. B.*, **64**, 583-639.

Symonds, M. and Moussalli, A.(2010) A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaikes information criterion. *Behavioral Ecology and Sociobiology*, **65**, 13-21.

Tsybakov, A.(2009) *Introduction to Nonparametric Estimation.* Springer, NY.

Wald, A.(1949) Note on the consistency of the maximum likelihood estimate *Ann. Math. Stat.*, **20**, 595-601.

Walker, A. M.(1969). On the asymptotic behavior of posterior distributions. *JRSSB*, **31**, 80-88.

Wolpert, D.(1997). No free lunch theorems for optimization. *IEEE Ttrans. Evol.Comp.*, **1**, 67-82.

Yanagihara, H., Wakaki, H. and Fujikoshi, Y(2012) A consistency property of the AIC for multivariate linear models when the dimension and sample size are large. *Tech. Rep. 12-08*, Dept. of Mathematics, Hiroshima University.

Yang, Y.(2005) Can the strengths of AIC and BIC be shared? *Biometrika*, **92**, 937-950.

Yu, B.(1996) Lower bounds on expected redundancy for nonparametric classes. *IEEE Trans. Inform. Theory*, **42**, 272-275.

Yuan, A. and Clarke, B.(1999) A minimally informative likelihood for decision analysis *Can. J. Stat.* **23**, 876-898.

Yuan, A. and Clarke, B.(1999) An information criterion for likelihood selection *IEEE Trans. Inform. Theory* **45**, 562-571.

Zellner, A.(1986) Bayesian estimation and prediction using asymetric loss functions. *J. Amer. Stat. Assoc.* **81**, 446-451.