An Information Criterion for Likelihood Selection

A. Yuan and B. Clarke, *Member, IEEE*

***Abstract*— For a given source distribution, we establish properties of the conditional density achieving the rate distortion function lower bound as the distortion parameter varies. In the limit as the distortion tolerated goes to zero, the conditional density achieving the rate distortion function lower bound becomes degenerate in the sense that the channel it defines becomes error-free. As the permitted distortion increases to its limit, the conditional density achieving the rate distortion function lower bound defines a channel which no longer depends on the source distribution.**

In addition to the data compression motivation, we establish two results—one asymptotic, one nonasymptotic—showing that the the conditional densities achieving the rate distortion function lower bound make relatively weak assumptions on the dependence between the source and its representation. This corresponds, in Bayes estimation, to choosing a likelihood which makes relatively weak assumptions on the data generating mechanism if the source is regarded as a prior.

Taken together, these results suggest one can use the conditional density obtained from the rate distortion function in data analysis. That is, when it is impossible to identify a "true" parametric family on the basis of physical modeling, our results provide both data compression and channel coding justification for using the conditional density achieving the rate distortion function lower bound as a likelihood.

***Index Terms*— Likelihood selection, mutual information, rate distortion.**

I. Introduction

THE Shannon mutual information (SMI) $I(X;Y)$ arises naturally in several settings, including redundancy in source coding, risk in statistical decision theory, rate of transmission in channel coding, and rate of compression in data compression. Even though these settings appear to be very different, they may have a common underlying structure. For instance, Kanaya and Nakagawa [19] used the parallel between rate distortion theory and decision theory to give conditions ensuring that the probability an average loss exceeds a prescribed value goes to zero. This is analogous to to Shannon's rate distortion theorem.

Because the SMI is a measure of dependence between the source and the output, rate distortion is also related to channel transmission. Indeed, the quantities defining optima in these settings are both derived from the SMI. Recall that if one maximizes the SMI over marginal distributions for a source X using a fixed conditional distribution for Y given X to define a channel, the result is the capacity. If, instead, one fixes a marginal distribution for a source X and minimizes the SMI over a class of conditional distributions for Y given X subject to a distortion constraint, the result is the rate-distortion function. The source X has the same interpretation in both cases, although we treat it differently depending on whether we want to transmit it or compress it.

By contrast, the conditional distribution of Y given X, with density denoted by $p(y\,|\,x)$, serves two functions. First, in the channel transmission setting, $p(y\,|\,x)$ is a channel in the usual sense. We send $X = x$, and the receiver gets $Y = y$ which should be decoded to give x. Second, in the data compression setting, $p(y|x)$ is regarded as a description of how X is represented by a codebook with codewords y. In this case, X is a source that is to be represented by as few bits as possible subject to a specified amount of inaccuracy. The data compression problem can be reformulated as the counterintuitive task of seeking the test channel $p(y|x)$ that transmits information as slowly as possible, subject to the distortion bound. (The upper bound on the distortion ensures that we necessarily transmit some useful information.) Here, we use the test channel interpretation of data compression and relate it to statistical decision theory.

More formally, we recall that the SMI is defined from the relative entropy, or the Kullback–Leibler number. The relative entropy between two densities p, q with respect to the same dominating measure, on the same sample space is

$$D(p(\cdot)\|q(\cdot)) = \int p(x)\log\frac{p(x)}{q(x)}\,dx.$$

Here, we have used the Lebesgue measure as the dominating measure and we recall that the relative entropy, although not a metric, does have metric-like properties (see Csiszár [11]–[13]). The SMI is the relative entropy between a joint distribution for two random variables and the product of their marginals,

$$\begin{aligned} I(X;Y) &= \int p(x,y)\log\frac{p(x,y)}{p(x)p(y)}\,dx\,dy \\ &= D(p_{XY}\|p_X \times p_Y) \end{aligned}$$

where subscripts and arguments indicate the random variable a density describes.

As a special case, we can imagine sending a message θ across a channel $p(x|\theta)$ n times independently. Suppose we permit the n receivers to improve their decoding by pooling the messages $x^n = (x_1, \cdots, x_n)$ they receive. In this context, X is a random variable Θ taking values in $\boldsymbol{R^k}$, and X^n is the vector variable $X^n = (X_1, \cdots, X_n)$. This gives

$$I(\Theta; X^n) = \int p(x^n\,|\,\theta)w(\theta)\log\frac{p(x^n\,|\,\theta)}{m(x^n)}\,dx^n\,d\theta$$

Manuscript received December 20, 1996; revised June 23, 1998.

A. Yuan is with the Statistics Research Laboratory, Department of Anesthesia, Massachussets General Hospital, Clinic #3, Boston, MA 02114 USA.

B. Clarke is with the Department of Statistics, University of British Columbia, Vancouver, BC, Canada V6T 1Z2.

Communicated by P. Moulin, Associate Editor for Nonparametric Estimation, Classification, and Neural Networks.

Publisher Item Identifier S 0018-9448(99)01392-9.

which we recognize as the SMI between the parameter and a sample of size n, X^n. Here, $w(\theta)$ is the density of the source distribution,

$$m_n = m(x^n) = \int w(\theta)p(x^n \mid \theta)\, d\theta$$

is the mixture density, and

$$w(\theta \mid x^n) = p(x^n \mid \theta)w(\theta)/m(x^n)$$

is the posterior density. By Bayes rule, we have

$$I(\Theta; X^n) = E_m D(w(\cdot \mid X^n)||w(\cdot)) = E_w D(p(\cdot \mid \theta)||m_n(\cdot)).$$

So, the SMI is the expected relative entropy between a posterior and the prior that generated it, or the expected relative entropy between an n-fold product of the channel and the marginal density for the received messages. It is seen that maximizing the middle quantity over the source gives the capacity—the SMI for which the source differs most from the conditional distribution of Θ given X^n. Moreover, minimizing the middle quantity over a class of conditional distributions for X^n given Θ gives the rate distortion function—the SMI for which the source differs least from the conditional density of Θ given X^n.

For simplicity, we assume that X and θ are continuous and unidimensional. When either is discrete, it will be enough to replace the integration with a summation. (When either has finite dimension greater than one, the properties we use continue to hold.) Let $L(x, \theta)$ be the distortion from representing θ by x. We minimize the SMI over the class $\mathcal{P}_l$ of conditional densities $p(x|\theta)$ which satisfy

$$\iint p(x \mid \theta)w(\theta)L(x, \theta)\, dx\, d\theta \leq l. \tag{1.1}$$

Here $l > 0$ bounds the expected distortion we will tolerate in representing θ by X. Note that the optimization is over the conditional densities $p(x|\theta)$ directly, not over mixtures such as $m(x)$ and that the integral is over both the source and the output. Now, the minimal value of the the SMI over $\mathcal{P}_l$

$$R(l) = \inf_{p \in \mathcal{P}_l} I(\Theta, X) \tag{1.2}$$

is the rate distortion function (see Berger [3], Blahut [8], and Cover and Thomas [10]). Since the SMI has been used to represent the information in a sample, as discussed in Ibragimov and Hasminsky [18], Bernardo [4], and Efroimovich [15], we refer to the density achieving the minimum in (1.2) as the minimally informative likelihood (MIL). This term is justified by the explanation of the minimization in (1.2) provided in Berger [3, p. 23].

Informally, our main results are three properties of the family of solutions optimizing (1.2). First, we show that the conditional density of the source given the output, $w(\theta|x)$, converges to the source $w(\theta)$ when the distortion is permitted to increase. This means that the class of test channels is so large that the optimal channel is trivial, or equivalently, a large amount of compression has occurred. Second, we show that when the distortion l shrinks to zero, $w(\theta|x)$ degenerates to point mass at x. That is, the optimal channel is perfect. However, no data compression has occurred, so we have represented the source exactly. In this case, the class of test channels is too small. Our third main result shows that in the limit of a large number of cooperative receivers, the expected relative entropy distance between a source w and $w(\theta|x^n)$ tends to zero. Equivalently, the Shannon mutual information goes to zero.

