A Minimally Informative Likelihood for Decision Analysis: Illustration and Robustness

(Abbreviated Title: Minimal Information Likelihoods)

By Ao YUAN and Bertrand S. CLARKE

Massachusetts General Hospital and University of British Columbia

Abstract

The authors discuss a class of likelihood functions involving weak assumptions on data generating mechanisms. These likelihoods may be appropriate when it is difficult to propose models for the data. The properties of these likelihoods are given and it is shown how they can be computed numerically by use of the Blahut-Arimoto algorithm. The authors then show how these likelihoods can give useful inferences using a data set for which no plausible physical model is apparent. The plausibility of the inferences is enhanced by the extensive robustness analysis these likelihoods permit.

Résumé

Les auteurs montrent comment il est possible, en l'absence de modèle naturel pour des observations, de construire une classe de fonctions de vraisemblance à partir d'hypothèses très faibles concernant l'origine des données. Ils présentent les propriétés de ces vraisemblances à information minimale et expliquent comment les calculer à l'aide de l'algorithme de Blahut-Arimoto. Ils illustrent la faisabilité et l'utilité de cette approche au moyen d'un exemple concret. Comme cette méthode se prête bien à une étude de robustesse, les conclusions auxquelles elle conduit sont d'autant plus plausibles.

Key words and phrases: information, rate distortion function, Blahut-Arimoto algorithm, robustness.

AMS 1991 subject classifications: Primary 62B10, 62F15; secondary 62F35, 62-07.

1. INTRODUCTION

Consider a parameter $\theta \in \mathbb{R}^k$ with density $w(\theta)$ with respect to Lebesgue measure and let $X^n = (X_1, \ldots, X_n)$ be a vector of observations whose conditional density given θ is denoted $p(x^n|\theta)$. The Shannon Mutual Information (SMI) between Θ and X^n, usually denoted $I(\Theta, X^n)$, is the relative entropy

$$D(f||g) = \int f \log(f/g) dx d\theta = \int p(x|\theta) w(\theta) \log \{p(x|\theta)/m(x)\} \, dx d\theta \tag{1}$$

between the joint density $f(x^n, \theta) = p(x^n|\theta)w(\theta)$ of the pair (X^n, Θ) and the product $g(x^n, \theta) = m(x^n)w(\theta)$ of their marginals.

As the sample size n tends to infinity, maximizing (1) over w leads to the reference prior method of Bernardo (1979). It has also been studied by Ibragimov and Hasminsky (1973), Efroimovich (1980), and Clarke and Barron (1994), among others. Rather than fixing a likelihood and maximizing over w, we fix w and minimize over a class of likelihoods $\mathcal{P}_\ell$ with Bayes risk for estimating θ bounded by ℓ. The result of minimizing (1) is the rate distortion function (see Berger, 1971). Rewriting (1) as $E_m D\{w(\cdot|X^n)||w(\cdot)\}$, we see that minimizing it gives the likelihood updating the prior the least in an asymptotic sense. We call this the minimally informative likelihood (MIL) because it makes the weakest assumptions permitted within $\mathcal{P}_\ell$ about the relationship between the data and the parameter.

Here we use MIL's to analyze a data set for which it is difficult to propose a physically plausible parametric family. This automation of likelihood selection means one can produce parameter estimates readily, but that confidence intervals or credible sets will be larger than would be found using a valid parametric family. However, even though conclusions may be weak, they can answer some questions of interest. In addition, the generality of the model generating method presented here permits diverse modeling strategies, based on several parametrizations and summarizations of data. Indeed, since all MIL's satisfy analogous optimality criteria, the models they generate start with equal plausibility. Thus, it is fair to evaluate the robustness of inferences to modeling strategy.

The MIL procedure is a special case of the maximum entropy procedure (see Jaynes, 1982, Soofi *et al.*, 1995) and has a physical meaning in terms of data compression as well as codelength. Independently of the data, but not of other assumptions, the MIL provides a parametric family one can use for inference. Formalizing the minimization requires that we fix a prior, a loss function, a value ℓ, the number k of components in θ, and what we will mean by X. The variable X may be univariate or multivariate; it may be a response or a summary statistic. Given those inputs, a formula for the MIL can be found in Blahut (1972a, Chapter 8). However, its values can only be determined computationally, for instance by the Blahut-Arimoto algorithm (see Blahut, 1972b, Arimoto, 1972).

Section 2 formally defines MIL's. In Section 3, we use MIL's in two different ways to re-analyze the data from Nader and Reboussin (1994). In Section 4, we define a third modeling strategy and compare it to the first two. In a short concluding section we discuss our results, in particular the implications for modeling. Proofs from Sections 2

and 4 are gathered together in Appendices I and II.

2. THE MIL METHOD

2.1 Background and description. Suppose $\theta = (\theta_1, \ldots, \theta_k) \in I\!R^k$ is an assessment of location for p_θ. Our task is to associate a parametric family $p(x^n|\theta)$ to X^n. Although the form of $p(\cdot|\theta)$ is unknown, for a given k an experimenter may still formulate pre-experimental conceptions into the prior density $w(\theta)$. First suppose $X = X_i$ is a single outcome. We identify an MIL for this X and then discuss the extension to an n-fold MIL for a data vector. Here we treat the continuous case but comment that the discrete case is similar if one replaces integration with summation.

To obtain a meaningful minimum of the SMI requires that we restrict the collection of conditional densities over which the minimization is done. So, let $L(x, \theta)$ be a loss function, fix w, and minimize $I(\Theta, X)$ over the class of conditional densities $\mathcal{P}_\ell$ defined to be the set of parametric families of densities on a measure space $(\mathcal{X}, \theta)$ that satisfy

$$\int\int p(x|\theta)w(\theta)L(x,\theta)dxd\theta \leq \ell. \tag{2}$$

The extra parameter ℓ bounds the Bayes risk incurred by estimating θ with $X = x$. Now, the MIL is the likelihood achieving

$$R(\ell) = \inf_{p\in\mathcal{P}_\ell} I(\Theta, X). \tag{3}$$

Expression (3) is the rate distortion function (see Cover and Thomas, 1991, and Blahut, 1987). It is shown in Blahut (1972b) that the minimum in (3) is achieved by

$$p^*_\lambda(x|\theta) = p^*(x|\theta) \propto m^*(x)e^{-\lambda L(x,\theta)}, \tag{4}$$

where $m^*_\lambda(x) = m^*(x)$ is determined by the expression

$$\int \frac{e^{-\lambda L(x,\theta)}w(\theta)}{\int m^*(y)e^{-\lambda L(y,\theta)}dy}d\theta = 1 \tag{5}$$

for x's which have $m^*(x) > 0$, and $\lambda \geq 0$ is determined by ℓ. Here $m^*(x) = \int p^*(x|\theta)w(\theta)d\theta$ is the marginal density.

For n independent outcomes one can use the product of the univariate MIL's. If one can not assume independence then one would optimize the analogous quantities replacing the univariate X by a random vector X^n. In this case, x and y in expressions (2) to (5) would be replaced by $x^n = (x_1, \ldots, x_n)$ and $y^n = (y_1, \ldots, y_n)$, and $L(x, \theta)$ by $L_n(x^n, \theta) = \sum L(x_i, \theta)$, the loss from estimating θ n times.

Closed form MIL's exist only in special cases.

Example 1. Suppose L is squared error loss and w is $N(\mu, \sigma^2)$. When $\ell < \sigma^2$, one can derive that $m^*(\cdot)$ is $N(\mu, \sigma^2 - 1/2\lambda)$ and then verify that $p^*(\cdot|\theta)$ is

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

see Yuan and Clarke (1999). In this case, as the prior variance decreases to $1/2\lambda$, the MIL variance decreases to zero and as the prior variance increases to infinity the variance of the MIL increases to $1/2\lambda$.

