A MINIMALLY INFORMATIVE LIKELIHOOD APPROACH TO BAYESIAN INFERENCE AND DECISION ANALYSIS

by

AO YUAN

B.Sc., Sichuan University, China, 1982

M.Sc., Sichuan University, China, 1989

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

DEPARTMENT OF STATISTICS

We accept this thesis as conforming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

1997

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Statistics

The University of British Columbia
Vancouver, Canada

Date June 16, 1996

DE-6 (2/88)

A Minimally Informative Likelihood Approach to Bayesian Inference and Decision Analysis

Abstract

For a given prior density, we minimize the Shannon Mutual Information between a parameter and the data, over a class of likelihoods defined by bounding a Bayes risk by a 'distortion parameter'. This gives a conditional distribution for the data given the parameter which provides optimal data compression, or equivalently, is minimally informative for a type of location parameter. These optimal likelihoods cannot, in general, be obtained in closed form. However, they can be found numerically. Moreover, we give two statistical senses in which the optimal likelihoods form parametric families which make the weakest possible assumptions on the data generating mechanism. In addition, we establish properties of this parametric family that characterize its behavior as the distortion parameter varies. We argue that the parametric families identified here may lead to a default technique for some settings in initial data analysis. We partially characterize the settings in which our techniques may be expected to provide useful answers. In particular, we argue that if one is interested in performing certain Bayesian hypothesis tests on a parameter that locates a typical region for the response, then our technique may provide weak but nevertheless useful inferences.

We also investigated the robustness of inferences to modeling strategies for paired, blocked data.

Contents

List of Tables

List of Figures

Basic Notations and Definitions

The following notations and definitions are used throughout the thesis.

1. x^n stands for $(x_1, ..., x_n)$.

Acknowledgements

I would like to thank my supervisor Bertrand S. Clarke for guiding me to the unexplored territory, for his inspiration, for his constant encouragement and the financial support.

I am indebted to Professor Harry Joe for his invaluable advise and suggestions, to Nancy Heckman, Paul Gustafson and all the members of my Supervisory Committee for their comments and help. I would like to thank Christine Graham, our Department secretary, for her constant help.

Finally, I would like to thank the University of British Columbia and the Department of Statistics for its financial support.

Chapter 1

Introduction

1.1 The Minimally Informative Likelihood Problem

In this thesis we investigate an information theoretic criterion for likelihood selection. It is based on minimizing information: The information being minimized is the information implicit in the likelihood. This is counter-intuitive because usually one wants a likelihood which is as informative as possible. However, it must be remembered that fundamentally the likelihood is as arbitrary, at least initially, as any other statistical construct. More to the point, for the sake of being conservative, one wants to assume as little as possible because it is hard to assess whether the assumptions one has made are acceptable for the application. Indeed, it is an empirical fact that no models are demonstrably exact for any real phenomenon. In practice, if one has a genuinely valid parametric family then inferences made using it will likely be stronger than those made using a likelihood representing weaker assumptions. We formalize this by obtaining minimally informative likelihoods. Even though they provide weaker inferences, their conservatism is useful. For instance, rejecting a hypothesis using a minimally informative likelihood is a stronger form of rejection than rejecting under a true likelihood. Estimation using a minimally informative likelihood bypasses much of the argumentation essential to justifying a model – which is generally done cursorily at best, that is without reference to the detailed physical basis of the phenomenon which is often unknown.

The actual criterion we study is a minimization over a set of 'good' likelihoods defined by a Bayes risk bound. Thus, to use this criterion one must identify a quantity or parameter, choose a prior density for it, and choose a loss function. In addition, one must choose a bound for the Bayes risk. Specifically, the criterion we are to optimize is the Shannon mutual information (SMI) between the likelihood and the prior. i.e. the likelihood we seek

is

$$p^* = \arg\min_{p \in \mathcal{P}} I(\Theta, X^n)$$

where

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

is the Shannon mutual information between the random variable X^n and the parameter Θ,

$$m(x^n) = \int p(x^n|\theta) w(\theta) d\theta$$

is the data marginal density and

$$\mathcal{P} = \{ \ p(\cdot|\theta) : \int \int p(x^n|\theta) L(x^n, \theta) w(\theta) dx^n d\theta \leq l \ \}.$$

Here $L(\cdot, \cdot)$ is the loss function and l is the specified risk tolerance bound. Note the SMI is the expected Kullback-Leibler divergence between the posterior and the prior

$$I(\Theta, X^n) = E_m D(w(\cdot|X^n)||w(\cdot)),$$

where for any pair of densities $p(\cdot)$ and $q(\cdot)$,

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

is the Kullback-Leibler divergence between $p(\cdot)$ and $q(\cdot)$. It is not a distance, but has some distance-like properties. It measures the discrepancy between the two densities.

Thus, we are seeking the likelihood, in the given class, which updates the prior the least in that it gives a posterior as close to the prior as possible. In this sense, it is minimally informative for the parameter among the class of likelihoods. We call such a likelihood the Minimally Informative Likelihood (MIL). It is the most conservative, and in case no knowledge of the data distribution available, it can be used for an initial data analysis.

Since $I(\Theta, X^n) = H(\Theta) - H(\Theta|X^n)$, where H denotes the entropy, our method has some connections with the maximum entropy method. In fact, minimizing the SMI over likelihoods which we do here, is the same as maximizing the second term which is the conditional entropy, where the conditioning is on the data.

1.2 Formulation of the MIL Problem

As we noted earlier, in Bayesian analysis, usually one assumes a known likelihood and a known prior density, so that all inferences are obtained from the posterior distribution of the parameter. Sometimes, however, we have pre-experimental knowledge about the parameter that we can quantify in a prior distribution, but little knowledge about the likelihood or the distribution of the data conditionally on the parameter. For example, past experience with phenomena similar to a phenomenon under investigation, and expert opinion may be used to suggest a prior, but do not generally provide enough information to suggest a likelihood. In addition, modeling a physical problem may suggest a particular loss function, or it can be chosen based on convenience. We also require a bound on the allowable Bayes risk under that loss function; this can be set by the experimenter as we will discuss later. This bound controls the Bayes risk of using the data itself say $\mathbf{X}$, as an estimator for the unknown parameter θ. The likelihood our optimization produces therefore depends on the prior, the loss function and the bound; its robustness to these inputs can also be assessed. We will discuss the choice of these quantities in more detail in Section 1.2.3.

As a specific example, we might want to estimate the mean precipitation in a given month over the long term in British Columbia based on the daily observations in that month. Past records in B.C. can be used to help us to formulate the prior distribution. (If no such records were available, a Bayesian might just specify a set of basic beliefs about the parameter, say its location, dispersion, etc, and choose some standard distribution to fit these beliefs.) For a moderately small number of observations, if a phenomena which, like the weather, is not well understood, it may be hard to identify a reasonable family of distributions. So the MIL method while may sensitive to the physical details of a phenomena, will at least provides an option for the user who has limited knowledge. The loss function can be partially specified by the cost of under estimation and over estimation. For instance, excessive rainfall might lead to flooding or other damage to crops which can be assessed financially. Too little rainfall might necessitate irrigation which has an approximately known cost. The Bayes risk bound can be chosen by the experimenter according to the practical precision requirements; the larger the bound, the less accurate, but the more flexible the model will be. Or it can be a suitable positive number no greater than the minimum average loss over parameter, which

in the case of squared error loss, is just the prior variance, since by Proposition 3.1, there is a unique MIL for each Bayes risk bound in that range.