The structure of the rest of this paper is as follows. In the next section, following Blahut [6], we state the solution to the rate-distortion function optimization problem. Using this, we derive the solution for a normal source under squared error loss. After describing the Blahut–Arimoto algorithm (see Blahut [7], Arimoto [2]) we show that a unique solution to the problem exists. In Section III we show how Blahut's solutions, the MIL's, depend on the "distortion" parameter l which determines the size of the class over which the SMI is minimized. There are two cases, one as l tends to zero and the other as l increases. In Section IV, we give two formal senses in which the MIL's can be regarded as minimally informative, but not entirely uninformative. Section V discusses the implications for statistical analysis that follow from the results in Section IV. We comment that the proofs of Theorems 1, 2, and 3 are somewhat technical so we have only described the main steps. Full details are available in Yuan [22].

II. Solving the Optimization Problem

What we have called a minimally informative likelihood, an MIL, is the conditional density which achieves the rate-distortion function lower bound defined in (1.2). The calculus of variations argument for the case of discrete sources is covered in detail in Berger [3, Sec. 2.5]. The reasoning carries over to the continuous case: The problem is the same as described in Berger [3, p. 30] and the variational argument for fixed l can be found in Berger [3, pp. 30–31] and Blahut [6, pp. 58–60, 214–221]. The result of minimizing (1.2) is

$$p^*_\lambda(x \mid \theta) = \frac{m^*(x)e^{-\lambda L(x,\theta)}}{\int m^*(y)e^{-\lambda L(y,\theta)}\, dy} \tag{2.1}$$

where λ and $m^*(x)$ are determined by the equations

$$\iint p^*_\lambda(x \mid \theta)w(\theta)L(x, \theta)\, dx\, d\theta = l \tag{2.2}$$

and

$$\int \frac{e^{-\lambda L(x,\theta)}w(\theta)}{\int m^*(y)e^{-\lambda L(y,\theta)}\, dy}\, d\theta \leq 1 \tag{2.3}$$

with equality in (2.3) for those x such that $m^*(x) > 0$.

For continuous sources w, the minimization to get (2.1) is valid for any source w that has a density that integrates to one, any distortion $L(\theta, x)$ which is positive, continuous in its two arguments and zero when $x = \theta$, and any $l \in (0, \overline{l})$, where $\overline{l}$ is defined before its use in Theorem 3.2. The calculus of variations technique produces (2.1) as the minimum of the SMI subject to the distortion constraint (2.2) and to the constraint that the function of x and θ integrate to one over x. The calculus of variations procedure in effect "differentiates"

with respect to the function $p(\cdot|\cdot)$ evaluated at specific x and θ values (see Blahut [6, proof of Theorem 4, part c, p. 463]. Blahut [7, p. 216] establishes that the minimum from the calculus of variations argument, given in (2.1), is nonnegative, and therefore is a conditional density. (Inequality (2.3) is obtained as a necessary and sufficient condition on m^* to verify that a density of the form (2.1) minimizes (1.2), see Blahut [7, p. 217].) By the uniqueness guaranteed in Proposition 2.1 below, there is no other minimum.

McEliece [21], Blahut [8], and Cover and Thomas [10] provide statements and proofs of many important properties of the rate-distortion function. They also provide closed-form examples for the binomial, the normal, and for finite probability space in general. Here, we focus on the class of solutions identified by Blahut [6] so we verify that Blahut's solutions are reasonable for the case that $w(\cdot)$ is a *Normal*(μ, σ^2) density and L is squared error loss. Cover and Thomas [10, p. 343] have essentially done this for a binary source; their treatment of the normal case did not include this.

From the form of (2.1), one expects that the MIL will be normal. This turns out to be the case subject to the restriction $l < \sigma^2$, i.e., the amount of distortion that can be tolerated must be less than the variance of the source distribution. For $l \geq \sigma^2$, the rate distortion function is zero, see Cover and Thomas [10, p. 344], so no unique solution exists. We see also that $l(\lambda) = 1/(2\lambda)$, so we get that λ must be greater than $1/(2\sigma^2)$. It will be seen that $m^*(\cdot)$ is $N(\mu, \sigma^2 - \frac{1}{2\lambda})$

$$p^*(\cdot\,|\,\theta) = N\left(\frac{\mu + \theta(2\lambda\sigma^2 - 1)}{2\lambda\sigma^2}, \frac{1}{2\lambda}\left(1 - \frac{1}{2\lambda\sigma^2}\right)\right)$$

and $l(\lambda) = \frac{1}{2\lambda}$. Clearly, if $\mu = 0$ then, in the limit as $\lambda\sigma^2$ goes to infinity, θ can be interpreted as the mean. More generally, any interpretation of θ will depend on the prior, and the loss L which determine (2.1).

Note that $m^*(\cdot)$ must satisfy

$$\int \frac{e^{-\lambda(x-\theta)^2} w(\theta)}{\int m^*(y) e^{-\lambda(y-\theta)^2}\, dy}\, d\theta \leq 1 \tag{2.4}$$

with equality in (2.4) for those x with $m^*(x) > 0$. With some foresight, set $m^*(y) = C\exp\{-(ay - b)^2\}$ for some real constants a and b, such that the ratio of $e^{-\frac{(\theta-\mu)^2}{2\sigma^2}}$ and $\int m^*(y) e^{-\lambda(y-\theta)^2}\, dy$ is a constant. Now, the exponent of $m^*(y)e^{-(y-\theta)^2}$ is

$$-[(ay-b)^2 + \lambda(y-\theta)^2] = -[(a^2+\lambda)y^2 - 2(ab+\lambda\theta)y + b^2 + \lambda\theta^2]$$

which is

$$-(a^2+\lambda)\left(y - \frac{ab+\lambda\theta}{a^2+\lambda}\right)^2 - \left[b^2 + \lambda\theta^2 - \frac{(ab+\lambda\theta)^2}{a^2+\lambda}\right].$$

Requiring that

$$\left[b^2 + \lambda\theta^2 - \frac{(ab+\lambda\theta)^2}{a^2+\lambda}\right] = \frac{(\theta-\mu)^2}{2\sigma^2}$$

holds for all θ gives $a^2 = \lambda/(2\lambda\sigma^2 - 1)$ and $b = a\mu$. Thus, we have

$$m^*(x) = \frac{|a|}{\sqrt{\pi}} e^{-(ax-b)^2} = \frac{1}{\sqrt{2\pi\left(\sigma^2 - \frac{1}{2\lambda}\right)}} e^{-\frac{(x-\mu)^2}{2(\sigma^2 - \frac{1}{2\lambda})}}$$

which is a *Normal*$(\mu, \sigma^2 - \frac{1}{2\lambda})$ density. Now, (2.1) gives

$$p^*_\lambda(x\,|\,\theta) = \frac{1}{\sqrt{4\pi(a^2+\lambda)}} e^{-(a^2+\lambda)(x - \frac{ab+\lambda\theta}{a^2+\lambda})^2}.$$

After substituting for a and b, this is the density of a

$$\textit{Normal}\left(\frac{\mu + \theta(2\lambda\sigma^2 - 1)}{2\lambda\sigma^2}, \frac{1}{2\lambda}\left(1 - \frac{1}{2\lambda\sigma^2}\right)\right)$$

and

$$\begin{aligned} l(\lambda) &= \iint p^*_\lambda(x\,|\,\theta) w(\theta) L(x,\theta)\, dx\, d\theta \\ &= \frac{1}{2(a^2+\lambda)} + \left(\frac{a^2}{a^2+\lambda}\right)^2 \sigma^2 = \frac{1}{2\lambda}. \end{aligned}$$

Here $1/\lambda$ or l behaves like a dispersion parameter for $p^*_\lambda(\cdot\,|\,\theta)$ in addition to its role in defining $\mathcal{P}_l$. Also, for fixed θ, μ, λ, as $\sigma^2 \to \infty$, $p^*_\lambda(\cdot\,|\,\theta) \to N(\theta, \frac{1}{2\lambda})$, and hence its variance increases to $\frac{1}{2\lambda} = l(\lambda)$. For fixed θ, μ, σ^2, as $\lambda \to \infty$, $p^*_\lambda(\cdot\,|\,\theta) \to \zeta(\theta)$, the degenerate distribution at θ, consistent with Theorem 1 below. This provides a sense in which λ is also a smoothing parameter, ensuring that an MIL does not just concentrate at the data points.