If one permits the mixture m^* to be identically one on the whole real line, a solution to the optimization problem still exists if one uses the smallest value of r for which $w(\theta|x^r)$ is proper and if (2) is changed to

$$\int L(x, \theta) w(\theta|x^r) d\theta \leq \ell,$$

redefining $\mathcal{P}_\ell$ appropriately. By (4), optimizing

$$R(\ell) = \inf_{p \in \mathcal{P}_\ell} \int w(\theta|x^r) \log\left\{\frac{w(\theta|x^r)}{w(\theta)}\right\} d\theta$$

gives $p^*(x|\theta) = \sqrt{\lambda/\pi}\exp\{-\lambda(x-\theta)^2\}$. Expression (5) is satisfied with equality on the real line and the posterior for Θ given x is $N(x, 1/2\lambda)$. One can show that $\lambda \geq 1/2\ell$. Since the SMI is $I(\Theta, X) = \log(\lambda/\pi) - (1/2)$, we have $R(\ell) = -(1/2)\{1 + \log(2\pi\ell)\}$. The minimum is achieved by $\lambda = 1/2\ell$ so the MIL is $N(\theta, \ell)$.

Example 2. Choose w to be $N(\mu, \sigma^2)$, and set $\mu = 0$ and $\sigma = 1$ for convenience. Define L by $L(x, \theta) = (x_1 + x_2 - 2\theta)^2$. It can be verified that for $\lambda = 3/8$, $p^*(x_1, x_2|\theta)$ is given explicitly by using $m^*(\cdot, \cdot) \sim N\{\underline{0}, (1/2\lambda)I_2\}$ to get

$$p^*(\cdot, \cdot|\theta) \sim N\left\{(2\theta/3)\underline{1}, 4/9\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}\right\},$$

where $\underline{0}$ and $\underline{1}$ represent vectors whose entries are all 0's or 1's, and I_k is the $k \times k$ identity matrix. The upper bound on the Bayes risk occurs for $\lambda = 3/8$ and $\ell(3/8) = 4/3$.

Example 3. Consider an n-dimensional data set with a summary statistic and a one-dimensional parameter using the loss function $L(x^n, \theta) = (\bar{x} - \theta)^2$ with a $N(0, 1)$ prior density for θ. For $\lambda = \lambda_n = (n+1)/2$ one can verify that $m^*(\cdot, \ldots, \cdot) \sim N_n(\underline{0}, A_n)$, and A_n is the matrix given by

$$A_n^{-1} = 2\lambda\{(n+1)I_n - \underline{1}^t\}/n^2.$$

The corresponding MIL is an n-dimensional independent multivariate normal

$$p^*(\cdot, \ldots, \cdot|\theta) \sim N_n\left\{\frac{n}{n+1}\,\theta\,\underline{1}, \frac{n^2}{(n+1)^2} I_n\right\}.$$

The bound on the Bayes risk in the constraint is $\ell(n) = n/(n+1) = 1-1/2\lambda_n$. In normal examples, θ often appears in the mean; more generally its interpretation depends on L.

2.2 Some key properties of the MIL method. One can verify that $p^*(x|\theta)$ as defined in (2) to (5) is well defined and unique and that the relative entropy between a posterior formed from an n-fold MIL and a prior converges to zero as n increases to infinity. Conditions and other related results can be found in Yuan and Clarke (1999).

We comment that if one uses a loss function as an inference function, then the likelihood equation under the corresponding MIL is an estimating equation which yields consistent and asymptotically normal estimates. Also, one can perform hypothesis tests to infer values of λ. These results can be found in Yuan (1997, p. 57 and 78). Here, we give the following two propositions.

Let $\theta = \theta(\eta)$ have a continuous and positive derivative $\theta'(\eta)$.

Proposition 2.1. Let $\bar{p}^*(x|\eta)$ be the MIL with respect to loss function $\bar{L}(x,\eta) = L\{x,\theta(\eta)\}$ and prior $\bar{w}(\eta) = w\{\theta(\eta)\}\theta'(\eta)$. Then $\bar{p}^*(x|\eta) = p^*\{x|\theta(\eta)\}$.

Proof: See Appendix I.

Let $I_{p_n^*}(\theta)$ be the Fisher information of the MIL. From the definition of the MIL we have that $I_{p_n^*}(\theta) = \mathrm{Var}_{p_n^*}\{\frac{\partial}{\partial\theta}L_n(X^n,\theta)\}$. Since the MIL depends on n and is exchangeable, we denote the common correlation between coordinates X_i and X_j of X^n, for $i \neq j$ by $\rho_n(\theta) = \mathrm{Corr}_{p_n^*}\{\frac{\partial}{\partial\theta}L(X_i,\theta), \frac{\partial}{\partial\theta}L(X_j,\theta)\}$. We have the following.

Proposition 2.2.

(i) The MIL p_n^* is exchangeable in its arguments.

(ii) Let $\underline{\rho}$ denote lim inf ρ_n. Then p_n^* is asymptotically non-negatively correlated, that is, $\underline{\rho} \geq 0$.

(iii) Let $\sigma_n^2(\theta) = \mathrm{Var}_{p_n^*}\{L'(X_1,\theta)\}$, $L_n(x^n,\theta) = a_n \sum_{i=1}^n L(x_i,\theta)$. Then

$$I_{p_n^*}(\theta)/n = a_n^2\sigma_n^2(\theta)\{1 + (n-1)\rho_n(\theta)\}.$$

Proof: See Appendix I.

2.3. Computational approach. The main technique for obtaining an MIL is the Blahut-Arimoto algorithm, see Blahut (1972a, b) and Arimoto (1972). This algorithm is an instance of a broad class of algorithms based on alternate maximization. Csiszar and Tusnady (1984) have established the convergence properties of this class of algorithms. The alternate maximization algorithm that we implemented is described in Cover and Thomas (1991). In the present context, the procedure is as follows.

Fix L, λ, w and an initial m_0 so as to form

$$p_{1,\lambda}(x|\theta) = \frac{m_0(x)e^{-\lambda L(x,\theta)}}{\int m_0(y)e^{-\lambda L(y,\theta)}dy}. \tag{6}$$

Marginalizing in (6) gives

$$m_1(x) = \int p_{1,\lambda}(x|\theta)w(\theta)d\theta. \tag{7}$$

Next, replace m_0 in (6) by m_1 from (7) to form $p_{2,\lambda}(x|\theta)$. Now one obtains m_2 from $p_{2,\lambda}(\cdot|\theta)$ by mixing out θ. Thus, one generates a sequence $p_{i,\lambda}(x|\theta)$ for a given λ, x and θ. Csiszar (1974) showed that $p_{i,\lambda}(x|\theta)$ converges to $p^*_\lambda(x|\theta)$ as i tends to infinity. This $p^*_\lambda(x|\theta)$ is on the boundary of $\mathcal{P}_\lambda$. We assess convergence of $p_{i,\lambda}(x|\theta)$ to its limit $p^*_\lambda(x|\theta)$ in supremum norm, terminating the computation when $\sup_x |p_{i,\lambda}(x|\theta) - p_{i-1,\lambda}(x|\theta)|$ is small enough. This convergence is independent of m_0.

3. DATA ANALYSIS VIA MIL'S

In this section, products of univariate MIL's are used to re-analyze data from Nader and Reboussin (1994) in two different ways, corresponding to two different summarizations of the data.

3.1. Description of the data. The experimental data studied in Nader and Reboussin (1994) was collected to see whether a history of responding under schedules that reinforce either high or low response rates could modify previously established rates of cocaine self-administration. In this experiment, 8 monkeys were initially trained under the FI5 schedule: The first time that a monkey pulled a lever after having waited at least five minutes produced a cocaine injection.

Next, the monkeys were rated from one to eight based on response rates under the FI5 schedule. The monkeys were then paired; the two highest ratings gave the first block; the third and fourth highest gave the next and so on. Within each block, members were randomly assigned to one of two cocaine self-administration reinforcement schedules. Thus, four monkeys were trained under an FR50 schedule, that is for every fifty responses (lever pulls) the monkey got an injection of cocaine. The other four monkeys were trained under an IRT30 schedule; the monkeys were reinforced by a cocaine dose for lever presses at least thirty seconds apart.