For another example, consider the life time x_i's of light bulbs, or any life time data. Let θ be the mean life time. Suppose we have some historical data that can be used to specify the prior. A reasonable loss function may be taken the form $L(x,\theta) = L_1(x,\theta)I(x > \theta) + L_2(x,\theta)I(x < \theta)$ where L_1 is non-negative function which is non-decreasing in $|x-\theta|$ and L_2 is non-positive which is non-increasing in $|x-\theta|$ and such that the Bayes risk is positive for a class of distributions. It is known that if L_1 and L_2 are linear in $|x-\theta|$, which may be reasonable choices in some practical settings, then the Bayes estimators are percentiles.

The parameter occupies an unorthodox role in this strategy. Conventionally, one specifies a prior fully and seeks a parametric family conditional on it. Of course this is hard. Instead, we specify the parameter partially: we regard it as a sort of location parameter in the sense that because we optimize over $\mathcal{P}$, θ must be estimable by $\mathbf{X}$. (This of course, is in addition to information theoretic criterion that θ must be decodable from $\mathbf{X}$.) As such, θ is partially specified and is only fully specified by the optimization procedure. In Example 1.4.3, for instance, we choose a $N(0,1)$ prior for θ that we think of as being the data mean. In fact, it is not exactly the data mean: once we have the MIL, we see that θ is interpreted as a shift of the mean. Examining the form of the MIL, we see that the loss function is in the exponent. Thus, the nature of the loss function also strongly affects the detailed interpretation of the parameter θ. This imprecision in the interpretation of θ will not in general hamper the assignment or elicitation of a prior.

In practice, it is difficult to explain the difference between a mean and a median and in our method this degree of exactitude is usually glossed over anyway. Moreover, it is rare to be able to assign a prior to one sort of location parameter without having obvious implications for similar but different location parameters.

In applications, one is concerned with what the parameter represents. Consider a hypothetical problem in which one person wants to estimate the median and another wants to estimate the 99th percentile using the same loss function and the same prior. Noting that the procedure gives the same MIL for both cases one is concerned that the basic set up doesn't make sense.

The answer to this criticism is that the set up does indeed make sense but that no uniform statistical interpretation of the parameter exists. That is, you get to choose a loss function, a measure of distortion and a prior but do not get to choose the statistical interpretation for the parameter; the exact statistical interpretation of the parameter arises from the optimization procedure. All that one case say in general is that as a consequence of the choice of loss and allowed distortion one will optimize over a class of likelihoods for which the random variable as an estimator for θ has Bayes risk bounded by the distortion. The parameter θ is a location parameter in the sense that it can be estimated by X with Bayes risk bounded by λ and the loss appears in the exponent of expression (1.2.1.3).

For instance, if one chooses the squared error loss and a normal prior one does get a normal MIL with an identified mean that is a function of λ, θ, and the parameters in the normal prior. In this case, the interpretation of θ as a percentile depends on the values of the parameters in the prior and on λ. As noted in Section 2, if $\mu = 0$, then only in the limit of $\lambda\sigma^2$ going to infinity does one get θ as the mean.

A limitation arises if one insists on using a certain parameter, loss and prior. In general one cannot link the choice of the loss function and the interpretation of the parameter. For example, suppose one insists on estimating the 99th percentile under the squared error loss. Even if one uses a prior appropriate for the 99th percentile, the exact interpretation for the parameter from our method depends on the loss function through (1.3.1). Since the 99th percentile is usually far from the posterior mean (the Bayes estimator under squared error loss) one expects the 99th percentile to differ substantially from the θ in (1.3.1). We conjecture the only way to remedy this is to change the loss function so its Bayes estimator is close to the interpretation one wants. Thus, the range of interpretations that can emerge from our method is narrow. In practice, one should put a prior on a vaguely defined location parameter after choosing a loss function. (This gives an idea of what type of θ's can be estimated well by $\mathbf{X}$.)

Fundamentally, we have a way to choose λ, the prior and the loss function to get a likelihood. i.e., we have a hyperplane in the space of likelihoods parametrized by $(\lambda, w(\cdot), L(\cdot,\cdot))$. This hyperplane is the result of an information theoretic optimization – a "universally" optimal reduction of the information in the sense of the data compression as described in Chapter 2. This formulation reverses the usual decision theory approach. In the usual ap-

proach one specifies a loss function and a parametric family, optimizing to find an estimator. Here, we have a loss function and an estimator ($\mathbf{X}$) but we optimize to find a parametric family.

1.2.1 Definition of the Minimally Informative Likelihood

To choose a likelihood, we note that for a given prior $w(\theta)$ and a parameter $\theta \in \mathbf{R}^d$, $I(\Theta, X)$ can be minimized over certain classes of likelihoods. This minimization gives a likelihood for which the posterior is least changed from the prior, in an average sense. This is one sense in which the optimal likelihood can be regarded as minimally informative so we denote it by $p_{MIL}(x|\theta)$. Consequently, a product of optimal univariate likelihoods is an independence model which is minimally informative apart from the assumption of independence.

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; the properties we use continue to hold. Let $L_n(x^n, \theta) = \sum_{i=1}^n L(x_i, \theta)$ be the cumulative empirical loss for estimating θ based on the sample x^n, where $L(\cdot, \cdot)$ is the loss function. We minimize the SMI over the class $\mathcal{P}_l$ of likelihoods defined to be the set of parametric families of densities on a measure space $(\mathcal{X}^n, \Theta)$ which satisfy

$$\int \int p(x^n|\theta) w(\theta) L_n(x^n, \theta) dx^n d\theta \leq l_n. \tag{1.2.1.1}$$

Here $l_n > 0$ bounds the amount of Bayes risk we will tolerate for estimating θ by X^n.

In information theory, the minimal value of the the SMI over $\mathcal{P}_l$, for the $n = 1$ case

$$R(l) = \inf_{p \in \mathcal{P}_l} I(\Theta, X), \tag{1.2.1.2}$$

is the rate distortion function, see Cover and Thomas (1991). It is shown in Blahut (1972b) that the minimum in (1.2.1.2) is achieved by

$$p_\lambda^*(x|\theta) = \frac{m^*(x) e^{-\lambda L(x,\theta)}}{\int m^*(y) e^{-\lambda L(y,\theta)} dy}, \tag{1.2.1.3}$$

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

$$\int \int p_\lambda^*(x|\theta) w(\theta) L(x, \theta) dx d\theta = l \tag{1.2.1.4}$$

and

$$\frac{1}{m^*(x)}\int p_\lambda^*(x|\theta)w(\theta)d\theta = \int \frac{e^{-\lambda L(x,\theta)}w(\theta)}{\int m^*(y)e^{-\lambda L(y,\theta)}dy}d\theta \leq 1 \qquad (1.2.1.5)$$

with equality in (1.2.1.5) for those x such that $m^*(x) > 0$.

The general approach of using MIL's as a default suggest models of the form

$$p(x|\theta,\lambda) = C(\theta,\lambda)m(x)e^{-\lambda L(x,\theta)},$$

where $m(x) \geq 0$ and $C(\theta,\lambda)$ is the normalizing constant. We may in turn ask the question: given the loss function $L(\cdot,\cdot)$ and $m(\cdot)$, does there exist $w(\cdot)$ such that $m(x)$ is the mixture of $p(x|\theta,\lambda)$ and $w(\theta)$:

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

This is an integral equation problem.