Note that if one were to form the posterior using a *Normal*(μ, σ^2) prior for θ and the MIL

$$\textit{Normal}\left(\frac{\mu + \theta(2\lambda\sigma^2 - 1)}{2\lambda\sigma^2}, \frac{1}{2\lambda}\left(1 - \frac{1}{2\lambda\sigma^2}\right)\right)$$

for a single outcome, one would find $w(\theta\,|\,x)$ is *Normal*(X, l). That is, the posterior does not depend on μ and depends on σ only because $\sigma^2 > l$.

As suggested by this example, one cannot solve for the optimal $p_{\mathrm{MIL}}(x\,|\,\theta) = p^*_\lambda(x\,|\,\theta)$ explicitly, outside of a few special cases. However, one can obtain $p^*_\lambda(x\,|\,\theta)$ numerically by the Blahut–Arimoto algorithm (see Blahut [7], Arimoto [2]). This is a particular instance of the alternate minimization algorithm whose convergence was established by Csiszár [12] (see Cover and Thomas [10], Csiszár and Tusnady [14]). In the present context, the procedure is as follows. Choose an initial marginal density $m_0(\cdot)$, a prior $w(\cdot)$, and any positive λ. Let

$$p_{1,\lambda}(x\,|\,\theta) = \frac{m_0(x) e^{-\lambda L(x,\theta)}}{\int m_0(y) e^{-\lambda L(y,\theta)}\, dy} \tag{2.5}$$

and form

$$m_1(x) = \int p_{1,\lambda}(x\,|\,\theta) w(\theta)\, d\theta \tag{2.6}$$

from $w(\cdot)$. Next, replace $m_0(\cdot)$ in (2.5) by $m_1(\cdot)$ from (2.6) to form $p_{2,\lambda}(x\,|\,\theta)$. Now one obtains $m_2(\cdot)$ from $p_{2,\lambda}(\cdot\,|\,\theta)$ by mixing out θ. In this fashion one generates a sequence of $p_{n,\lambda}(x\,|\,\theta)$ for a given λ, x, and θ. It follows from Csiszár and Tusnady [14] that as n tends to infinity, $p_{n,\lambda}(x\,|\,\theta)$ converges to $p^*_\lambda(x\,|\,\theta)$. Finally, one can choose λ so that the equality in the constraint (1.1) is satisfied. Indeed, Blahut [6] shows that the minimum in (1.2) is achieved for this λ. That this procedure gives useful results in statistical applications is shown in Yuan and Clarke [23].

Our first result is a proposition guaranteeing that the solutions $p_\lambda(x\,|\,\theta)$ specified in (2.1) and (2.3) to (1.2) are unique.

Proposition 2.1: For each l, $R(l)$ has a unique minimizer $p^*(\cdot\,|\,\cdot)$ in $\mathcal{P}_l$.

Proof: Since $\mathcal{P}_l$ is convex, it is enough to show that $I(\Theta, X)$ is strictly convex on $\mathcal{P}_l$ as a functional of $p(\cdot\,|\,\cdot)$. Write $I(p) = I(\Theta, X)$. Now $\forall\, 0 \le \lambda \le 1$, and $p_1(\cdot\,|\,\cdot), p_2(\cdot\,|\,\cdot) \in \mathcal{P}_l$, with $p_1(\cdot\,|\,\cdot) \neq p_2(\cdot\,|\,\cdot)$, we have $I(\lambda p_1 + (1-\lambda)p_2)$ equals

$$\iint w(\theta)[\lambda p_1(x\,|\,\theta) + (1-\lambda)p_2(x\,|\,\theta)] \times \log \frac{\lambda p_1(x\,|\,\theta) + (1-\lambda)p_2(x\,|\,\theta)}{\lambda m_1(x) + (1-\lambda)m_2(x)}\, d\theta\, dx.$$

By the log-sum inequality, see Cover and Thomas [10], we have that $I(\lambda p_1 + (1-\lambda)p_2)$ is bounded from above by

$$\lambda \iint w(\theta) p_1(x\,|\,\theta) \log \frac{p_1(x\,|\,\theta)}{m_1(x)}\, d\theta\, dx + (1-\lambda) \iint w(\theta) p_2(x\,|\,\theta) \log \frac{p_2(x\,|\,\theta)}{m_2(x)}\, d\theta\, dx = \lambda I(p_1) + (1-\lambda) I(p_2). \qquad \square$$

III. Some Key Properties of Minimally Informative Likelihoods

Clearly, the MIL depends on the choice of l or λ used to define $\mathcal{P}_l$. Next, we prove two theorems that show how the size of $\lambda = \lambda(l)$ in (2.1) affects the behavior of the MIL. We write $p^*_\lambda(x\,|\,\theta)$ for the MIL, $m^*_\lambda(x)$ for the marginal density for the output, and $w^*_\lambda(\theta\,|\,x)$ for the channel. Let $\zeta(\theta)$ denote unit mass at θ, and $\xrightarrow{D}$ denote convergence in distribution. When needed, $\mu(\cdot)$ is Lebesgue measure on R^1. First, we characterize the behavior of the MIL for λ large.

Theorem 3.1:

i) The marginal density for X from $p^*_\lambda(x\,|\,\theta)$ is $m^*_\lambda(x)$, where

$$m^*_\lambda(x) = \int p^*_\lambda(x\,|\,\theta) w(\theta)\, d\theta.$$

Let S be the support of $w(\cdot)$, with interior S^0, and let C be the set of points in S at which w is continuous. Assume $L(x,\theta) = L(|x-\theta|)$ is strictly increasing in $|x-\theta|$, with $L(0) = 0$, and $L(s+t) \ge L(s) + L(t)$, for all $s \ge 0, t \ge 0$. Then as $\lambda \to \infty$, we have the following.

ii) The marginal density for X satisfies

$$m^*_\lambda(x) \to \begin{cases} w(x), & \text{if } x \in S \cap C \\ 0, & \text{for a.e. } \mu(\cdot)\, x \in S^c. \end{cases} \tag{3.1}$$

iii) and the conditional densities satisfy

$$p^*_\lambda(x\,|\,\theta) \xrightarrow{D} \zeta(\theta), \qquad \forall \theta \in S^0 \cap C \tag{3.2}$$

and

$$w^*_\lambda(\theta\,|\,x) \xrightarrow{D} \zeta(x), \qquad \forall x \in S^0 \cap C. \tag{3.3}$$

Remark We do not include a proof of (3.1) here because it is similar to a theorem in Berger [3, p. 103, Theorem 4.3.4, expression 4.3.53]. Our conditions are slightly different because they use the continuity of the solutions (2.1) to get a pointwise limit; Berger [3] obtains (3.1) in absolute mean.

Proof:

i) From (2.1) we see that $p^*_\lambda(x\,|\,\theta) > 0$ implies $m^*_\lambda(x) > 0$. So

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

by (2.3). If $p^*_\lambda(x\,|\,\theta) = 0$, then $m^*_\lambda(x) = 0$, we get the same result.

iii) To prove (3.2), let $\phi_\lambda(\cdot)$ be the characteristic function of $p^*_\lambda(\cdot\,|\,\theta)$. We have the expression at the bottom of this page.