Following the 65th session under FR50 or IRT30, availability of cocaine was again scheduled under FI5 for sixty consecutive sessions. The primary dependent variable was response rate which was the total number of responses during the session divided by the session length in minutes.

Part of the data analysis presented in Nader and Reboussin (1994) was a repeated measures analysis of variance for the rate data. Pairs and treatment group are between subject effects, session is a within subject effect. Nader and Reboussin (1994) sought

linear trends in the rate over the 60 sessions and asserted that the apparent nonlinearity did not affect the conclusions substantially. This model did not reject the hypothesis that the mean rates are the same in both groups, though the p-value was in the suggestive range. However, it did reveal a highly significant difference between the mean linear trends in the two groups over the sixty sessions.

Other models examined by Nader and Reboussin (1994) gave similar conclusions. This included one model with an AR(1) component over the 60 sessions, one allowing some curvature in the trend over sessions, and several excluding the early sessions. In these cases, the conclusions were much the same: not quite significant mean difference between groups, highly significant difference in linear trends.

Here, we augment this analysis with two other models based on MIL's. We treat the animals as paired and look for training differences in the mean response rates. We find a significant difference in the mean rates which the earlier data analysis did not establish.

3.2. Model I. Label the pairs of monkeys (1, 2), (3, 4), (5, 6) and (7, 8), where odd labels indicate training under FR50 and the even labels indicate training IRT30. Let y_{ij} be the datum on rate for the i-th monkey on the j-th day where $i = 1, \ldots, 8$ and $j = 1, \ldots, 60$. For each i, let $\bar{y}_i = \frac{1}{60}\sum_{j=1}^{60} y_{ij}$ be the sample mean of the rate data from the i-th monkey. Write the differences of the means within each pair as $x_1 = \bar{y}_1 - \bar{y}_2$, $x_2 = \bar{y}_3 - \bar{y}_4$, $x_3 = \bar{y}_5 - \bar{y}_6$, and $x_4 = \bar{y}_7 - \bar{y}_8$. Now, we have $x_1 = 0.8001695$, $x_2 = 4.098834$, $x_3 = 9.423503$ and $x_4 = 1.791167$.

For model I, suppose the expected values of the x_i's are the same and treat this as an approximation to the parameter of interest θ. By assuming dependence is only within pairs and that baseline FI5 lever pressing rates cancel in the difference we retain only the training effects in $x_1, \ldots, x_4$ which are i.i.d. For a prior $w(\theta)$, a loss function L, and a value λ, we use a product of univariate MIL's. So, we get a posterior for θ given (x_1, x_2, x_3, x_4) from which we can make inferences about the expectation of x_i.

Formally, we find the posterior density

$$w^*(\theta|x_1, x_2, x_3, x_4) = \frac{p^*(x_1|\theta)p^*(x_2|\theta)p^*(x_3|\theta)p^*(x_4|\theta)w(\theta)}{\int p^*(x_1|\eta)p^*(x_2|\eta)p^*(x_3|\eta)p^*(x_4|\eta)w(\eta)d\eta}. \tag{8}$$

We used the iterative method described in Subsection 2.3 to approximate $p^*(x|\theta)$ and then graphed (8) for various w's, L's and λ's. In particular, we chose w to be $U(-15, 20)$ or any of $N(-2, 1)$, $N(0, 1)$, or $N(2, 1)$; we choose $L(x, \theta) = (x - \theta)^2$, or $L(x, \theta) = |x - \theta|$. Choosing λ in [1/10, 10] led to feasible computations and we suggest [1/2,5] is a reasonable range for λ in this problem.

Our results for these cases were generally consistent: The posterior density was unimodal and essentially concentrated on the positive half line, with mode between 4 and 8.5, and posterior variance typically around 1. Thus, we infer that there is a significant effect from FR50 and IRT30 training, with the rate for the IRT30 group being much less than the rate for the FR50 group. Fig. 1a shows some of the posteriors we obtained.

Note that if we had 6 or 600 sessions rather than 60, the results would not change. This is a deficiency of our analysis, but not of the method: We could have used a

60-variate MIL for the 60-variate data rather than reducing 60-variate data to a real number. However, the former approach would have been too computationally time consuming. Model II below has the same deficiency for the same reason. However, because the sessions are highly dependent the summary statistics may contain essentially all the useful information in the data.

3.3 Model II. Again, we average over the 60 data points for each monkey. Instead of taking differences within pairs, however, we use products of MIL's for the two groups separately. Thus, there are two parameters in the product of the likelihoods for the two groups. Given a bivariate posterior, marginalization gives a posterior for the difference in the parameters. For now, we assume $w(\theta_1, \theta_2) = w_1(\theta_1)w_2(\theta_2)$.

One can obtain four means $x = (x_1, x_2, x_3, x_4)$ for the FR50 group and four means $y = (y_1, y_2, y_3, y_4)$ for the IRT30 group. Regard these as independent and identical outcomes from their respective distributions. The data from the FR50 group give (2.6495, 5.0032,19.0287,3.3857) as a value for x and the data from the IRT30 group give (1.8493,0.9043, 9.6052,1.5945) for y. From w_1, we get the MIL $p_1^*(x|\theta_1) = \prod_{i=1}^4 p_1^*(x_i|\theta_1)$, and from w_2 we get the MIL $p_2^*(y|\theta_2) = \prod_{i=1}^4 p_2^*(y_i|\theta_2)$, temporarily suppressing the dependence on λ. Now, we can form the two dimensional posterior

$$w(\theta_1, \theta_2|x, y) = \frac{w_1(\theta)p_1^*(x|\theta_1)w_2(\theta)p_2^*(y|\theta_2)}{m^*(x, y)} = w_1(\theta_1|x)w_2(\theta_2|y),$$

where $m^*(x, y)$ is the marginal from $p(x, y|\theta_1, \theta_2) = p_1^*(x|\theta_1)p_2^*(y|\theta_2)$ and w, and $w_1(\theta_1|x)$ and $w_2(\theta_2|y)$ are the corresponding one dimensional marginals. Using $\psi = \theta_1 + \theta_2$ and $\phi = \theta_1 - \theta_2$ in the bivariate posterior and integrating over ψ gives a posterior for ϕ.

For $N(0, 1)$ priors and squared error loss we found that as λ increases, the posterior concentrates on positive values of θ. However, our implementation of the Blahut-Arimoto algorithm is inadequate for λ much above 0.09 or much below 0.0001. The limited range of λ is more serious for Model II than for Model I. The problem may be that as λ increases, the factors $w_1(\theta_1|x)$ and $w_2(\theta_2|y)$ concentrate at different points. Fig. 1b shows posteriors for $\lambda = 0.0001$, 0.09. Intermediate values of λ give posteriors roughly between these two.

3.4. Robustness of the analysis. Posteriors for two λ's using a $U(-10, 15)$ prior in place of a normal prior are shown in Fig. 2. For a reasonable range of λ's, inference about the direction of the difference between the FR50 and IRT30 groups is the same as from Fig. 1. The posteriors in Fig. 2 have nearly the same modes; the posteriors in Fig. 1a do not. We attribute the insensitivity to λ in Fig. 2 to the non-informativity of the uniform. The posterior based on a normal prior may be more sensitive to λ because the MIL compensates for prior information in part by reflecting the opposite of that information.

Using absolute value for L instead of squared error does not change the results much. Fig. 3 shows posteriors for two λ's using a $N(0, 1)$ prior. The posterior concentrates on

the positive real line, although the peak is distorted due to the 'corner' of the absolute value function.

Results for Model II are similar. Indeed, if the λ in Model II is chosen to be as close as possible to the values used for Model I (without exceeding 0.09) then the two models give the same inferences. Agreement between different models increases the plausibility of the conclusion.