Note that $\lambda = 0$ in (1.2.1.3) is associated with $p_\lambda^*(x|\theta) = m^*(x)$ which is independent of θ. We will see later that the corresponding l, under suitable conditions, is infinity. In this sense, the constraint (1.2.1.1) vanishes and the SMI assumes its minimum of zero for any distribution that is independent of θ. When $n > 1$, the foregoing holds with x and $L(x,\theta)$ replaced by x^n and $L_n(x^n,\theta)$ respectively. Although we have taken x to be real valued, the procedure is valid more generally. In particular it is valid for x taking any values in R^k. It is this generality which will help permit the formulation of diverse models in Chapter 4.

In the definition of the MIL, $1/\lambda$ or l behaves like a dispersion parameter for $p_\lambda^*(\cdot|\theta)$ in addition to its role in defining $\mathcal{P}_l$. This will be discussed in an example in Section 1.4.

Apart from a few special cases, one cannot solve for the optimal $P_{MIL}(x|\theta) = p_\lambda^*(x|\theta)$ explicitly. However, one can obtain $p_\lambda^*(x|\theta)$ numerically by the Blahut-Arimoto algorithm, see Section 1.3.

This information theoretic technique produces a likelihood $p_\lambda^*(x|\theta)$ which is optimal within the class $\mathcal{P}_l$ of parametric families. It is optimal in the sense of making the weakest assumptions consistent with estimating by X with a Bayes risk bounded by l. Thus, in general $p_\lambda^*(x|\theta)$ is not a "true" likelihood. In particular, we require only that X be not a bad estimator for θ, where θ is a location-type parameter. It is a location only in the sense that it summarizes X in a data compression context or permits the effective decoding of X in a channel transmission context.

1.2.2 Minimally Informative Distributions

When considering λ as a parameter, the MIL is an optimal parametric family within a class of parametric families. Sometimes it is interesting and practical for a given prior density $w(\theta)$, to ask for an optimal distribution within a given parametric family $p(\cdot|\theta,\eta)$, where θ is the parameter of interest and η is an additional parameter, over which we are optimizing. In this case, we still assume the data distribution is *iid*. The SMI is a function of the additional parameter η:

$$I(\eta) = \int\int \prod_{i=1}^{n} p(x_i|\theta,\eta)w(\theta)\log\frac{\prod_{i=1}^{n} p(x_i|\theta,\eta)}{m(x^n|\eta)}d\theta dx^n,$$

$$= \int D(p^n_{\theta,\eta}(\cdot)||m_n(\cdot|\eta))w(\theta)d\theta = I(\Theta;X^n|\eta)$$

and $I(\Theta;X^n|\eta)$ is the conditional SMI. where

$$m(x^n|\eta) = \int \prod_{i=1}^{n} p(x_i|\theta,\eta)w(\theta)d\theta.$$

We are to minimize $I(\eta)$ over η. This will lead to more understanding of the behavior of the minimum information approach.

For large sample size n and any fixed value of η, we can use the approximation by B. Clarke and A. Barron (1990)

$$D(p^n(\cdot|\theta,\eta)||m_n(\cdot|\eta)) = \frac{d}{2}\log\frac{n}{2\pi e} + \log\frac{\sqrt{|\gamma(\theta|\eta)|}}{w(\theta)} + o(1), \qquad (1.2.2.1)$$

where d is the dimension of the parameter θ and $\gamma(\theta|\eta)$ is the Fisher information of the likelihood $p(x|\theta)$. So, by this formula, we can get an approximate minimally informative distribution, by asymptotically minimizing $\gamma(\theta|\eta)$ over η.

We denote the minimizer of $\gamma(\theta|\eta)$ by η^* and call the corresponding distribution $p(\cdot|\theta,\eta^*)$ the *minimally informative distribution* (MID). We will give some examples of MID's in Section 1.4.

1.2.3 The Quantities that Determine the MIL

Since the MIL requires a fixed prior $w(\cdot)$, a specified loss function $L(\cdot,\cdot)$ and a given Bayes risk bound l, we must specify these quantities before the construction of the MIL. There

are numerous methods for selecting the prior. If we have historical data, it may be used to suggest a prior. If we have some vague knowledge of the likelihood, for example the Fisher information $\gamma(\theta)$, one can use Bernardo's reference prior which is based on $\gamma(\theta)$ (see Bernardo, 1979). In practice, to specify a prior distribution, one usually first specifies some basic beliefs about the prior, for example its range, location, dispersion, and then chooses some standard distribution to meet these constraints.

Also, if the belief in the prior distribution is not strong, we may choose priors sequentially. That is let $w_0(\cdot)$ be the initial prior which may be very flat, get $p_0^*(\cdot|\cdot)$ from it by the MIL procedure; use $w_1(\cdot) \propto p_0^*(data|\cdot)w_0(\cdot)$ as the next stage prior and so on. Or divide the data into two parts $\mathbf{x} = (\mathbf{y}, \mathbf{z})$, use $\mathbf{y}$ as a training set to get $w(\theta|\mathbf{y})$, treating it as the prior and $\mathbf{z}$ as the data for inference.

The loss function can be chosen in several ways. The best way is to use an experimenter's understanding to formulate L so that the loss is a matter of modeling. In practice, however, one often chooses a loss function subjectively or for mathematical convenience. For a single continuous observation, the usual choice is the squared error loss or absolute error loss. For the binary case one may choose the 0-1 loss. For a sample of size n, we may use the average or weighted average loss for each observation.

As for the Bayes risk bound l, it may be chosen based on how much risk is tolerable for the specific problem, or chosen informally in the same way for an optimal smoothing parameter. (We will see later that l behaves much like a smoothing parameter in the nonparametric context.)

A practical and simple way to choose a proper value for the parameter λ is to find, possibly by grid search, the λ_0 based on which the corresponding posterior updated by $p_{\lambda_0}^*$ is closest, in the Kullback-Leibler sense, for instance, to the prior. This is consistent to the ideal with the MIL.

Since l and λ, in general, determine each other (see (ii) of Theorem 3.3.2), sometimes choosing λ is more convenient. Since λ and l have a sort of reciprocal relationship, we may roughly choose $\lambda \propto 1/l$. From the structure of the MIL, we see that λ^{-1} behaves somewhat like the dispersion of the distribution, thus, as another alternative we may choose $\lambda \approx 1/(Q(0.75) - Q(0.25))$, where Q(0.25) and Q(0.75) are the first and third sample quartile respectively. In the example computed in Chapter 4 we choose λ on the basis of how the

posteriors look – a heuristic approach which we find more convincing that the automatic techniques we have listed.

1.3 The Blahut-Arimoto Iterative Procedure

In Section 1.2.1 we stated that in general, there is no closed form solution for the MIL's, but it is computationally possible to obtain MIL's through an iterative procedure, the Blahut-Arimoto algorithm. Recall that the MIL is the minimizer of the rate-distortion function $R(l)$ for given distortion or Bayes risk l. It is shown in Blahut (1972), that the minimum in $R(l)$ is achieved uniquely by

$$p^*(x|\theta) = \frac{m^*(x)e^{-\lambda L(x,\theta)}}{\int m^*(y)e^{-\lambda L(y,\theta)}dy} \tag{1.3.1}$$

where $m^*(x)$ is determined by the equation

$$\int \frac{e^{-\lambda L(x,\theta)}w(\theta)}{\int m^*(y)e^{-\lambda L(y,\theta)}dy}d\theta \leq 1 \tag{1.3.2}$$

with equality for those x such that $m^*(x) > 0$, and $\lambda \geq 0$ is determined by l through the equality in the constraint (1.2.1). Note that $m^*(x)$ is just the marginal density for the data from $p^*(x|\theta)$: $m^*(x) = \int p^*(x|\theta)w(\theta)d\theta$.