For $\theta \in S^0$ and appropriate choices of δ and $a > 0$, we can use Step 4 [22, Proof of Theorem 3.3.1] to get

$$\frac{\int_{[\theta-\delta,\theta+\delta]^c} m^*_\lambda(y) e^{-\lambda L(y,\theta)}\, dy}{\int_{[\theta-\delta,\theta+\delta]} m^*_\lambda(y) e^{-\lambda L(y,\theta)}\, dy} \le \frac{2e^{-\lambda[L(\delta)-L(\delta/2)]}}{a\delta} \to 0, \qquad \text{as } \lambda \to \infty.$$

This can be used to show that

$$\phi_\lambda(t) = \frac{\int_{[\theta-\delta,\theta+\delta]} m^*_\lambda(x) e^{-\lambda L(x,\theta)} \cos(xt)\, dx}{\int_{[\theta-\delta,\theta+\delta]} m^*_\lambda(y) e^{-\lambda L(y,\theta)}\, dy} + i \frac{\int_{[\theta-\delta,\theta+\delta]} m^*_\lambda(x) e^{-\lambda L(x,\theta)} \sin(xt)\, dx}{\int_{[\theta-\delta,\theta+\delta]} m^*_\lambda(y) e^{-\lambda L(y,\theta)}\, dy} + o(1).$$

Write the first term as $J_1(\lambda, t)$ and the second term as $iJ_2(\lambda, t)$. It can be shown that

$$\lim_{\delta \to 0} J_1(\lambda, t) = \cos(\theta t)$$

and

$$\lim_{\delta \to 0} J_2(\lambda, t) = \sin(\theta t)$$

holds for all λ. Thus if we first let $\delta \to 0$ and then let $\lambda \to \infty$, we get $\lim_{\lambda\to\infty} \phi_\lambda(t) = e^{i\theta t}$, the characteristic function of $\zeta(\theta)$.

To prove (3.3), let $\psi_\lambda(\cdot)$ be the characteristic function of $w^*_\lambda(\cdot\,|\,x)$. Now, $\psi_\lambda(\cdot)$ is (see the second expression at the top of the following page). As in Step 1 [22, Proof of Theorem 3.3.1], the second term tends to zero as λ tends to infinity. The same technique of proof used to obtain (3.1) gives

$$\psi_\lambda(t) = \frac{w(\zeta) e^{i\zeta t}}{m^*_\lambda(\eta)} \left(\frac{\int_0^\delta e^{-\lambda L(t)}\, dt}{\int_0^\delta e^{-\lambda L(s)}\, ds(1+o(1))} + o(1) \right) + o(1)$$

where $\zeta \in [x-\delta, x+\delta], \eta \in [\theta-\delta, \theta+\delta] \subset [x-2\delta, x+2\delta]$. By the same reasoning as in Step 4 [22, Proof of Theorem 3.3.1], ζ and η tend to x as δ tends to zero. Therefore, the ratio $w(\zeta)/m^*_\lambda(\eta)$ tends to 1 as in (3.1). Thus we have

$$\phi_\lambda(t) = \frac{\int_{[\theta-\delta,\theta+\delta]} m^*_\lambda(x) e^{-\lambda L(x,\theta)} e^{ixt}\, dx + \int_{[\theta-\delta,\theta+\delta]^c} m^*_\lambda(x) e^{-\lambda L(x,\theta)} e^{ixt}\, dx}{\int_{[\theta-\delta,\theta+\delta]} m^*_\lambda(y) e^{-\lambda L(y,\theta)}\, dy + \int_{[\theta-\delta,\theta+\delta]^c} m^*_\lambda(y) e^{-\lambda L(y,\theta)}\, dy}.$$

$$\int_{[x-\delta,x+\delta]} \frac{w(\theta)e^{-\lambda L(x,\theta)}e^{i\theta t}}{\int_{[\theta-\delta,\theta+\delta]} m_\lambda^*(y)e^{-\lambda L(y,\theta)}\,dy + \int_{[\theta-\delta,\theta+\delta]^c} m_\lambda^*(y)e^{-\lambda L(y,\theta)}\,dy}\,d\theta + \int_{[x-\delta,x+\delta]^c} \frac{w(\theta)e^{-\lambda L(x,\theta)}e^{i\theta t}}{\int m_\lambda^*(y)e^{-\lambda L(y,\theta)}\,dy}\,d\theta.$$

$\lim_{\lambda\to\infty}\psi_\lambda(t) = e^{ixt}$, a.e. with respect to $\mu(\cdot)$ for $x \in S$ and part iii) is proved. □

As noted in the Introduction, Theorem 3.1 shows that as λ goes to infinity, the observation channel becomes totally informative in the sense of converging to point masses. Our next result confirms that λ increases as l decreases and the reverse, and then guarantees that in the limit as λ goes to zero, the observation channel becomes totally uninformative. More precisely, let

$$\bar{l} = \inf_x \int w(\theta)L(x,\theta)\,d\theta$$

and set

$$x_0 = \arg\inf_x \int w(\theta)L(x,\theta)\,d\theta.$$

It will be seen that when $l > \bar{l}$, the method breaks down, because there is no necessary relationship between the data and the estimand θ.

Theorem 3.2: Assume $L(\cdot,\theta)$ is not constant. Then we have

i) For $l \in (0,\bar{l}), p_\lambda^*(x\,|\,\theta)$ exists uniquely

$$\inf_{p(\cdot\,|\,\theta)\in\mathcal{P}_l} I_p(\Theta,X) = I_{p_\lambda^*}(\Theta,X) > 0$$

and is a continuous, decreasing function of l.

ii) For $l \in (0,\bar{l}), \lambda$ and l determine each other uniquely. We can therefore write $l = l(\lambda)$, or $\lambda = \lambda(l)$.

iii) For $l \in [\bar{l},\infty]$

$$\inf_{p(\cdot\,|\,\theta)\in\mathcal{P}_l} I_p(\Theta,X) = 0$$

and the infimum is achieved by any $p(x) \in \mathcal{P}_l$ which is independent of θ.

iv) Assume

$$D(m_{\lambda_2}^*\|m_{\lambda_1}^*) + D(m_{\lambda_1}^*\|m_{\lambda_2}^*) < \infty, \qquad \text{for } 0 < \lambda_1 \le \lambda_2$$

then $l(\lambda_2) \le l(\lambda_1)$, i.e., $l(\cdot)$ is a decreasing function.

Under conditions of Theorem 3, we have the following:

v) $l(\lambda) \to 0$, as $\lambda \to \infty$.

vi) $l(\lambda) \to \bar{l}$, as $\lambda \to 0$.

vii) Let $P_\lambda^*(\cdot\,|\,\theta), M_\lambda^*(\cdot), W(\cdot)$ and $W_\lambda^*(\cdot\,|\,x)$ be the probability measures corresponding to $p_\lambda^*(x\,|\,\theta), m_\lambda^*(x), w(\theta)$ and $w_\lambda^*(\theta\,|\,x)$, respectively. If $M_\lambda^*(\cdot) \to M_0(\cdot)$ in distribution for some probability measure $M_0(\cdot)$ as $l \to \bar{l}$ (or $\lambda \to 0$), then

$$P_\lambda^*(\cdot\,|\,\theta) \xrightarrow{D} M_0(\cdot) \tag{3.4}$$

$$W_\lambda^*(\cdot\,|\,x) \xrightarrow{D} W(\cdot). \tag{3.5}$$

viii) Under conditions of (vii), if $\bar{l} < \infty$, then $M_0(\cdot) = \zeta(x_0)$.