4. ROBUSTNESS AGAINST MODELING STRATEGY

The inferential similarity of Models I and II motivates our investigation of robustness to modeling strategy. We define three strategies and seek conditions under which they will give similar inferences. Two of these models are the general cases of Model I and Model II. The third, Model III, we did not use, but is equally plausible. Some of our results are for MIL's only, others are more general.

4.1. General definition of the models. Suppose there are two data sets X^n and Y^n and the task is to make inferences about a unidimensional parameter θ. The parameter represents location, but no more precise interpretation need be enforced. Thus, Model I uses an MIL determined by a prior $w(\theta)$ for Z^n where $Z^n = X^n - Y^n$. We write $p_{(1)}(z^n|\theta)$ for the likelihood and $w_{(1)}(\theta|z^n)$ for the posterior. Model II uses two parameters, θ_1 and θ_2, and two MIL's determined by priors $w_1(\theta_1)$ and $w_2(\theta_2)$ for the two i.i.d. sets of data X^n and Y^n. We write $p_1^*(x^n|\theta_1)$ and $p_2^*(y^n|\theta_2)$, and combine them to obtain the joint posterior $w(\theta_1, \theta_2|X^n, Y^n)$. Using $(\theta, \phi) = (\theta_1 - \theta_2, \theta_1 + \theta_2)$ and marginalizing gives $w_{(2)}(\theta|X^n, Y^n)$.

A third model is formed by taking differences in the data as in Model I and in the parameter as in Model II. Thus, Model III starts with MIL's $p_1(x^n|\theta_1)$ and $p_2(y^n|\theta_2)$ for X^n and Y^n. Using the transformation $Z^n = X^n - Y^n, S^n = X^n + Y^n$ and integrating over S^n gives the density $p_{(3)}(z^n|\theta_1, \theta_2)$ for Z^n, which leads to a bivariate posterior. Using the transformation $\theta = \theta_1 - \theta_2$, $\phi = \theta_1 + \theta_2$ and integrating over ϕ gives the posterior $w_{(3)}(\theta|z^n)$ for θ.

4.2. Equivalence of the three modeling strategies. The two results in this subsection are general. The first says that for certain priors the inferences from Models I, II and III will be identical, on average. Let the priors for Model I, Model II and Model III be, respectively, $w_{(1)}(\theta_1) = w(\theta_1)$, $w_{(2)}(\theta_1, \theta_2) = w_1(\theta_1)w_2(\theta_2) = w_{(3)}(\theta_1, \theta_2)$, and assume w is the density for $\theta = \theta_1 - \theta_2$. Our result is the following.

Proposition 4.1 With expectations taken with respect to the mixture densities $m_{(1)}$, $m_{(2)}$ and $m_{(3)}$ in the three models we have

$$Ew_{(1)}(\theta|Z^n) = Ew_{(2)}(\theta|X^n, Y^n) = Ew_{(3)}(\theta|Z^n) = w(\theta).$$

Proof: See Appendix II.

Under two extra conditions we can show the stronger result that the posteriors from Models I, II, and III are identical in a pointwise sense. In this proposition, as in the last, the quantities are continuous so that slight deviations from the hypotheses are expected to produce only slight deviations from the conclusion. So, again, choose $w_{(1)}(\theta)$ to be the integral of $(1/2)w_1\{(\phi+\theta)/2\}w_2\{(\phi-\theta)/2\}$ over ϕ, and write the likelihood $p_{(1)}(z|\theta)$ for Model I as

$$\frac{1}{4}\int\int p(s,z|\phi,\theta)\pi(\phi)dsd\phi = \frac{1}{4}\int\int p_1\left(\frac{s+z}{2}|\frac{\phi+\theta}{2}\right)p_2\left(\frac{s-z}{2}|\frac{\phi-\theta}{2}\right)\pi(\phi)dsd\phi, \tag{9}$$

for some density π. Our result is the following.

Proposition 4.2 Suppose the joint likelihood for (S,Z) factors into a function of (z,θ) and a function that does not involve θ, i.e.,

$$p(s,z|\phi,\theta) = g(z,\theta)h(z,s,\phi) \tag{10}$$

for some functions g and h, and that the joint prior satisfies

$$w\left(\frac{\phi+\theta}{2},\frac{\phi-\theta}{2}\right) = w_1(\theta)w_2(\phi)$$

for some w_1 and w_2, then,

$$w_{(1)}(\cdot|Z) = w_{(2)}(\cdot|X,Y) = w_{(3)}(\cdot|Z).$$

We comment that if $w_1 = w_2$ is the standard normal then the condition in Proposition 4.2 is satisfied because $p_1(\cdot|\theta_1) = Normal(\theta_1,\sigma^2)$ and $p_2(\cdot|\theta_2) = Normal(\theta_2,\sigma^2)$.

Proof: See Appendix II.

4.3 Robustness against Modeling Strategies. Propositions 4.1 and 4.2 suggest it is rare for Models I, II, III to give identical inferences. However, we expect that in practical sense inferences from equally plausible models will be very similar. To address this point, we characterize the discrepancy between the posteriors from the different models. Yuan (1997) derives pointwise bounds on the L^1 distance between the posteriors from the three models.

In this subsection, we compare modeling strategies I, II, and III when MIL's are used in all three cases. So, consider the SMI's $I_0^*(n) = D\{p^*(Z^n|\cdot)w(\cdot)||m^*(Z^n)w(\cdot)\}$, $I_1^*(n) = D\{p_1^*(X^n|\cdot)w_1(\cdot)||m_1^*(X^n)w_1(\cdot)\}$, and $I_2^*(n) = D\{p_2^*(Y^n|\cdot)w_2(\cdot)||m_2^*(Y^n)w_2(\cdot)\}$, in which we have indicated the data involved in the relative entropy. These SMI's define a notion of typicality by their convergence to zero, see Yuan and Clarke (1999). To use this, let $C_i(n)$, for $i = 0,1,2$ be sequences of constants such that as $n \to \infty$ we have

$C_i(n) \to \infty$ and $C_i(n)I_i^*(n) \to 0$. Next, for $i = 0, 1, 2$, let S_i be subsets of the sample space defined by

$$S_0 = \{Z^n : D[w^*(\cdot|Z^n)||w(\cdot)] \leq C_0(n)I_0^*(n)\},$$
$$S_1 = \{X^n : D[w_1^*(\cdot|X^n)||w_1(\cdot)] \leq C_1(n)I_1^*(n)\},$$

and

$$S_2 = \{Y^n : D[w_2^*(\cdot|Y^n)||w_2(\cdot)] \leq C_2(n)I_2^*(n)\}.$$

We call such data sets canonical. For convenience, we write Θ, θ and p^* as Θ_0, θ_0 and p_0^*, and let P_i^* be the probability measure corresponding to $p_i^*, (i = 0, 1, 2)$. Next we show that for large sample sizes, the canonical sets have large $P_i^*(\cdot|\theta_i)$ probabilities, for all values of θ_i in a set of arbitrarily large probability under the prior distribution W_i.

Proposition 4.3 For any pre-assigned $\epsilon > 0$, there exist subsets $A_{i,n}(\epsilon)$ in the domain of $\Theta_i, i = 0, 1, 2$ (we write Θ_0 for Θ) so that as $n \to \infty$,

$$\forall \theta_i \in A_i, \quad P_i^*(S_i^c|\theta_i) \to 0, \quad \text{and} \quad \forall n \quad W_i(A_{i,n}^c) \leq \epsilon.$$

Proof: See Appendix II.

Proposition 4.3 accomplishes a key step in the proof of Theorem 4.4 below so as to give conditions under which data from canonical sets analyzed with MIL's using strategy I, II, or III results in posteriors, that are close to each other in variation distance. This implies that, for large sample sizes, credible sets and hypothesis tests from different modeling strategies will not be too different. Write $\mathcal{B}$ for the Borel field on the appropriate parameter space. We have the following.