The following three theorems are given in Blahut (1987). For self containment, we state them here and give an outline of the proof. The technique of proof is valid when X_i is discrete or continuous. Let

$$\mathcal{Q} = \{q : q(\cdot) \text{ is a probability density on } \mathcal{X}\},$$

and

$$\mathcal{R} = \{r(\cdot|\cdot) : \forall x, \ r(\theta|x) \text{ is a posterior density on } \Theta\}.$$

We have

Theorem 1.3.1 (Blahut, 1987).

$$(i) \qquad I(\Theta, X) = \inf_{q\in\mathcal{Q}} \int w(\theta)p(x|\theta)\log\frac{p(x|\theta)}{q(x)}d\theta dx.$$

$$(ii) \qquad I(\Theta, X) = \sup_{r \in \mathcal{R}} \int w(\theta) p(x|\theta) \log \frac{r(\theta|x)}{w(\theta)} d\theta dx.$$

Theorem 1.3.2 (Blahut, 1987). $R(l)$ is decreasing on $[0, \infty)$ and is convex and hence continuous on $[0, \underline{r}]$, where

$$\underline{r} = \inf_x \int w(\theta) L(x, \theta) d\theta.$$

Theorem 1.3.1 says that the inequality constraints in the definition of $R(l)$ can be replaced by an equality constraint, since $R(l)$ is deceasing and continuous. This is significant because it means that we can use the equality constraint to introduce the Lagrange multiplier. That is, consider

$$R(l, \lambda) = \inf_p \Big[\int \int w(\theta) p(x|\theta) \log \frac{p(x|\theta)}{\int w(\xi) p(x|\xi) d\xi} dx d\theta$$
$$+ \lambda \Big(\int \int w(\theta) p(x|\theta) L(x, \theta) dx d\theta - l \Big) \Big],$$

for each λ. The minimum of this expression will be achieved by some p^*_λ, which depends on λ, and λ is chosen so that $\int w(\theta) p^*_\lambda(x|\theta) L(x, \theta) dx d\theta = l$. Thus we have the following:

Theorem 1.3.3 (Blahut, 1987).

$$R(l) = -\lambda l + \inf_m \inf_p \Big[\int \int w(\theta) p(x|\theta) \log \frac{p(x|\theta)}{m(x)} d\theta dx + \lambda \int \int w(\theta) p(x|\theta) L(x, \theta) dx d\theta \Big],$$

where $m(\cdot)$ is a probability density.

For fixed $p(\cdot|\cdot)$, the expression in the square brackets is minimized by choosing

$$m(x) = \int w(\theta) p(x|\theta) d\theta.$$

For fixed $m(\cdot)$, the expression in the square brackets is minimized by choosing

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

For more details of proof, see Blahut (1972a, b).

Based on the above double minimization, the MIL can be evaluated by the following iterative procedure, see Blahut (1972a) and Arimoto (1972): choose an arbitrary density function $m_0(\cdot)$ and form $p_0(\cdot|\cdot)$ by setting

$$p_0(x|\theta) = \frac{m_0(x)e^{-\lambda L(x,\theta)}}{\int m_0(y)e^{-\lambda L(y,\theta)}dy}, \tag{1.3.3}$$

where λ is chosen to achieve the equality in (1.2.1.1). Then, form the next step $m_1(\cdot)$ marginal by

$$m_1(x) = \int w(\theta)p_0(x|\theta)d\theta, \tag{1.3.4}$$

and the next step $p_1(\cdot|\cdot)$ by replacing $m_0(\cdot)$ by $m_1(\cdot)$ in (1.3.3) and continue this fashion. After n step iteration, we get $p_n(\cdot|\cdot)$. Is was shown by Csiszar (1974), that

$$\lim_{n\to\infty} p_n(x|\theta) = p^*(x|\theta), \forall x \text{ and } \theta. \tag{1.3.5}$$

1.4 Some Closed Form Examples

Except a few special cases, the rate distortion function $D(l)$ and its corresponding minimizer $p^*(x|\theta)$ cannot be evaluated in closed form, but can, in general, be carried out by the Blahut-Arimoto iterative procedure. Here we show several examples, some for the discrete case and some for the continuous case, in which the MIL $p^*(x|\theta)$ can be obtained in closed form.

Example 1.4.1. For a discrete one-dimensional example, we take the binary symmetric source, that is the prior $w(\theta)$ takes α and $1-\alpha$ for some $\alpha \in (0,1)$ respectively at 0 and 1, and zero at any other point. We take the loss to be the probability-of-error loss, that is $L(x,\theta) = 0$ for $x = \theta$, and 1 for $x \neq \theta$. For $\theta = 0, 1$, $p(x|\theta)$ is a probability mass function for x on $\{0,1\}$. This example is used in information theory to illustrate how the rate-distortion bound can be achieved by the corresponding channel, the MIL in our context. It also shows that even in the simple situation, a closed form MIL solution is relatively difficult to obtain. For $l \in [0, \min(\alpha, 1-\alpha)]$ the constraint is

$$l = \sum_{i=0}^{1}\sum_{j=0}^{1} w(i)p(i|j)L(i,j) = (1-\alpha)p(0|1) + \alpha p(1|0) \tag{1.4.1}$$

The MIL in this case is (see Cover and Thomas, 1991)

$$p^*(x|0) = \begin{cases} \frac{(1-\alpha-l)(1-l)}{(1-\alpha)(1-2l)} & \text{if } x = 0, \\ \frac{l(\alpha-l)}{(1-\alpha)(1-2l)} & \text{if } x = 1, \end{cases} \qquad p^*(x|1) = \begin{cases} \frac{l(1-\alpha-l)}{\alpha(1-2l)} & \text{if } x = 0, \\ \frac{(1-l)(\alpha-l)}{\alpha(1-2l)} & \text{if } x = 1, \end{cases}$$

And the corresponding $m^*(\cdot)$ is

$$m^*(0) = \frac{1-\alpha-l}{1-2l}, \quad m^*(1) = \frac{\alpha-l}{1-2l},$$

so we can determine, for each $l \in [0, \min(\alpha, 1-\alpha)]$, the corresponding λ in the formula for $p^*(x|\theta)$ by the relationship

$$p^*(0|0) = \frac{m^*(0)}{m^*(0) + m^*(1)e^{-\lambda}},$$

which gives $\lambda = \ln \frac{1-l}{l}$. $I_{p^*}(X, \Theta)$ is a decreasing function of l, since for larger l, the class $\mathcal{P}_l$ is larger, and hence the infimum over the class is smaller. For $l = 0$, the corresponding $I_{p^*}(X, \Theta)$ achieves its maximum value $H(\Theta)$, the entropy of Θ. For l larger than $\min(\alpha, 1-\alpha)$, the corresponding SMI is zero and is achieved by any distribution which is independent of θ (see Cover and Thomas, 1991).