Proof:

i) This follows from Proposition 1 and Blahut [8, Theorem 6.3.2].

ii) By (2.2), l is determined uniquely by λ. On the other hand, for $l \in (0,\bar{l})$, we know that there is a unique $p_\lambda^*(\cdot\,|\,\theta)$. By way of contradiction, if there is a $\lambda' \ne \lambda$ such that

$$p_{\lambda'}^*(\cdot\,|\,\theta) = p_\lambda^*(\cdot\,|\,\theta)$$

then by i) of Theorem 3.1, $m_{\lambda'}^*(\cdot) = m_\lambda^*(\cdot)$. Using (2.1) gives that $\exp -(\lambda' - \lambda)L(x,\theta)$ is constant as a function of x, which is impossible.

iii) In [8, Proof of Theorem 6.3.2] it was shown that the rate distortion function is zero for $l = \bar{l}$. By i), it is decreasing in l. Clearly, if $p(\cdot)$ is independent of θ, then $I_p(\Theta,X) = 0$.

iv) Consider $p_\lambda^*(x\,|\,\theta)$ for two values λ_1 and λ_2. For fixed θ

$$\begin{aligned} D(p_{\lambda_2}^*\|p_{\lambda_1}^*)(\theta) &= \log\left(\frac{\int m_{\lambda_1}^*(y)e^{-\lambda_1 L(y,\theta)}\,dy}{\int m_{\lambda_2}^*(y)e^{-\lambda_2 L(y,\theta)}\,dy}\right) \\ &\quad + (\lambda_1 - \lambda_2)\int L(x,\theta)p_{\lambda_2}^*(x\,|\,\theta)\,dx \\ &\quad + \int p_{\lambda_2}^*(x\,|\,\theta)\log\frac{m_{\lambda_2}^*(x)}{m_{\lambda_1}^*(x)}\,dx. \end{aligned}$$

Adding this to the corresponding expression with λ_1 and λ_2 interchanged gives

$$\begin{aligned} &D(p_{\lambda_2}^*\|p_{\lambda_1}^*)(\theta) + D(p_{\lambda_1}^*\|p_{\lambda_2}^*)(\theta) \\ &= (\lambda_1 - \lambda_2)\left(\int L(x,\theta)p_{\lambda_2}^*(x\,|\,\theta)\,dx \right. \\ &\quad \left. - \int L(x,\theta)p_{\lambda_1}^*(x\,|\,\theta)\,dx\right) + \int p_{\lambda_2}^*(x\,|\,\theta)\log\frac{m_{\lambda_2}^*(x)}{m_{\lambda_1}^*(x)}\,dx \\ &\quad + \int p_{\lambda_1}^*(x\,|\,\theta)\log\frac{m_{\lambda_1}^*(x)}{m_{\lambda_2}^*(x)}\,dx. \end{aligned} \tag{3.6}$$

Averaging over θ in (3.6) and using the technique of Proposition 2.1 finishes the proof.

v) Using (2.2) it is enough to show

$$\overline{\lim_{\lambda\to\infty}} \int p_\lambda^*(x\,|\,\theta)L(x,\theta)\,dx = 0, \qquad \forall\theta \in S^0 \cap C. \tag{3.7}$$

A continuity argument on x and θ satisfying $|x-\theta| \le \delta$ bounds part of the integral in (3.7). To bound the other part, shrink the domain of integration in the denominator to an interval around θ. Then, bound the loss function on the reduced domains of integration and bound the result by use of

$$b = \inf_{y\in[\theta-\delta',\theta+\delta']} w(y)$$

and

$$B_\lambda = \left\{y \in [\theta-\delta',\theta+\delta']\,|\,m_\lambda^*(y) \ge \frac{b}{2}\right\}.$$

vi) Let $\lambda(0) = \lim_{l \to l} \lambda(l)$, then from iii) we know that $I_{p^*_{\lambda(0)}}(\Theta, X) = 0$, and so $p^*_{\lambda(0)}$ is independent of θ. That is, $\lambda(0) = 0$ or it cannot be independent of θ.

vii) For (3.4), it is enough to prove that for all $A \in \mathcal{B}$, the Borel algebra on R^1, as $\lambda \to 0$ $P^*_\lambda(A \,|\, \theta) \to M_0(A)$. Let $\epsilon > 0$ and choose a, b, λ_0 to ensure that for λ_o, the integral of (2.1) over A can be bounded by

$$P^*_\lambda(A \,|\, \theta) \leq \frac{M^*_\lambda(A)}{\int_{[a,b]} e^{-\lambda_0 L(y,\theta)} M^*_\lambda(dy)}. \tag{3.8}$$

Since $e^{-\lambda_0 L(y,\theta)}$ is bounded and continuous on $[a, b]$, and

$$M^*_\lambda(\cdot) \xrightarrow{D} M_0(\cdot)$$

(3.8) is bounded from above by $(M_0(A) + \epsilon)/(1 - \epsilon)$.

For a lower bound, integrating (2.1) over A and bounding the integral in the denominator by 1, we get

$$P^*_\lambda(A \,|\, \theta) \geq \int_{A \cap [a,b]} e^{-\lambda_0 L(x,\theta)} M^*_\lambda(dx)$$

for appropriately chosen a, b, and λ_0. So, for λ small, we have $P^*_\lambda(A \,|\, \theta) \geq M_0(A) - \epsilon$, establishing (3.4).

To get (3.5) for λ small, integrate the posterior from the proof of (3.3) over a set A and choose a, b, and λ_0 so that

$$W^*_\lambda(A \,|\, x) \leq \int_A \frac{W(d\theta)}{\int_{[a,b]} e^{-\lambda_0 L(y,\theta)} M^*_\lambda(dy)} \leq \frac{W(A)}{(1 - \epsilon)}. \tag{3.9}$$

For the inequality the other way, for small λ, choose a, b,, and λ_o small enough that

$$W^*_\lambda(A \,|\, x) \geq \int_{A \cap [a,b]} e^{-\lambda_0 L(x,\theta)} W(d\theta) \geq W(A) - \epsilon.$$

viii) By way of contradiction, suppose $M_0(\cdot) \neq \zeta(x_0)$. We can choose a, b, c, and d independent of λ so that by (2.1) we have

$$l(\lambda) \geq \int_{[a,b]} \int_{[c,d]} L(x,\theta) P^*_\lambda(dx \,|\, \theta) W(d\theta)$$

and the limit of the right-hand side, as $\lambda \to 0$, is

$$\int_{[a,b]} \int_{[c,d]} L(x,\theta) M_0(dx) W(d\theta) \geq \iint L(x,\theta) M_0(dx) W(d\theta) - \epsilon_1. \tag{3.10}$$

Now, since $M_0(\cdot) \neq \zeta(x_0)$, the right-hand side of (3.11) is strictly greater than

$$\inf_x \int L(x,\theta) W(d\theta) + \epsilon_1 = \bar{l} + \epsilon_2$$

for some $\epsilon_2 > 0$. By (3.4), the continuity of $L(x,\theta)$ and $l(\lambda) \to \bar{l}$ as $\lambda \to 0$, we get

$$\lim_{\lambda \to 0} l(\lambda) \geq \lim_{\lambda \to 0} \int_{[a,b]} \int_{[c,d]} L(x,\theta) P^*_\lambda(dx \,|\, \theta) W(d\theta) = \int_{[a,b]} \int_{[c,d]} L(x,\theta) M_0(dx) W(d\theta) > \bar{l} + \epsilon_2$$

a contradiction, so (3.5) follows. □

When it exists, let

$$x_0 = \arg\inf_x \int w(\theta) L(x,\theta)\, d\theta.$$

If $L(x,\theta) = (x - \theta)^2$, for example, then $w(\cdot)$ must have a finite second moment for x_0 to exist. In particular, if the prior $w(\cdot)$ is $N(\mu, \sigma^2)$, then

$$\int w(\theta) L(x,\theta)\, d\theta = \sigma^2 + (\mu - x)^2$$

so

$$\inf_x \int w(\theta) L(x,\theta)\, d\theta = \sigma^2$$

and the infimum is achieved at $x_0 = \mu$. If $w(\cdot)$ is *Exponential*(α, μ), where μ is the location parameter, then

$$\int w(\theta) L(x,\theta)\, d\theta = (x - \mu - 1/\alpha)^2 + 1/\alpha^2$$

and

$$\inf_x \int w(\theta) L(x,\theta) d\theta = 1/\alpha^2$$

and the infimum is achieved at $x_0 = \mu + 1/\alpha$. This ensures that Theorem 3.2 can be used, in some cases, to identify the limits of MIL's in practice.