Theorem 4.4 Let Model I be obtained from the prior w as in Proposition 4.1, and suppose Models II and III are obtained by using the priors w_1 and w_2 as described in Section 4.1. Suppose also that $I_i^*(n) \to 0$ as $n \to 0$, for $i = 0, 1, 2$. Then, if Models I, II, and III are formed from MIL's, the posteriors from these models conditional on canonical data satisfy

$$\sup_{B \in \mathcal{B}} |W_{(i)}(B|X^n, Y^n) - W_{(j)}(B|X^n, Y^n)| \leq C_{i,j}(n) + o_{p_1^*+p_2^*,i,j}(1),$$

for $1 \leq i, j \leq 3$, where $C_{i,j}(n) \to 0$ as $n \to \infty$. The error terms are $o_{p_1^*+p_2^*,1,2}(1) = 0$, $o_{p_1^*+p_2^*,1,3}(1) = o_{p_1^*+p_2^*,2,3}(1) = o_{p_1^*(\cdot|\theta_1)}(1) + o_{p_2^*(\cdot|\theta_2)}(1)$, for $\theta_1 \in A_1, \theta_2 \in A_2$ where convergences in probability are assessed in the appropriate mixture density, and A_1 and A_2 have prior probability that can be made arbitrarily close to 1 as n increases.

Proof: See Appendix II.

We now have that Models I, II, and III are similar but not identical. So, we ask which pair of them is closest. The next result gives that Models II and III are the closest, and

Models I and III differ most. Since Models I and III are conditional on Z^n whereas Models II and III involve marginalizing out a parameter, marginalization in the parameters may have a smaller effect than changes in summary statistics. Moreover, Proposition 4.5 below suggests that the conclusions from Model III will be close to those of Model II, but not as strong, because Model II conclusions were weaker than Model I conclusions.

Proposition 4.5 For Models I, II and III as defined in subsection 4.1, we have

$$E_{(X^n,Y^n)}\left[D\{w_{(2)}(\cdot|X^n,Y^n)||w_{(1)}(\cdot|Z^n)\} - D\{w_{(2)}(\cdot|X^n,Y^n)||w_{(3)}(\cdot|Z^n)\}\right] > 0, \quad (11)$$

$$E_{(X^n,Y^n)}\left[D\{w_{(3)}(\cdot|Z^n)||w_{(1)}(\cdot|Z^n)\} - D\{w_{(3)}(\cdot|Z^n)||w_{(2)}(\cdot|X^n,Y^n)\}\right] > 0, \quad (12)$$

and

$$E_{(X^n,Y^n)}\left[D\{w_{(1)}(\cdot|Z^n)||w_{(3))}(\cdot|Z^n)\} - D\{w_{(1)}(\cdot|Z^n)||w_{(2)}(\cdot|X^n,Y^n)\}\right] > 0, \quad (13)$$

where expectations are with respect to the marginal distribution of (X^n, Y^n).

Proof: See Appendix II.

Finally, we note that the average discrepancy between the models which are furthest apart (I and III) can be taken as an assessment of robustness for the three modeling strategies. To simplify the problem, we may fix the priors. For Model I, we take the prior

$$w_{(1)}(\theta) = \frac{1}{2}\int w_1\left(\frac{\phi+\theta}{2}\right) w_2\left(\frac{\phi-\theta}{2}\right) d\phi,$$

and the likelihood to be

$$p_{(1)}(z|\theta) = \frac{1}{4}\int\int p_1\left(\frac{s+z}{2}|\frac{\phi+\theta}{2}\right) p_2\left(\frac{s-z}{2}|\frac{\phi-\theta}{2}\right) ds d\phi.$$

Now the question becomes the robustness of the likelihood pairs $\{p_1(\cdot|\theta_1), p_2(\cdot|\theta_2)\}$ against the three modeling strategies. Thus, we could use $R = E_{m_{(3)}(z)} D\{w_{(3)}(\cdot|Z)||$ $w_{(1)}(\cdot|Z)\}$, or we could take expectations inside the arguments of the relative entropy and use $R' = D[w_{(3)}(\cdot)||E_{m_{(3)}(z)}\{w_{(1)}(\cdot|Z)\}]$ to assess robustness. The log-sum inequality, see Cover and Thomas (1991), gives $R' \leq R$ suggesting that R is a stronger measure of robustness.

5. DISCUSSION

The main point of this paper was to demonstrate that MIL's can be used to answer physically meaningful problems and that robustness against the modeling strategy may substitute for physical modeling in some cases. That is, strong robustness properties —

in particular against modeling strategy — may substitute for knowing the likelihood as a technique for ensuring inferences are valid.

In our re-analysis of data from Nader and Reboussin (1994) we found a significant difference in mean rates, a slight improvement on the results reported there. We suggest that this improvement occurred because MIL's can give likelihoods for use with summary statistics. In this case, using the summary statistics better overcame the noninformativity of the MIL. This is consistent with the fact that the results from Model I seemed more conclusive than the results of Model II: Model II does not make direct use of the pairing.

In our analysis of robustness against the modeling strategy we found three models which we showed would be in broad agreement. We characterized when they were equivalent and then determined which pairs were closer than other pairs. Together, our results suggest that robustness against modeling strategy may be a useful check on the plausibility of analyses in general.

APPENDIX I: PROOFS FROM SECTION 2

Proof of Proposition 2.1: Let $\bar{I}(\eta, X)$ be the mutual information between η and X, and write $\bar{w}(\eta) = w(\eta)\theta'(\eta)$. We have

$$\int\int p(x|\theta)w(\theta)\log\left\{\frac{p(x|\theta)}{m(x)}\right\}dxd\theta = \int\int \bar{p}(x|\eta)\bar{w}(\eta)\log\left\{\frac{\bar{p}(x|\eta)}{\bar{m}(x)}\right\}dxd\eta,$$

where $\bar{m}(x) = \int \bar{p}(x|\eta)\bar{w}(\eta)d\eta = m(x)$, and therefore $I(\Theta, X) = \bar{I}(\eta, X)$. By (4) and (5) the $\bar{p}^*(x|\eta)$ which minimizes $\bar{I}(\eta, X)$ is given by

$$\bar{p}^*(x|\eta) = \frac{\bar{m}^*(x)e^{-\lambda\bar{L}(x,\eta)}}{\int \bar{m}^*(y)e^{-\lambda\bar{L}(y,\eta)}dy},$$

where $\bar{m}^*(\cdot)$ satisfies:

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

with equality for x in the support of $\bar{m}^*$. However, m^* has the same support as $\bar{m}^*$ so $m^* = \bar{m}^*$. Thus $\bar{p}^*(x|\eta) = p^*\{x|\theta(\eta)\}$ as claimed.

Proof of Proposition 2.2: (i) This follows from the symmetry of the optimization procedure.

(ii) By way of contradiction, suppose $\underline{\rho} < 0$. If c_2 denote the variance of $\frac{\partial}{\partial\theta}L(x, \theta)$, then

$$\mathrm{Var}_{p_n^*}\left\{\sum_{i=1}^{n}\frac{\partial}{\partial\theta}L(X_i, \theta)\right\} \leq nc_2 + n(n+1)\rho_n c_2,$$

If ρ_n is negative for large n then a variance is negative, a contradiction.