Example 1.4.2. This is a discrete two-dimensional example. We use it to examine dependence in an MIL based on generalizing the last example. We will see that here, the dependence is not so high that the two random variables represent the same information. The prior $w(\theta)$ is the same as in Example 1.4.1. Let the loss function be $L_2(x_1, x_2, \theta) = L(x_1, \theta) + L(x_2, \theta)$, where $L(\cdot, \cdot)$ is the same as in Example 1.4.1. By (ii) of Proposition 3.1, the MIL is permutation symmetric in x_1 and x_2, so $p^*_\lambda(0, 1|0) = p^*_\lambda(1, 0|0)$ and $p^*_\lambda(1, 0|1) = p^*_\lambda(0, 1|1)$. Now, constraint (1.2.1.4) is

$$l = 2\Big(\alpha p^*_\lambda(0, 1|0) + \alpha p^*_\lambda(1, 1|0) + (1-\alpha)p^*_\lambda(0, 0|1) + (1-\alpha)p^*_\lambda(0, 1|1)\Big). \tag{1.4.2}$$

To get the MIL, we first find the corresponding m^*. For this, first we note that $m^*(x_1, x_2)$ is also permutation symmetric in its two arguments. Let $\beta_1 = m^*(0, 0), \beta_2 = m^*(0, 1)$ and $\beta_3 = m^*(1, 1)$. In (1.2.1.5) we take $x_1 = x_2 = 0, x_1 = x_2 = 1$ and $x_1 = 1, x_2 = 0$ (or $x_1 = 0, x_2 = 1$) in turn to get the following three equations

$$\frac{\alpha}{\beta_1 + 2e^{-\lambda}\beta_2 + e^{-2\lambda}\beta_3} + \frac{(1-\alpha)e^{-2\lambda}}{e^{-2\lambda}\beta_1 + 2e^{-\lambda}\beta_2 + \beta_3} = 1 \tag{1.4.3},$$

$$\frac{\alpha e^{-2\lambda}}{\beta_1 + 2e^{-\lambda}\beta_2 + e^{-2\lambda}\beta_3} + \frac{1-\alpha}{e^{-2\lambda}\beta_1 + 2e^{-\lambda}\beta_2 + \beta_3} = 1 \tag{1.4.4}$$

and

$$\frac{\alpha e^{-\lambda}}{\beta_1 + 2e^{-\lambda}\beta_2 + e^{-2\lambda}\beta_3} + \frac{(1-\alpha)e^{-\lambda}}{e^{-2\lambda}\beta_1 + 2e^{-\lambda}\beta_2 + \beta_3} = 1. \tag{1.4.5}$$

Multiply both sides of these equations respectively by

$$(\beta_1 + 2e^{-\lambda}\beta_2 + e^{-2\lambda}\beta_3)(e^{-2\lambda}\beta_1 + 2e^{-\lambda}\beta_2 + \beta_3)$$

we get

$$(\beta_1 + 2e^{-\lambda}\beta_2 + e^{-2\lambda}\beta_3)(e^{-2\lambda}\beta_1 + 2e^{-\lambda}\beta_2 + \beta_3)$$

$$= e^{-2\lambda}\beta_1 + 2[\alpha e^{-\lambda} + (1-\alpha)e^{-3\lambda}]\beta_2 + [\alpha + (1-\alpha)e^{-4\lambda}]\beta_3, \tag{1.4.6}$$

$$(\beta_1 + 2e^{-\lambda}\beta_2 + e^{-2\lambda}\beta_3)(e^{-2\lambda}\beta_1 + 2e^{-\lambda}\beta_2 + \beta_3)$$

$$= [(1-\alpha) + \alpha e^{-4\alpha}]\beta_1 + 2[\alpha e^{-3\lambda} + (1-\alpha)e^{-\lambda}]\beta_2 + e^{-2\lambda}\beta_3 \tag{1.4.7}$$

and

$$(\beta_1 + 2e^{-\lambda}\beta_2 + e^{-2\lambda}\beta_3)(e^{-2\lambda}\beta_1 + 2e^{-\lambda}\beta_2 + \beta_3)$$

$$= [(1-\alpha)e^{-\lambda} + \alpha e^{-3\lambda}]\beta_1 + 2(1-\alpha)e^{-2\lambda}\beta_2 + [\alpha e^{-\lambda} + (1-\alpha)e^{-3\lambda}]\beta_3. \tag{1.4.8}$$

Let

$$\begin{aligned}
a_1 &= 1 - \alpha - e^{-2\lambda} + \alpha e^{-4\lambda}, \\
a_2 &= 2[(1-2\alpha)e^{-\lambda} + (2\alpha - 1)e^{-3\lambda}], \\
a_3 &= 1 + \alpha - e^{-2\lambda} - \alpha e^{-4\lambda}, \\
b_1 &= -(1-\alpha) + (1-\alpha)e^{-\lambda} + \alpha e^{-3\lambda} - \alpha e^{-4\lambda}, \\
b_2 &= 2[-(1-\alpha)e^{-\lambda} - \alpha e^{-3\lambda} + (1-\alpha)e^{-2\lambda}
\end{aligned}$$

and

$$b_3 = -\alpha e^{-\lambda} + e^{-2\lambda} - (1-\alpha)e^{-3\lambda}.$$

Subtracting (1.4.6) from (1.4.7) and (1.4.7) from (1.4.8), we get respectively

$$a_1\beta_1 + a_2\beta_2 = a_3\beta_3$$

$$b_1\beta_1 + b_2\beta_2 = b_3\beta_3.$$

Thus we get

$$\beta_1 = \frac{a_3b_2 - a_2b_3}{a_1b_2 - a_2b_1}\beta_3, \quad \beta_2 = -\frac{a_3b_1 - a_1b_3}{a_1b_2 - a_2b_1}\beta_3.$$

Also, by the relationship $\beta_1 + 2\beta_2 + \beta_3 = 1$, we get

$$\beta_3 = \frac{a_1b_2 - a_2b_1}{a_1(b_2 + 2b_3) - a_2(b_1 + b_3) + a_3(b_2 - 2b_1)}.$$

Plugging the values of β_1, β_2 and β_3 into (1.4.2) we choose the value of λ to satisfy the constraint for some values of l. Thus we can specify the β_i's completely and by (1.2.1.3) get the MIL $p^*_\lambda(x_1, x_2|\theta)$ as

$$p^*_\lambda(0,0|0) = \beta_1/C(0), \qquad p^*_\lambda(0,1|0) = p^*_\lambda(1,0|0) = \beta_2 e^{-\lambda}/C(0),$$

$$p^*_\lambda(1,1|0) = \beta_3 e^{-2\lambda}/C(0), \qquad p^*_\lambda(0,0|1) = \beta_1 e^{-2\lambda}/C(1),$$

$$p^*_\lambda(0,1|1) = p^*_\lambda(1,0|1) = \beta_2 e^{-\lambda}/C(1), \qquad p^*(1,1|1) = \beta_3/C(1),$$

where

$$C(0) = \beta_1 + 2\beta_2 e^{-\lambda} + \beta_3 e^{-2\lambda}, \qquad C(1) = \beta_1 e^{-2\lambda} + 2\beta_2 e^{-\lambda} + \beta_3.$$

To investigate the dependence in $p^*(x_1, x_2|\theta)$, we can calculate the Pearson correlation coefficient between X_1 and X_2. Let $p^*_{\lambda,1}(x_1|\theta)$ and $p^*_{\lambda,2}(x_2|\theta)$ be the marginals of $p^*_\lambda(x_1, x_2|\theta)$. We get

$$p^*_{\lambda,1}(0|0) = p^*_{\lambda,2}(0|0) = \frac{\beta_1 + \beta_2 e^{-\lambda}}{C(0)}, \qquad p^*_{\lambda,1}(1|0) = p^*_{\lambda,2}(1|0) = \frac{\beta_2 e^{-\lambda} + \beta_3 e^{-2\lambda}}{C(0)},$$

$$\begin{aligned} \mathrm{Var}_{\theta=0}(X_1) &= \mathrm{Var}_{\theta=0}(X_2) = p^*_{\lambda,1}(1|0)(1 - p^*_{\lambda,1}(1|0)) \\ &= \frac{(\beta_2 e^{-\lambda} + \beta_3 e^{-2\lambda})(\beta_1 + \beta_2 e^{-\lambda})}{C^2(0)}, \end{aligned}$$

and

$$\begin{aligned} \mathrm{Cov}_{\theta=0}(X_1, X_2) &= p^*_\lambda(1,1|0) - p^*_{\lambda,1}(1|0)p^*_{\lambda,2}(1|0) \\ &= \frac{(\beta_1\beta_3 - \beta_2^2)e^{-2\lambda}}{(\beta_2 e^{-\lambda} + \beta_3 e^{-2\lambda})(\beta_1 + \beta_2 e^{-\lambda})}. \end{aligned}$$