IV. Two Formal Senses in Which the Conditional Density Is Minimally Informative

In this section, we provide two senses in which the conditional density which achieves the rate distortion function lower bound can be regarded as minimally informative. Suppose that we replace X from the last section with a multivariate random variable X^n and consider

$$\mathcal{P}_n = \left\{ p_n(x^n \,|\, \theta) : \iint p_n(x^n \,|\, \theta) w(\theta) L_n(x^n, \theta)\, dx^n\, d\theta \leq l_n \right\}. \tag{4.1}$$

Denote the MIL for X^n by $p_{\mathrm{MIL}}(x^n \,|\, \theta)$, that is, write

$$p_{\mathrm{MIL}}(x^n \,|\, \theta) = \arg\min_{p \in \mathcal{P}_n} I(\Theta, X^n).$$

Similar to the univariate case handled in Blahut [6], one can obtain a form for the MIL based on the loss function L. For given prior w, this is

$$p_{\mathrm{MIL}}(x^n \,|\, \theta) = \frac{m^*_n(x^n) e^{-\lambda_n L_n(x^n,\theta)}}{\int m^*_n(y^n) e^{-\lambda_n L_n(y^n,\theta)}\, dy^n} \tag{4.2}$$

where $m^*_n(x^n)$ is determined by

$$\int \frac{e^{-\lambda_n L_n(x^n,\theta)} w(\theta)}{\int m^*_n(y^n) e^{-\lambda_n L_n(y^n,\theta)} dy^n}\, d\theta \leq 1 \tag{4.3}$$

with equality for x^n's such that $m^*_n(x^n) > 0$, and $\lambda_n \geq 0$ is determined by l_n. We will see that a posterior formed from the parametric family (4.2) and the source w is asymptotically the same as w in a relative entropy sense. That is, the data update w trivially. In addition, we will see that use of the

MIL, $p^*_\lambda(x\,|\,\theta)$, gives the weakest inferences possible among the elements of $\mathcal{P}_n$. Note that this differs from the results in Section III which assumed a univariate X.

Our first result in this section is the asymptotic equivalence of w and the posterior based on w and (4.2). Note that $p_{\mathrm{MIL}}(x^n\,|\,\theta)$ has a dependence structure determined in part by n and λ_n, and that p_{MIL} typically cannot be given in closed form. To prove results about the MIL we use an independence density $p_n(\cdot\,|\,\theta)$ in $\mathcal{P}_n$. The expected relative entropy between the posterior based on this density and w and the prior is bounded by the relative entropy between the posterior based on the MIL and w and the prior. We prove the first expected relative entropy tends to zero as n goes to infinity.

Our result is for the case that the distortion is

$$L_n(x^n,\theta) = a_n \sum_{i=1}^{n} L(x_i,\theta)$$

for given $L(\cdot,\cdot)$. For convenience, we absorb the l_n into a_n in $\mathcal{P}_n$ and assume $l_n = 1$, for all n. The average loss for fixed x is now

$$r(x) = \int w(\theta)L(x,\theta)\,d\theta.$$

Its supremum and infimum are

$$\underline{r} = \inf_x r(x) \qquad \bar{r} = \sup_x r(x).$$

Theorem 4.1: Assume that $r(x)$ is bounded, that $L(\cdot,\cdot)$ is continuous in both arguments, and that $\lim na_n$ exists and is s. Now, if $\underline{r}s < 1$ we have

$$E_{m_{p_n^*}} D\big(w_{p_n^*}(\cdot\,|\,X^n)\|w(\cdot)\big) \to 0.$$

Remark: This result contrasts with the fact that for independent and identically distributed (i.i.d.) data distributed according to a density $p(x\,|\,\theta)$ where θ is a d-dimensional parameter, we have $I(\Theta,X^n) = (d/2)\ln n + o(1)$.

Proof:

Step 1: First we prove that there exists a probability density $q(\cdot)$ such that the new parametric family p_n for X^n defined by

$$p_n(x^n\,|\,\theta) = \frac{e^{-L_n(x^n,\theta)}\prod_{i=1}^{n} q(x_i)}{\int e^{-L_n(y^n,\theta)}\prod_{i=1}^{n} q(y_i)\,dy^n} \tag{4.4}$$

is an element of $\mathcal{P}_n$ for n large enough. The main steps are as follows. Choose a constant $b \in (\underline{r},\bar{r})$ such that $bs < 1$ and a probability density $q(\cdot)$ so that $\int q(x)L(x,\theta)\,dx < \infty$ for all θ and $\int w(\theta)q(x)L(x,\theta)\,dx\,d\theta = b$. Now, $p_n(x^n\,|\,\theta)$ satisfies

$$\iint p_n(x^n\,|\,\theta)w(\theta)L_n(x^n,\theta)\,dx^n\,d\theta = na_n \iint \frac{w(\theta)q(x)e^{-a_nL(x,\theta)}L(x,\theta)}{\int q(y)e^{-a_nL(y,\theta)}\,dy}\,dx\,d\theta. \tag{4.5}$$

Denote the double integral in (4.5) by $I(a_n)$. Since $na_n \to s$, and $sb < 1$, to see that $p_n(x^n\,|\,\theta) \in \mathcal{P}_n$ for all large n, it is enough to show $I(a_n) \to b$. By a sequence of standard inequalities we can derive that

$$|I(a_n) - b| \le 2\int w(\theta)q(x)L(x,\theta)\,dx\,d\theta = 2b < \infty.$$

Finally, a slightly involved argument gives that $p_n(x^n\,|\,\theta) \in \mathcal{P}_n$ for all large n. This argument uses the Dominated Convergence Theorem three times; once pointwise in θ and then two more times for the numerator and denominator of the resulting bound.

Step 2: We prove the assertion of the theorem. Let $q(x^n) = \prod_{i=1}^{n} q(x_i)$ be the density used in Step 1 to define $p_n(x^n\,|\,\theta)$ and write

$$m_{p_n}(x^n) = \int p_n(x^n\,|\,\theta)w(\theta)\,d\theta.$$

Now, the nonnegative quantity $E_{m_{p_n^*}} D(w_{p_n^*}(\cdot\,|\,x^n)\|w(\cdot))$ is bounded from above by

$$-\iint w(\theta)p_n(x^n\,|\,\theta)L_n(x^n,\theta)\,dx^n\,d\theta \tag{4.6}$$

$$-\int w(\theta)\log\left(\int q(y^n)e^{-L_n(y^n,\theta)}\,dy^n\right)d\theta \tag{4.7}$$

$$-\iint w(\theta)p_n(x^n\,|\,\theta) \times \log\left(\int \frac{e^{-L_n(x^n,\xi)}w(\xi)}{\int q(y^n)e^{-L_n(y^n,\xi)}\,dy^n}\,d\xi\right)dx^n\,d\theta. \tag{4.8}$$

The term (4.6) is $-na_nI(a_n) \to -sb$. It is enough to show (4.7) $\to sb$ and (4.8) $\to 0$. For (4.7), since $-\log(\cdot)$ is convex, we have

$$0 < -\log\left(\int q(y^n)e^{-L_n(y^n,\theta)}\,dy^n\right) \le na_n\int q(y)L(y,\theta)\,dy < \infty \tag{4.9}$$

for any θ and n. Denoting the integral on the right of (4.9) by $a(\theta)$ we get

$$\int a(\theta)W\,(d\theta) = b$$

and we have $L_n(Y^n,\theta) \to sa(\theta)$, almost surely with respect to q for any θ. Now, a convergence in probability argument gives

$$\int q(y^n)e^{-L_n(y^n,\theta)}\,dy^n \to e^{-sa(\theta)}$$

for each θ. Taking logs, one sees that the Dominated Convergence Theorem implies that (4.6) and (4.7) will cancel.