(iii) The Dominated Derivative Theorem gives

$$\frac{\partial}{\partial\theta}\int m_n^*(y^n)e^{-L_n(y^n,\theta)}dy^n = -\int m_n^*(y^n)e^{-L_n(y^n,\theta)}L_n'(y^n,\theta)dy^n.$$

Apply the quotient rule to p_n^* from (4) with $\lambda_n = 1$. Using the last expression gives

$$\frac{\partial}{\partial\theta}p_n^*(X^n|\theta) = -L_n'(X^n,\theta)p_n^*(X^n|\theta) + p_n^*(X^n|\theta)\int L_n'(y^n,\theta)p_n^*(y^n|\theta)dy^n. \qquad (A.1)$$

Using (A.1), the Fisher information is

$$I_{p_n^*}(\theta) = E_{p_n^*}\left\{\frac{\frac{\partial}{\partial\theta}p_n^*(X^n|\theta)}{p_n^*(X^n|\theta)}\right\}^2 = E_{p_n^*}\left\{L'(X^n,\theta) - E_{p_n^*}L'(X^n,\theta)\right\}^2. \qquad (A.2)$$

where $X^n = (X_{n,1},\ldots,X_{n,n})' \sim p_n^*(x^n|\theta)$. Expression (A.2) is a variance so we have

$$\begin{aligned} I_{p_n^*}(\theta) &= \mathrm{Var}_{p_n^*}\left\{a_n\sum_{i=1}^n L'(X_{n,i},\theta)\right\} \\ &= a_n^2\left[\sum_{i=1}^n \mathrm{Var}_{p_n^*}\left\{L'(X_{n,i},\theta)\right\} + \sum_{i\neq j}^n \mathrm{Cov}_{p_n^*}\left\{L'(X_{n,i},\theta), L'(X_{n,j},\theta)\right\}\right] \\ &= na_n^2\sigma_n^2(\theta)\{1+(n-1)\rho_n(\theta)\}. \end{aligned}$$

APPENDIX II: PROOFS FROM SECTION 4

Proof of Proposition 4.1: Since $w_{(1)} = w$ by definition, we get $E_{Z^n}w_{(1)}(\theta|Z^n) = w(\theta)$.

For $w_{(2)}$, observe that the joint marginal $m_{(2)}(x^n,y^n)$ is

$$\int\int p_1(x^n|\theta_1)p_2(y^n|\theta_2)w_1(\theta_1)w_2(\theta_2)d\theta_1 d\theta_2 = m_1(x^n)m_2(y^n).$$

Now, since $w_{(2)}(\theta|X^n,Y^n)$ is the integral over ϕ of the product of $w_1(\frac{\phi+\theta}{2}|X^n)$ and $w_2(\frac{\phi-\theta}{2}|Y^n)$, the independence of X^n and Y^n under $m_{(2)}$ gives

$$Ew_{(2)}(\theta|X^n,Y^n) = \frac{1}{2}\int Ew_1\left(\frac{\phi+\theta}{2}|X^n\right)Ew_2\left(\frac{\phi-\theta}{2}|Y^n\right)d\phi = w(\theta),$$

because w is the convolution of w_1 and w_2.

For $w_{(3)}$, we note $w_{(3)}(\theta|Z^n)$ is

$$\frac{1}{m_{(3)}(Z^n)}\frac{1}{4}\int\int p_1\left(\frac{s^n+Z^n}{2}|\frac{\phi+\theta}{2}\right)p_2\left(\frac{s^n-Z^n}{2}|\frac{\phi-\theta}{2}\right)ds^n w_1\left(\frac{\phi+\theta}{2}\right)w_2\left(\frac{\phi-\theta}{2}\right)d\phi.$$

Taking the expectation with respect to $m_{(3)}(Z^n)$ cancels the denominator. Then, integrating over z^n reduces the expression to the convolution w.

Proof of Proposition 4.2: Using (9) and (10) to factor the likelihood in Model I, we get the posterior

$$w_{(1)}(\theta|Z) = \frac{g(Z,\theta)\int\int h(Z,s,\phi)\pi(\phi)dsd\phi w_1(\theta)}{\int g(Z,\xi)\int\int h(Z,s,\phi)\pi(\phi)dsd\phi w_1(\xi)d\xi} \propto g(Z,\theta)w_1(\theta). \tag{A.3}$$

By (10), the likelihood $p_{(2)}(x,y|\theta_1,\theta_2)$, for Model II, factors into $g(z,\theta)h(z,s,\phi)$ and thereby gives the posterior

$$w_{(2)}(\theta|X,Y) = \frac{\frac{1}{2}g(Z,\theta)\int h(Z,S,\phi)w(\frac{\phi+\theta}{2},\frac{\phi-\theta}{2})d\phi}{\frac{1}{2}\int\int g(Z,\xi)h(Z,S,\phi)w(\frac{\phi+\xi}{2},\frac{\phi-\xi}{2})d\phi d\xi} \propto g(Z,\theta)w_1(\theta). \tag{A.4}$$

By (10), the likelihood $p_{(3)}(z|\theta_1,\theta_2)$ for Model III is $g(z,\theta)\int h(z,s,\phi)ds$. Thus, the posterior is

$$w_{(3)}(\theta|Z) = \frac{g(Z,\theta)\int\int h(Z,s,\phi)dsw(\frac{\phi+\theta}{2},\frac{\phi-\theta}{2})d\phi}{\int g(Z,\xi)\int\int h(Z,s,\phi)dsw(\frac{\phi+\xi}{2},\frac{\phi-\xi}{2})d\phi d\xi} \propto g(Z,\theta)w_1(\theta). \tag{A.5}$$

Taken together, (A.3), (A.4) and (A.5) complete the proof.

Proof of Proposition 4.3: We only prove the statement for $i = 1$; the other cases are similar. Markov's inequality gives

$$P_1^*(S_1^c|\theta_1) \leq \frac{1}{C_1(n)I_1^*(n)} E_{P_1^*} D\left\{w_1^*(\cdot|X^n)||w_1(\cdot)\right\}. \tag{A.6}$$

To see that (A.6) goes to zero, we begin by showing that for any pre-assigned $\epsilon > 0$ there is a set $A_{1,n}$, such that $e^{-L(x^n,\theta_1)}h(n,\theta_1)$ is bounded by a number B on $A_{1,n}$ uniformly in n and x^n with $W_1(A_{1,n}^c) \leq \epsilon$, where $h(n,\theta)$ is the inverse mixture $1/\int m_1^*(t^n)e^{-L_n(t^n,\theta_1)}d\theta_1$.

By way of contradiction, suppose there is an ϵ_0 and a sequence x^n so that for some sequence $N(n) \to \infty$ as $n \to \infty$, there is a sequence of sets $A'_{1,n}$ such that

$$e^{-L(x^n,\theta_1)}h(n,\theta_1) \geq N(n), \quad \text{on} \quad A'_{1,n}, \quad \text{and} \quad W_1(A_{1,n}^{'c}) \geq \epsilon_0.$$

Then,

$$\int e^{-L(x^n,\theta_1)}h(n,\theta_1)w(\theta_1)d\theta_1 \geq \int_{A'_{1,n}} e^{-L(x^n,\theta_1)}h(n,\theta_1)w(\theta_1)d\theta_1 \geq N(n)\epsilon_0.$$

The right-hand side goes to infinity, contradicting the bound (5) defining the MIL.

By the definition of h we have

$$\forall n, \quad \forall\theta_1 \in A_{1,n}, \quad \forall x^n \qquad \frac{p_1^*(x^n|\theta_1)}{m_1^*(x^n)} = e^{-L(x^n,\theta_1)}h(n,\theta_1) \leq B, \tag{A.7}$$

for some $B > 0$. So, by (A.6), (A.7), and the definition of the relative entropy we have that $\forall \theta_1 \in A_{1,n}$ and $\forall n$

$$P_1^*(S_1^c|\theta_1) \leq \frac{1}{C_1(n)I_1^*(n)} B \int \int p_1^*(x^n|\xi) w(\xi) \log \frac{p_1^*(x^n|\xi)}{m_1^*(x^n)} d\xi dx^n.$$

Since the numerator is bounded by $BI_1^*(n)$, the right-hand side is $B/C_1(n)$ which goes to zero.