So

$$\mathrm{Corr}_{\theta=0}(X_1, X_2) = \frac{e^{-2\lambda}(\beta_1\beta_3 - \beta_2^2)(\beta_1 + 2\beta_2 e^{-\lambda} + \beta_3 e^{-2\lambda})^2}{(\beta_2 e^{-\lambda} + \beta_3 e^{-2\lambda})^2(\beta_1 + \beta_2 e^{-\lambda})^2}.$$

Similarly,

$$\mathrm{Corr}_{\theta=1}(X_1, X_2) = \frac{e^{-2\lambda}(\beta_1\beta_3 - \beta_2^2)(\beta_1 e^{-2\lambda} + 2\beta_2 e^{-\lambda} + \beta_3)^2}{(\beta_2 e^{-\lambda} + \beta_3)^2(\beta_1 e^{-2\lambda} + \beta_2 e^{-\lambda})^2}.$$

We see that as $\lambda \to 0$,

$$\mathrm{Corr}_{\theta=0}(X_1, X_2) \quad \text{and} \quad \mathrm{Corr}_{\theta=1}(X_1, X_2) \to \frac{(\beta_1\beta_3 - \beta_2^2)(\beta_1 + 2\beta_2 + \beta_3)}{(\beta_2 + \beta_3)^2(\beta_1 + \beta_2)^2};$$

and that as $\lambda \to +\infty$,

$$\mathrm{Corr}_{\theta=0}(X_1, X_2) \quad \text{and} \quad \mathrm{Corr}_{\theta=1}(X_1, X_2) \to \frac{(\beta_1\beta_3 - \beta_2^2)}{\beta_2^2}.$$

Example 1.4.3. For a continuous one-dimensional MIL example, choose $w(\cdot)$ to be a $N(\mu, \sigma^2)$ density and L to be squared error loss. From the form of p^*, one expects that the minimally informative likelihood will be normal. Indeed, the maximum entropy distribution under a second moment constraint, is a normal, which is similar to the p^* here. This turns out to be the case subject to the restriction $l < \sigma^2$, i.e., the amount of Bayes risk 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 (1991, Chapter 13), so no unique solution exists. We see also that for this range of l, $l(\lambda) = 1/(2\lambda)$, so we get that λ must be greater than $1/(2\sigma^2)$. It will be seen that $m^*(\cdot)$ is $N(\mu, \sigma^2 - \frac{1}{2\lambda})$, and $p^*(\cdot|\theta)$ is $N((1 - \frac{1}{2\lambda\sigma^2})\theta + \frac{1}{2\lambda\sigma^2}\mu, \frac{1}{2\lambda}(1 - \frac{1}{2\lambda\sigma^2}))$, 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 determines p^*.

To identify $p^*(\cdot|\theta)$ and the relationship between the tolerable risk bound l and the Lagrange multiplier λ in this case, we use three steps.

Step 1. we identify the $m^*(\cdot)$ which satisfies the constraint.

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

$$\int \frac{e^{-\lambda(x-\theta)^2} w(\theta)}{\int m^*(y) e^{-\lambda(y-\theta)^2} dy} d\theta \leq 1 \tag{1.4.9}$$

with equality in (1.4.9) 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 $w(\theta) = C\exp\{-\frac{(\theta-\mu)^2}{2\sigma^2}\}$ 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

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

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

$$a^2 = \frac{\lambda}{2\lambda\sigma^2 - 1}, \qquad b = a\mu.$$

Thus we have

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

which is recognized as a $N(\mu, \sigma^2 - \frac{1}{2\lambda})$ density, and it satisfies (1.4.9).

Step 2. we identify the MIL $p^*(\cdot|\theta)$ in this case.

Now, the expression for p^* gives

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

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

After substituting for a and b, the last expression is seen to be a $N((1 - \frac{1}{2\lambda\sigma^2})\theta + \frac{1}{2\lambda\sigma^2}\mu, \frac{1}{2\lambda}(1 - \frac{1}{2\lambda\sigma^2}))$ density. Note that $E_{p^*_\lambda}(X|\theta)$ is not θ, it's a weighted average of μ and θ.

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 $\lambda \to \infty$ (or $l \to 0$), the family $\mathcal{P}_l$ shrinkages to a single member $\zeta(\theta)$, the degenerate distribution at θ. We see $p^*_\lambda(\cdot|\theta) \to \zeta(\theta)$, which is consistent with above reasoning. This provides a sense in which λ is also a smoothing parameter, ensuring that a minimally informative density does not just concentrate at the data points.

Also, we investigate the relationship between l and λ. From the constraint for the Bayes risk, we have

$$l(\lambda) = \int\int p^*_\lambda(x|\theta) w(\theta) L(x,\theta) dx d\theta = \frac{1}{2(a^2+\lambda)} + (\frac{a^2}{a^2+\lambda})^2\sigma^2 = \frac{1}{2\lambda}.$$

Lastly, simple computation gives the corresponding posterior $w^*(\theta|x)$ is $N(x, 1/(2\lambda)) = N(x, l)$ which is still in the same normal family as the prior, but with the prior mean and variance (μ, σ^2) been updated to (x, l), any other likelihood in the class $\mathcal{P}_l$ will update the prior more by the expected Kullback-Leibler measure.

Example 1.4.4. This example is also for a one-dimensional parameter however we consider the general n-dimensional data case to understand more about the structure of the MILs. If no data summarization is possible one can, in principle, use the dependence model $p^*(\cdot|\theta)$ to be derived shortly to form a posterior. It is seen that this $p^*(\cdot|\theta)$ bears a superficial resemblance to the intuition behind shrinkage estimators.

As in the last example we choose a standard normal prior $w(\cdot)$ and comment that our calculations can be extended to an arbitrary $N(\mu, \sigma^2)$. Consider the loss function

$$L(x^n, \theta) = (\frac{1}{n}\sum_{i=1}^{n} x_i - \theta)^2;$$

arguably $\tilde{L}(x^n, \theta) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \theta)^2$ is a more natural choice. However, it is difficult to obtain closed form results for $\tilde{L}(\cdot, \cdot)$. As before, even for $L(\cdot, \cdot)$, we can only obtain closed form expressions for selected values of λ. For ease of calculation, choose $\lambda = \lambda_n = (n+1)/2$ (this choice of λ makes the computation simpler and results in a closed form for the MIL; other choices of λ may not give a closed form expression for the MIL). We show that the marginal density for the data is an n-dimensional normal

$$m^*(\cdot) \sim N_n(\mathbf{0}, A_n^{-1}),$$

where $\mathbf{0}$ is the n-dimensional zero vector, and A_n is the variance-covariance matrix given by

$$A_n^{-1} = \frac{2\lambda}{n^2}\begin{pmatrix} n & -1 & \dots & -1 \\ -1 & n & \dots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \dots & n \end{pmatrix}.$$

It is seen that A_n is positive definite with determinant

$$|A_n^{-1}| = (\frac{2\lambda}{n^2})^n (n+1)^{n-1}.$$

The corresponding MIL is a n dimensional independent multivariate normal

$$p^*(\cdot|\theta) \sim N_n\Big(\frac{n}{n+1}\theta \mathbf{1}_n, \frac{n^2}{(n+1)^2} I_n\Big),$$

where $\mathbf{1}_n$ is a n-vector of 1's and I_n is the n dimensional identity matrix. Here symmetry is also achieved as predicted by Proposition 3.1. The bound on the Bayes risk in the constraint is

$$l(n) = \frac{n}{(n+1)^2} = \frac{1}{2\lambda_n}(1 - \frac{1}{2\lambda_n}).$$

In this case, the marginal density is a dependence model and the MIL is an independence model. We regard this case as unusual. The corresponding posterior $w^*(\theta|x^n)$ is $N(\overline{x}, \frac{1}{n+1})$, does not follow the result of Theorem 3.1.1, since the loss function is not of the form there. From this example, we see that the form of the loss function affects the MIL a lot.