Finally, (4.8) is nonpositive: Regard (4.8) as an expectation with respect to $q(x^n)$ rather than $p_n(\cdot\,|\,\theta)$ and use the inequality $-\log x \le (1/x) - 1$. □

Next, we turn to a nonasymptotic sense in which the MIL is minimally informative. Following Csiszár [13], the tangent

hyperplane determined by $w(\theta)$ and $w_{p_n^*}(\theta \mid x^n)$ is

$$H(x^n, w, w_{p_n^*}) = \Big\{ w' : \int w'(\theta) \log \frac{w_{p_n^*}(\theta \mid x^n)}{w(\theta)}\, d\theta = D\big(w_{p_n^*}(\cdot \mid x^n) \| w(\cdot)\big) \Big\}.$$

Let $p \in \mathcal{P}_n$ be another density. The tangent hyperplane determined by $w(\theta)$ and $w_p(\theta \mid x^n)$ is

$$H(x^n, w, w_p) = \Big\{ w' : \int w'(\theta) \log \frac{w_p(\theta \mid x^n)}{w(\theta)}\, d\theta = D(w_p(\cdot \mid x^n) \| w(\cdot)) \Big\}.$$

These two tangent hyperplanes divide the whole space of priors into subspaces. One of them is $S(x^n, w, w_{p_n^*}, w_p)$, defined to be

$$\Big\{ w' : \int w'(\theta) \log \frac{w_{p_n^*}(\theta \mid x^n)}{w(\theta)}\, d\theta \le D\big(w_{p_n^*}(\cdot \mid x^n) \| w(\cdot)\big), \int w'(\theta) \log \frac{w_p(\theta \mid x^n)}{w(\theta)}\, d\theta \ge D(w_p(\cdot \mid x^n) \| w(\cdot)) \Big\}.$$

Let $S_n(w, w_{p_n^*}, w_p) = \cap_{x^n} S(x^n, w, w_{p_n^*}, w_p)$ and let w_0 be a member of $S_n(w, w_{p_n^*}, w_p)$. We show that the MIL updates $w(\cdot)$ to a posterior $w_{p^*}(\theta \mid x^n)$ further from any target density w_0 in Kullback–Leibler distance than any other member $p(x^n \mid \theta)$ in $\mathcal{P}_n$ does. This means that the MIL requires more data than any other member of $\mathcal{P}_n$ does to achieve the same accuracy of estimation. To get a result for individual x^n's, let

$$U\big(w, w_{p_n^*}, w_p\big) = \big\{ x^n : D\big(w_{p_n^*}(\cdot \mid x^n) \| w(\cdot)\big) \le D(w_p(\cdot \mid x^n) \| w(\cdot)) \big\}.$$

Since

$$E_{m_{p_n^*}} D(w_{p_n^*}(\cdot \mid x^n) \| w(\cdot)) \le E_{m_{p_n}} D(w_p(\cdot \mid x^n) \| w(\cdot))$$

it is likely that for some x^n, $U(w, w_{p_n^*}, w_p) \neq \phi$.

Theorem 4.2:

i) If

$$x^n \in U(w, w_{p_n^*}, w_p)$$

and

$$w_0 \in S(x^n, w, w_{p_n^*}, w_p)$$

then

$$D\big(w_0(\cdot) \| w_{p_n^*}(\cdot \mid x^n)\big) \ge D(w_0(\cdot) \| w_p(\cdot \mid x^n)).$$

ii) If for some n, $w_0 \in S_n(w, w_{p_n^*}, w_p)$, then

$$E_{m_{p_n^*}} D\big(w_0(\cdot) \| w_{p_n^*}(\cdot \mid X^n)\big) \ge E_{m_p} D(w_0(\cdot) \| w_p(\cdot \mid X^n)).$$

Proof:

i) We have

$$\int w'(\theta) \log \frac{w_{p_n^*}(\theta \mid x^n)}{w(\theta)}\, d\theta = D(w'(\cdot) \| w(\cdot)) - D(w'(\cdot) \| w_{p_n^*}(\cdot \mid x^n))$$

and the corresponding expression with $w_p(\theta \mid x^n)$ in place of $w_{p_n^*}(\theta \mid x^n)$. This implies

$$D(w_0(\cdot) \| w(\cdot)) \le D\big(w_0(\cdot) \| w_{p_n^*}(\cdot \mid x^n)\big) + D\big(w_{p_n^*}(\cdot \mid x^n) \| w(\cdot)\big)$$

and the corresponding (reversed) inequality with $w_p(\theta \mid x^n)$ in place of $w_{p_n^*}(\theta \mid x^n)$. Taken together these two inequalities give i).

ii) Since $w_0 \in S_n(w, w_{p_n^*}, w_p)$, we have that

$$D(w_0(\cdot) \| w(\cdot)) \le D(w_0(\cdot) \| w_{p_n^*}(\cdot \mid X^n)) + D\big(w_{p_n^*}(\cdot \mid X^n) \| w(\cdot)\big).$$

We also have the reverse inequality for $w_p(\theta \mid X^n)$ in place of $w_{p_n^*}(\theta \mid X^n)$. Taking expectations with respect to $m_{p_n^*}$ and m_{p_n}, respectively, gives two inequalities which, taken together, give ii) in view of the definition of the MIL. □

V. Discussion

We have studied the solutions to the problem of minimizing the Shannon Mutual Information (SMI) over a class of test channels defined by a distortion constraint. That is, we have studied the class of conditional densities which achieve the rate-distortion function lower bound. The solutions are parametric families that depend on a distortion function, an allowable distortion, and the source distribution itself. While the rate-distortion function quantifies the amount of data compression that can be optimally achieved, the conditional density that achieves the rate-distortion function lower bound has the weakest dependence between the source and the output within the class of conditional densities over which we have minimized.

Our main results are properties of these conditional densities. For a fixed source and distortion function, we have shown that as the distortion permitted increases, the channel becomes ever less informative. By contrast, when the distortion permitted decreases, the channel becomes ever more informative. The limits of totally informative and totally uninformative can be characterized as trivial: The first is just the source itself; the second is a limiting distribution independent of the source.

Our third main result is that the optimal channels are minimally informative in an asymptotic sense. If we minimize the SMI between a source and n i.i.d. random variables that depend on it, we get an n-variate density. If we take the limit as n increases of the SMI between the source and an n variate random variable distributed according to the optimal n variate density, we find this limit to be zero. By Bayes rule, this SMI is the expected relative entropy between the source and the conditional distribution of the source given the random variable. That is, the n variate random variable contains ever less information about the source as n increases, i.e., as the

number of receivers increases. By contrast, the rate-distortion function has a fixed number of receivers and minimizes over replications of the data compression procedure.

An alternative way to view the procedure proposed here is that it is a variant on maximum entropy. Indeed, the SMI $I(\Theta, X)$ is the difference between an entropy $H(\Theta)$ and a conditional entropy $H(\theta \mid X)$. Minimizing the SMI with a fixed distribution for Θ over the class of conditional densities used for the rate-distortion function is equivalent to maximizing a conditional entropy over the same class. In addition, this class of densities has a statistical meaning. It is the class of likelihoods for which the Bayes risk of estimating the parameter θ by the random variable X itself is bounded by a real number λ.