Proof of Theorem 4.4: We prove the conclusion for models I and II and then sketch the proof for models I and III, and II and III. First, we expand $W_{(1)}(B|X^n, Y^n)$ and $W_{(2)}(B|X^n, Y^n)$ as $W(B)$ plus negligible error terms. For any $B \in \mathcal{B}$, we have for Model I

$$\left| \int_B \left\{ (w_{(1)}^*(\theta|Z^n) - w(\theta) \right\} d\theta \right| \leq \int |w_{(1)}^*(\theta|Z^n) - w(\theta)| d\theta \leq \frac{1}{\sqrt{2 \ln 2}} \sqrt{D\{w^*(\cdot|Z^n)||w(\cdot)\}},$$

which is bounded by $\sqrt{C_0(n)I_0^*(n)/2 \log 2}$, so that

$$|W_{(1)}(B|Z^n) - W(B)| \leq \frac{1}{\sqrt{2 \ln 2}} \sqrt{C_0(n)I_0^*(n)}. \tag{A.8}$$

The right-hand side is $o(1)$ as $n \to \infty$, for X^n, Y^n giving $Z^n \in S_0$. Similarly,

$$W_{(2)}(B|X^n, Y^n) = \int_B \frac{1}{2} \int w_1 \left(\frac{\phi + \theta}{2} \right) w_2 \left(\frac{\phi - \theta}{2} \right) d\phi d\theta + J_{2,1}(X^n) + J_{2,2}(X^n, Y^n),$$

where the first term on the right-hand side is $W(B)$ and the error terms are

$$J_{2,1}(X^n) = \int_B \frac{1}{2} \int \left\{ w_1^* \left(\frac{\phi + \theta}{2} | X^n \right) - w_1 \left(\frac{\phi + \theta}{2} \right) \right\} w_2 \left(\frac{\phi - \theta}{2} \right) d\phi d\theta,$$

$$J_{2,2}(X^n, Y^n) = \int_B \frac{1}{2} \int w_1^* \left(\frac{\phi + \theta}{2} | X^n \right) \left\{ (w_2^* \left(\frac{\phi - \theta}{2} | Y^n \right) - w_2 \left(\frac{\phi - \theta}{2} \right) \right\} d\phi d\theta.$$

They arise by adding and subtracting $W(B)$ and $\int_B \frac{1}{2} \int w_1^*(\frac{\phi+\theta}{2}|X^n) w_2(\frac{\phi-\theta}{2}) d\phi d\theta$. The median point theorem of integration and the canonicality of the data give bounds on the error terms. We have

$$|J_{2,1}(X^n)| \leq \int \int |w_1^*(\theta_1|X^n) - w_1(\theta_1)| w_2(\theta_2) d\theta_1 d\theta_2 \leq \frac{1}{\sqrt{2 \ln 2}} \sqrt{C_1(n)I_1^*(n)}, \tag{A.9}$$

since the relative entropy dominates the square of L^1 distance (Csiszar, 1967). Thus, $J_{2,1}(X^n)$ tends to zero as n tends to infinity. By the same reasoning we get

$$|J_{2,2}(X^n, Y^n)| \leq \frac{1}{\sqrt{2 \ln 2}} \sqrt{C_2(n)I_2^*(n)}. \tag{A.10}$$

So, $J_{2,2}(X^n, Y^n)$ tends to zero as $n \to \infty$. By (A.8), (A.9) and (A.10), if we set $C_{1,2} = (1/\sqrt{2\ln 2})\left\{\sqrt{C_0(n)I_0^*(n)} + \sqrt{C_1(n)I_1^*(n)} + \sqrt{C_2(n)I_2^*(n)}\right\}$, which goes to zero as n increases, we get Theorem 4.4 for Models I and II.

Now we sketch the conclusion for models I and III, II and III. First write $W_{(3)}(B|Z^n)$ as $W(B)$ plus an error term, and note $w_{(3)}(\theta|Z^n)$ equals

$$\frac{1}{4}\int\int w_1^*\left(\frac{\phi+\theta}{2}|\frac{s^n+Z^n}{2}\right) w_2^*\left(\frac{\phi-\theta}{2}|\frac{s^n-Z^n}{2}\right)\frac{m_1^*(\frac{s^n+Z^n}{2})m_2^*(\frac{s^n-Z^n}{2})}{m_{(3)}^*(Z^n)}ds^n d\phi, \quad (A.11)$$

where

$$m_1^*\left(\frac{s^n+Z^n}{2}\right) = \int\int p_1^*\left(\frac{s^n+Z^n}{2}|\xi_1\right) w_1(\xi_1)d\xi_1,$$
$$m_2^*\left(\frac{s^n-Z^n}{2}\right) = \int\int p_2^*\left(\frac{s^n-Z^n}{2}|\xi_2\right) w_2(\xi_2)d\xi_2,$$

and

$$m_{(3)}^*(Z^n) = \frac{1}{2}\int m_1^*\left(\frac{s^n+Z^n}{2}\right) m_2^*\left(\frac{s^n-Z^n}{2}\right) ds^n.$$

By integrating over s^n for fixed $Z^n = z^n$ it is seen that

$$h(s^n, Z^n) = \frac{1}{2}\frac{m_1^*(\frac{s^n+Z^n}{2})m_2^*(\frac{s^n-Z^n}{2})}{m_{(3)}^*(Z^n)},$$

is a probability density. Now integrating (A.11) over B, and adding and subtracting $\int\int_B \frac{1}{2}\int w_1^*(\frac{\phi+\theta}{2}|\frac{s^n+Z^n}{2})w_2(\frac{\phi-\theta}{2})h(s^n, Z^n)d\phi d\theta ds^n$ and $W(B)$, gives $W_{(3)}(B|Z^n)$ is

$$W(B) + \int J_{2,1}\left(\frac{s^n+Z^n}{2}\right)h(s^n, Z^n)ds^n + \int J_{2,2}\left(\frac{s^n+Z^n}{2}, \frac{s^n-Z^n}{2}\right)h(s^n, Z^n)ds^n.$$

By the definition of S_1 in Proposition 4.3, the first term on the right-hand side is bounded by

$$\left|\int_{2S_1-Z^n}\int J_{2,1}\left(\frac{s^n+Z^n}{2}\right)h(s^n, Z^n)ds^n\right| + \left|\int_{2S_1^c-Z^n}\int J_{2,1}\left(\frac{s^n+Z^n}{2}\right)h(s^n, Z^n)ds^n\right|,$$

where $2S_1 - Z^n$ is the set of all s^n's, for fixed $Z^n = z^n$, such that $(s^n + Z^n)/2 \in S_1$. As in (A.9), the first term in the right-hand side above is bounded by $\sqrt{C_1(n)I_1^*(n)/2\ln 2}$. For the second term on the right-hand side, note that $J_{2,1}$ is bounded and $h(\cdot, Z^n)$ is a density, so Proposition 4.3 asserts $P_1^*(2S_1^c - Z^n)$ tends to zero as $n \to \infty$. Thus, the second term on the right-hand side is $o_{P_1^*}(1)$.

Similarly, the third term on the right-hand side, involving $J_{2,2}$, is bounded by

$$\left|\int_{2S_2+Z^n} J_{2,2}\left(\frac{s^n+Z^n}{2}, \frac{s^n-Z^n}{2}\right)h(s^n, Z^n)ds^n\right|$$

$$+\left|\int_{2S_2^c+Z^n} J_{2,2}\left(\frac{s^n+Z^n}{2},\frac{s^n-Z^n}{2}\right)h(s^n,Z^n)ds^n\right|,$$

by (A.10). The first term above is bounded by $\sqrt{C_1(n)I_1^*(n)/2\ln 2}$. For the second term, the argument is nearly as before: $J_{2,1}(\cdot,\cdot)$ is bounded and $h(\cdot,Z^n)$ is a density, and Proposition 4.3 implies the set $2S_2^c+Z^n$ is $o_{p_2^*}(1)$, as is the second term in the right-hand side above. Thus we have

$$|W_{(3)}(B|Z^n)-W(B)|\le b(n)+o_{p_1^*}(1)+o_{p_2^*}(1), \tag{A.12}$$

where $b(n)\to 0$ as $n\to\infty$. By (A.8) and the last bound, we get the conclusion for Models I and III. By the last bound and the analogous bound (from (A.8), (A.9), and (A.10) as before) we get the conclusion for Models II and III.