To verify the forms of $m^*(\cdot,...,\cdot)$ and $p^*(\cdot,...,\cdot|\theta)$, first note that the support of $m^*(\cdot,...,\cdot)$ is the entire n-dimensional Euclidean space, so we must check that the equality in (1.2.1.5) holds for all vectors $(x_1,...,x_n)$. Using the conjectured form of $m^*(\cdot,...,\cdot)$ we have

$$\int m^*(y^n)e^{-\lambda L(y^n,\theta)}dy^n$$

$$= \frac{(\sqrt{2\lambda})^n(n+1)^{\frac{n-1}{2}}}{n^n(\sqrt{2\pi})^n}\int...\int \exp\Big]\{-\lambda\Big(n\sum_{i=1}^n(\frac{y_i}{n})^2 - \sum_{i\neq j}\frac{y_i}{n}\frac{y_j}{n}\Big)$$

$$-\lambda(\sum_{i=1}^n\frac{y_i}{n}-\theta)^2\Big\}dy_1...dy_n$$

$$= \frac{(\sqrt{2\lambda})^n(n+1)^{\frac{n-1}{2}}}{n^n(\sqrt{2\pi})^n}\int...\int \exp\Big\{-(n+1)\lambda\Big(\sum_{i=1}^n(\frac{y_i}{n}-\frac{\theta}{n+1})^2\Big)$$

$$-\lambda\frac{\theta^2}{n+1}\Big\}dy_1...dy_n = \frac{1}{\sqrt{n+1}}e^{-\lambda\frac{\theta^2}{n+1}}.$$

Choosing $\lambda = (n+1)/2$, we get

$$\int \frac{e^{-\lambda(\frac{1}{n}\sum_{i=1}^n x_i-\theta)^2}\frac{1}{\sqrt{2\pi}}e^{-\frac{\theta^2}{2}}}{\int\int m^*(y^n)e^{-\lambda(\frac{1}{n}\sum_{i=1}^n y_i-\theta)^2}dy_1...dy_n}d\theta$$

$$= \frac{\sqrt{n+1}}{\sqrt{2\pi}}\int e^{-\frac{n+1}{2}(\theta-\frac{1}{n}\sum_{i=1}^n x_i)^2}d\theta = 1, \quad \forall x_1,...,x_n,$$

because the prior density $w(\theta)$ cancels out. Thus (1.2.1.5) is satisfied.

Now, by (1.2.1.3),

$$\begin{aligned}
p^*(x^n|\theta) &= \frac{m^*(x^n)\exp\{-\lambda(\frac{1}{n}\sum_{i=1}^n x_i-\theta)^2\}}{\int...\int m^*(y_1,...y_n)\exp\{-\lambda(\frac{1}{n}\sum_{i=1}^n y_i-\theta)^2\}dy_1...dy_n} \\
&= \frac{(\sqrt{2\lambda})^n(n+1)^{\frac{n}{2}}}{n^n(\sqrt{2\pi})^n}e^{\theta^2/2}\exp\Big\{n\sum_{i=1}^n(\frac{x_i}{n})^2 - \sum_{i\neq j}\frac{x_i}{n}\frac{x_j}{n} \\
&\quad +\sum_{i=1}^n(\frac{x_i}{n})^2 + \sum_{i\neq j}\frac{x_i}{n}\frac{x_j}{n} - 2\sum_{i=1}^n\frac{x_i}{n}\theta+\theta^2\Big\} \\
&= \frac{(\sqrt{2\lambda})^n(n+1)^{\frac{n}{2}}}{n^n(\sqrt{2\pi})^n}e^{\theta^2/2}\exp\Big\{\lambda(n+1)\sum_{i=1}^n(\frac{x_i}{n}-\frac{\theta}{n+1})^2 - \frac{\lambda\theta^2}{n+1}\Big\} \\
&= \frac{(n+1)^n}{n^n(\sqrt{2\pi})^n}\exp\Big\{-\frac{1}{2}\frac{(n+1)^2}{n^2}\sum_{i=1}^n(x_i-\frac{n}{n+1}\theta)^2\Big\},
\end{aligned}$$

which is the claimed multivariate normal distribution.

Finally, we derive an expression for the bound on the Bayes risk. It is

$$l_n((n+1)/2) = \int \cdots \int p^*(x^n|\theta)L(x'\theta)w(\theta)dx^n d\theta$$

$$= \frac{(n+1)^n}{n^n(\sqrt{2\pi})^n}\frac{1}{\sqrt{2\pi}}\int \cdots \int \exp\Big\{-\frac{1}{2}\frac{(n+1)^2}{n^2}\sum_{i=1}^{n}(x_i - \frac{n}{n+1}\theta)^2\Big\}$$

$$(\frac{1}{n}\sum_{i=1}^{n}x_i - \theta)^2 \exp\{-\theta^2/2\}dx_1...dx_n d\theta$$

$$= \frac{(n+1)^n}{n^n(\sqrt{2\pi})^n}\frac{1}{\sqrt{2\pi}}\int \Big(\int \cdots \int \exp\Big\{-\frac{1}{2}\frac{(n+1)^2}{n^2}\sum_{i=1}^{n}(x_i - \frac{n}{n+1}\theta)^2\Big\}$$

$$\frac{1}{n^2}[\sum_{i=1}^{n}(x_i - \frac{n}{n+1}\theta) + \frac{n}{n+1}\theta]^2 dx_1...dx_n\Big)\exp\{-\theta^2/2\}d\theta$$

$$= \frac{(n+1)^n}{n^n(\sqrt{2\pi})^n}\frac{1}{\sqrt{2\pi}}\int \Big(\int \cdots \int \exp\Big\{-\frac{1}{2}\frac{(n+1)^2}{n^2}\sum_{i=1}^{n}(x_i - \frac{n}{n+1}\theta)^2\Big\}$$

$$\frac{1}{n^2}\Big[\sum_{i=1}^{n}(x_i - \frac{n}{n+1}\theta)^2 + \sum_{i\neq j}(x_i - \frac{n}{n+1}\theta)(x_j - \frac{n}{n+1}\theta)$$

$$+\frac{2n\theta}{n+1}\sum_{i=1}^{n}(x_i - \frac{n}{n+1}\theta)\Big]dx_1...dx_n\Big)\exp\{-\theta^2/2\}d\theta$$

$$= \frac{(n+1)^n}{n^n(\sqrt{2\pi})^n}\frac{1}{\sqrt{2\pi}}\int \Big(\int \cdots \int \exp\Big\{-\frac{1}{2}\frac{(n+1)^2}{n^2}\sum_{i=1}^{n}(x_i - \frac{n}{n+1}\theta)^2\Big\}$$

$$\frac{1}{n^2}\sum_{i=1}^{n}(x_i - \frac{n}{n+1}\theta)^2 dx_1...dx_n\Big)\exp\{-\theta^2/2\}d\theta$$

$$= \frac{n}{(n+1)^2}\frac{1}{\sqrt{2\pi}}\int e^{-\theta^2/2}d\theta = \frac{n}{(n+1)^2}.$$

Examples of MID. We discussed the minimally informative distributions (MID) in Section 1.2.2. They can be considered as of the MILs from different point of view, in which we restrict the class of distributions to be optimized be some specified two parameter (they can be vectors) distribution. One is a parameter of interest, the other, while not a nuisance parameter serves only as an index for optimization. The optimization gives the member of the parametric family closest to the prior in the Kullback-Leibler distance. In the examples following, we will see that in most of the cases, MIDs can be solved in closed forms, and the computation is usually easier than that of the MIL's. The parameter value which achieves

the minimum in the SMI can be viewed as the *minimally informative estimation* of the additional parameter, it is the most conservative initial guess of the additional parameter value.