More generally, information-theoretic techniques have been used in various statistical contexts. The SMI has been used, for instance, by Bernardo [4]. He proposed using the capacity achieving source for a channel as a minimally informative prior when one uses the channel as a parametric family for data analysis. Here we have essentially reversed this. In contrast to Bernardo's reasoning, we have minimized the SMI over parametric families for a fixed prior: This provides optimal data compression and produces the conditional density that gives a posterior as close as possible to the prior. Conditional densities giving posteriors close to their priors are uninformative, and because they achieve the rate-distortion function lower bound we have called them minimally informative likelihoods, MIL's. Theorems 4.1 and 4.2 verify the validity of this interpretation.

Although this is initially counterintuitive, in fact it reflects the reality that usually a statistician is unable to formulate a physically plausible likelihood. An MIL in effect minimizes the strength of assumptions that go into the choice of likelihood even though this minimization is over likelihoods that have bounded Bayes risk, and so cannot be too pathological. Unlike other methods for likelihood selection, this method directly optimizes over possible relationships between outcome values and parameter values. This relationship is fundamental because one can only make inferences about the parameter from the data if the parameter and data are dependent. Indeed, the SMI between two random variables is zero if they are independent. The hope would be that using an MIL, or rate-distortion function achieving conditional density, would be conservative in the sense that even though it is not right, an investigator could use it to make weak but valid and useful inferences, when no other likelihood is available. The confidence intervals, while consistent, might be wider than optimal; rejecting a null hypothesis using an MIL would correspond to a better significance level under the true likelihood.

An earlier effort to find a minimally informative likelihood is due to Huber [16], [17]. For simplicity, assume the data follow a distribution of the form $F(x-\theta)$ where θ is unknown and F is a member of a class $\mathcal{P}$. Then, one can search $\mathcal{P}$ for the distribution function F_o with the smallest Fisher information. Huber [17] states that asymptotically efficient M-estimators for θ obtained by using F_o have minimax properties with respect to $\mathcal{P}$. Related problems were solved by Levit [20] and Bickel and Collins [5]. The minimax approach taken in these references maximizes over parameters and then minimizes over estimators. Here, decision theory chiefly enters in defining the set of likelihoods over which one will minimize the SMI. This set is defined to be those parametric families for which the random variable X itself as an estimator of the parameter θ has Bayes risk under the chosen prior bounded by l. Recalling that the SMI is itself a minimum over density estimators, see [1], the procedure here amounts to minimizing over likelihoods after having minimized over estimators.

In the context of the normal example of Section II, using an MIL amounts to a recommendation that if one truly believes a normal prior is appropriate and one knows little else, the only statistical option is to report an inflated posterior variance. Although not satisfying, this result is consistent with what one expects in the normal case with squared-error loss. If one chooses a different prior, one gets other likelihoods that do not admit closed-form expressions. More generally, one can use products of univariate MIL's to achieve better inferences; this would be minimally informative apart from the assumption of independence. Note that in this formulation, the interpretation of the parameter θ arises from L and the optimization; θ is a location parameter in general only in the sense that it can be estimated by X with Bayes risk bounded by λ. This reverses the role of estimator and likelihood, because it is the estimator that is fixed, and we are seeking parametric families for which it is good.

Once one has a prior and the MIL, one can construct credibility sets or find posterior probabilities for Bayesian hypothesis testing. Alternatively, one can use the MIL to find a maximum-likelihood estimator and form confidence intervals even though the optimality properties of the MIL are Bayesian. Indeed, because our perspective is information-theoretic, the technique is generally applicable. It can be used with most summary statistics, interacts well with parameter transformations, and can be used with dependent data. The major drawback at the present time is that coding algorithms to obtain the minimally informative likelihood, or posteriors or other estimators can be considerable.

Finally, an unexpected benefit of MIL's is that they permit a comprehensive robustness analysis. In addition to being able to assess the sensitivity of inferences to choice of loss, λ and prior, automating the choice of likelihood permits one to assess sensitivity to modeling strategy. Namely, use of MIL's permits one to compare the effects of using different summary statistics and different numbers of parameters. For instance, one might have an independent sequence of paired data and want to estimate the difference in the locations. One can marginalize a bivariate likelihood to get a model for the difference in each pair. This gives a univariate posterior for a single parameter generating credibility sets for the difference in means in one sense. Alternatively, one can condition on all the data in a two-parameter likelihood to get a bivariate posterior. Now, one can marginalize in the posterior to get a credibility set for the difference in means in a different sense. One expects that taking differences in the data will be similar to taking differences in the parameters, but it is not clear that the two modeling strategies will always give compatible inferences.

REFERENCES

[1] J. Aitchison, "Goodness of prediction fit," *Biometrika*, vol. 62, pp. 547–554, 1975.

[2] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 14–20, Jan. 1972.

[3] T. Berger, *Rate Distortion Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1971.

[4] J. M. Bernardo, "Reference posterior distribution for Bayesian inference," *J. Roy. Statist. Soc., Ser. B*, no. 2, pp. 113–147, 1979.

[5] P. J. Bickel and J. R. Collins, "Minimizing Fisher information over mixtures of distributions" *Sankhya, Ser. A*, vol. 45, pp. 1–19, 1983.

[6] R. E. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 460–473, July 1972.

[7] ———, "A hypothesis testing approach to information theory," Ph.D. dissertation, Cornell Univ., Ithaca, NY, 1972.

[8] ———, *Principles and Practice of Information Theory*. Reading, MA: Addison-Wesley, 1987.

[9] B. Clarke and A. R. Barron, "Jeffreys' prior is asymptotically least favorable under entropy risk," *J. Statist. Planning Infer.*, vol. 41, pp. 37–60, 1994.

[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[11] I. Csiszár, "On topological properties of f-divergences," *Studia Sci. Math. Hungar.*, vol. 2, pp. 329–339, 1967.

[12] ———, "On the computation of rate distortion functions," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 122–124, 1974.

[13] ———, "I-divergence geometry of probability distributions and minimization problems," *Ann. Prob.*, vol. 3, no. 1, pp. 146–158, 1975.

[14] I. Csiszár and G. Tusnady, "Information geometry and alternating minimization procedures," *Statist. Decisions*, Supplement Issue, vol. 1, pp. 205–237, 1984.

[15] I. Efroimovich, "Information contained in a sequence of distributions," *Probl. Inform. Transm.*, vol. 15, pp. 24–39, 1980.

[16] P. J. Huber, "Robust estimation of a location parameter," *Ann. Math. Statist.*, vol. 35, pp. 73–101, 1964.

[17] ———, *Robust Statistics*. New York: Wiley, 1981.

[18] I. A. Ibragimov and R. Z. Hasmimsky, "On the information in a sample about a parameter," in *Proc. 2nd Int. Symp. Information Theory*. Budapest, Hungary: Akademiai Kaido, 1973, pp. 295–309.

[19] F. Kanaya and K. Nakagawa, "On the practical implication of mutual information for statistical decision making," *IEEE Trans. Inform. Theory*, vol. 37, pp. 1151–1156, July 1991.

[20] B. Y. Levit, "On asymptotic minimax estimates of the second order," *Theory Probab. Appl.*, vol. 25, no. 3, pp. 552–568, 1980.

[21] R. J. McEliece, *The Theory of Information and Coding*. Reading, MA: Addison-Wesley, 1977.

[22] A. Yuan, "A minimally informative likelihood approach to Bayesian inference and decision analysis," Ph.D. dissertation, Dept. Statistics, Univ. British Columbia, Vancouver, BC, Canada, 1997.

[23] A. Yuan and B. Clarke, "A minimally informative likelihood approach to Bayesian decision analysis and robustness agaisnt modeling strategy," submitted to *Can. J. Statist.*, 1998.