Proof of Proposition 4.5: The marginal density for Z^n, $m(z^n)$, is obtained from $m_1(x^n)m_2(y^n)$ by the transformation $Z^n=X^n-Y^n$, and $S^n=X^n+Y^n$, and integrating out s^n. Now

$$m(z^n)=\frac{1}{2}\int m_1\left(\frac{s^n+Z^n}{2}\right)m_2\left(\frac{s^n-Z^n}{2}\right)ds^n,$$

the same as the marginal density for model 3, $m_{(3)}(z^n)$. So, the left-hand side of (11) is

$$E_{(X^n,Y^n)}\int\left\{\int\frac{1}{2}w_1\left(\frac{\phi+\theta}{2}|X^n\right)w_2\left(\frac{\phi-\theta}{2}|Y^n\right)d\phi\right\}\times$$

$$\log\frac{\int\frac{1}{4m_{(3)}(Z^n)}\int w_1(\frac{\phi+\theta}{2}|\frac{s^n+Z^n}{2})w_2(\frac{\phi-\theta}{2}|\frac{s^n-Z^n}{2})m_1(\frac{s^n+Z^n}{2})m_2(\frac{s^n-Z^n}{2})d\phi ds^n}{w_{(1)}(\theta|Z^n)}d\theta.$$

Jensen's inequality gives the log sum inequality (see Cover and Thomas, 1991, p. 29). Using it to bound the right-hand side from below gives a quantity that is strictly positive outside of the degenerate case which we have ruled out. Thus, (11) follows, and the proof of (12) is similar.

The proof of (13) is harder. When $w_{(1)}$ is bounded above and $w_{(2)}$ is bounded below from zero we define

$$v=\sup_{(\theta,x^n,y^n)}\frac{w_{(1)}(\theta|z^n)}{w_{(2)}(\theta|x^n,y^n)},$$

$$A(x^n,y^n)=\left\{\theta:\log\frac{w_{(2)}(\theta|x^n,y^n)}{w_{(3)}(\theta|z^n)}<0\right\},$$

and

$$\overline{w}_{(k)}(\theta|X^n,Y^n)=w_{(k)}(\theta|X^n,Y^n)\chi_{A(X^n,Y^n)}(\theta),$$

for $k = 2,\ 3$. Now, combining the terms in the logarithm in (13), multiplying and dividing by $\overline{w}_{(2)}$, bounding from below by v and cutting the domain of integration down to $A(x^n, y^n)$ gives

$$v \int E_{(X^n,Y^n)} \left[\overline{w}_{(2)}(\theta|X^n, Y^n) \log \left\{ \frac{\overline{w}_{(2)}(\theta|X^n, Y^n)}{\overline{w}_{(3)}(\theta|X^n, Y^n)} \right\} \right] d\theta \qquad (A.13)$$

as a lower bound for the left-hand side of (13). The finiteness of

$$E_{(X^n,Y^n)} D\{w_{(2)}(\cdot|X^n, Y^n)||w_{(3)}(\cdot|X^n, Y^n)\}$$

ensures that $E_{(X^n,Y^n)} D\{\overline{w}_{(2)}(\cdot|X^n, Y^n)||\overline{w}_{(3)}(\cdot|X^n, Y^n)\}$ is finite also. By a discretization argument we can set up an application of the log-sum inequality. This gives the new lower bound

$$v \int \left[E_{(X^n,Y^n)} \overline{w}_{(2)}(\theta|X^n, Y^n) \log \left\{ \frac{E_{(X^n,Y^n)} \overline{w}_{(2)}(\theta|X^n, Y^n)}{E_{(X^n,Y^n)} \overline{w}_{(3)}(\theta|X^n, Y^n)} \right\} \right] d\theta - \epsilon,$$

for (A.13). Explicitly writing out the expectations in the argument of the logarithm shows they cancel so that the first term in zero. Since ϵ is arbitrary the proof is complete.

Acknowledgements. The authors would like to express their gratitude to the referees, the Associate Editor, Professors Nancy Reid and Christian Genest for their thoughtful and constructive suggestions on how to improve our paper. We are in their debt. This work was partially supported by the Natural Sciences and Engineering Research Council of Canada.

References

Arimoto, S. (1972). An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Trans. Inform. Theory*, IT-18, No.1, 14-20.

Berger, T. (1971). *Rate Distortion Theory.* Prentice-Hall, Englewood Cliffs, NJ.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. Ser. B*, 41, 113-147.

Blahut, R. E. (1972a). Computation of channel capacity and rate-distortion functions. *IEEE Trans. Inform. Theory*, IT-18, No. 4, 460–473.

Blahut, R. E. (1972b). *An Hypothesis Testing Approach to Information Theory.* Ph.D. Thesis, Cornell University, Ithaca, NY.

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

Clarke, B. S. and Barron A. R. (1994). Jeffreys' prior is asymptotically least favorable under entropy risk. *J. Statist. Plann. Inf.*, 41, 37-60.

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

Csiszar I. (1967). On topological properties of f-divergences. *Studia Sci. Math. Hungar.*, 2, 329-339.

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

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

Efroimovich, S. Yu. (1980). Information contained in a sequence of observations. *Problems Inform. Transmission* 15, 178-189.

Ibragimov, I. A. and Hasminsky, R. Z. (1973). *Statistical Estimation: Asymptotic Theory.* Springer-Verlag, New York.

Jaynes, E. T. (1982). On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70, No. 9, 939–952.

Nader, M. A. and Reboussin, D. M. (1994). The effects of behavioral history on cocaine self-administration by rhesus monkeys. *Psychopharmacology,* 115, 53-58.

Soofi, E. S., Ebrahimi N. and Habibullah M. (1995). Information distinguishability with application to analysis of failure data. *J. Amer. Statist. Assoc.* 90, 657–668.

Yuan, A. (1997). *A Minimally Informative Likelihood Approach to Bayesian Inference and Decision Analysis.* Ph. D. Thesis, Department of Statistics, University of British Columbia, Vancouver, Canada.

Yuan, A. and Clarke, B. S. (1999). An information criterion for likelihood selection. *IEEE Trans. Inform. Theory*, 45, No.2, 1-10.

Ao Yuan
Statistics Research Laboratory
Department of Anesthesia
Massachusetts General Hospital
32 Fruit Street, Clinics 3
Boston, MA 02114
email: yuan@srlb4.mgh.harvard.edu

Bertrand Clarke
Department of Statistics
University of British Columbia
6357 Agricultural Road, Room 333
Vancouver, B.C., Canada V6T 1Z2
email: bertrand@stat.ubc.ca

Figure Captions

Figure 1. Posteriors densities from MIL's. The posteriors plotted in (a) are from Model I using MIL's with w chosen to be $Normal(0, 1)$, L to be squared error loss, and $\lambda = .7$ (bold) or $\lambda = 1.5$ (solid). The posteriors plotted in (b) are from Model II using MIL's with $w_1 = w_2$ chosen to be $Normal(0, 1)$, L to be squared error loss, and $\lambda_1 = \lambda_2 = .0001$ (dots), or $\lambda_1 = \lambda_2 = .09$ (solid). Note that the posteriors assign most of their mass to the positive half line.

Figure 2. Posteriors for Model I Using a Uniform Prior. The posteriors plotted here were formed from MIL's using a $Uniform[-10, 15]$ prior, squared error loss, and $\lambda = .5$ (dots), or $\lambda = 5$ (solid). Despite the change in prior from those in Fig. 1, the posteriors continue to assign essentially all their mass to the positive halfline.

Figure 3. Posteriors for Model I Using Absolute Value Loss. The posteriors plotted here were formed from MIL's using a $Normal(0, 1)$ prior, $\lambda = .5$ (dots), $\lambda = 2$ (solid), and choosing L to be absolute value loss. The shape of the posteriors near their peak is different from those in Figs. 1 and 3, because the absolute value function is not differentiable. However, mass is still assigned essentially only to the positive half line.