Example 1.4.5. $w(\cdot) \sim N(0,1)$, $p(\cdot|\theta,\eta) \sim N(\theta,\eta^2)$. In this example, we can get a closed form solution.

$$m(x^n|\eta) = \frac{1}{(\sqrt{2\pi})^{n+1}\eta^n}\int \exp\Big\{-\frac{1}{2\eta^2}\sum_{i=1}^{n}(x_i-\theta)^2 - \frac{\theta^2}{2}\Big\}d\theta$$

$$= \frac{1}{(\sqrt{2\pi})^{n}\eta^{n-1}\sqrt{\eta^2+n}}\exp\Big\{-\frac{1}{2\eta^2}(\sum_{i=1}^{n}x_i^2 - \frac{n^2}{\eta^2+n}\overline{x}^2)\Big\}$$

$$\times\frac{\sqrt{\eta^2+n}}{\sqrt{2\pi}\eta}\int \exp\Big\{-\frac{\eta^2+n}{2\eta^2}(\theta - \frac{n}{\eta^2+n}\overline{x})^2\Big\}d\theta$$

$$= \frac{1}{(\sqrt{2\pi})^{n}\eta^{n-1}\sqrt{\eta^2+n}}\exp\Big\{-\frac{1}{2\eta^2(\eta^2+n)}[(\eta^2+n)\sum_{i=1}^{n}x_i^2 - (\sum_{i=1}^{n}x_i)^2]\Big\},$$

so,

$$\log\frac{\prod_{i=1}^{n}p(x_i|\theta,\eta)}{m(x^n|\eta)} = \log\frac{\sqrt{\eta^2+n}}{\eta} + \frac{1}{2\eta^2(\eta^2+n)}$$

$$\times[(\eta^2+n)\sum_{i=1}^{n}x_i^2 - (\sum_{i=1}^{n}x_i)^2] - \frac{1}{2\eta^2}\sum_{i=1}^{n}(x_i-\theta)^2,$$

thus,

$$I(\eta) = \log\frac{\sqrt{\eta^2+n}}{\eta} + \frac{1}{2\eta^2(\eta^2+n)}\int\int\Big[(\eta^2+n)\sum_{i=1}^{n}x_i^2 - (\sum_{i=1}^{n}x_i)^2\Big]$$

$$\times\prod_{i=1}^{n}p(x_i|\theta,\eta)w(\theta)dx^n d\theta - \frac{1}{2\eta^2}\int\int\sum_{i=1}^{n}(x_i-\theta)^2\prod_{i=1}^{n}p(x_i|\theta,\eta)w(\theta)dx^n d\theta$$

$$= \frac{1}{2}\log(1+\frac{n}{\eta^2}) + \frac{n(n-1)}{2\eta^2(\eta^2+n)}.$$

It is decreasing in η^2, so the minimum is achieved at $\eta^* = +\infty$, and $I(\eta^*) = 0$.

If we add the constraint (1.2.1.1) with $L(x^n,\theta) = \sum_{i=1}^{n}(x_i-\theta)^2$, we have

$$l \geq \int\int\prod_{i=1}^{n}p(x_i|\theta,\eta)\sum_{i=1}^{n}(x_i-\theta)^2 w(\theta)dx^n d\theta = n\eta^2,$$

add the corresponding $\eta^{*2} = l/n$, with $I(\eta^*) = \frac{1}{2}\log\frac{n(n+l)}{l} + \frac{n^3(n-1)}{2l^2(n^3+l^2)}$. The corresponding MID degenerates to a uniform distribution on $(-\infty,\infty)$ and the corresponding posterior $w(\theta|x^n,\eta^{*2})$ is $N(\frac{n\overline{x}}{\eta^{*2}+n},\frac{\eta^{*2}}{\eta^{*2}+n}) = N(\frac{n^2\overline{x}}{n^2+l},\frac{l}{n^2+l})$. Note here η^{*2} corresponds to the $1/\lambda$ for

the λ in the MIL. As n tends to infinity, η^{*2} tends to zero. (This corresponds to λ tends to infinity for the MIL.) The corresponding $p(\cdot|\theta, \eta^{*2})$ converges to the degenerate distribution at θ, and the corresponding posterior $w(\theta|x^n, \eta^{*2})$ converges to the degenerate distribution at $\overline{x}$; this result is a parallel to (iii) of Theorem 3.3.1.

The choice of the loss function is problem dependent. In some cases, the average squared error loss may be more reasonable. If we take the loss to be $\frac{1}{n}\sum_{i=1}^{n}(x_i-\theta)^2$ in the constraint (1.2.1), then $\eta^{*2} = l$, $I(\eta^*) = \frac{1}{2}\log(1+\frac{n}{l}) + \frac{n(n-1)}{2l(l+n)}$ and the corresponding MID is $N(\theta, l)$. This has the least concentration around θ and hence the is least informative for θ. The corresponding posterior $w(\theta|x^n, \eta^{*2})$ is $N(\frac{n\overline{x}}{l+n}, \frac{l}{l+n})$. This is the posterior which has the least mean Kullback-Leibler divergence, as the likelihoods varying in the class $\mathcal{P}$, from the prior $N(0,1)$. In fact, the general forms of the posterior $w(\theta|x^n, \eta^2)$ updated by other likelihood in the class has the form $N(\frac{n\overline{x}}{\eta^2+n}, \frac{\eta^2}{\eta^2+n})$, with $\eta^2 \leq l$. They have bigger Kullback-Leibler divergence from N(0,1) than that of $w(\theta|x^n, \eta^{*2})$, since the former is more concentrated around, roughly, $\overline{x}$. Also, we see that as l increases to infinity, the constraint varnishes. In this case, the MID tends to the uniform distribution on $(-\infty, \infty)$ as in the case of no constraint, and the corresponding $I(\eta^*)$ tends to zero. The corresponding posterior tends to $N(0,1)$, which is the same as the prior, i.e. the MID did not in fact update the prior in forming the posterior.

Example 1.4.6. Again, let $w(\cdot) \sim N(0,1)$ and suppose $p(x|\theta, \eta)$ is the logistic density

$$p(x|\theta,\eta) = \frac{\exp\{-(x-\theta)/\eta\}}{(1+\exp\{-(x-\theta)/\eta\})^2}, \quad -\infty < x, \theta < \infty, \quad 0 < \eta < \infty.$$

In this example, it's hard to get a closed form solution for η^*, so we use grid search. That is for each fixed η and n, we use the Monte Carlo simulation to calculate $I(\eta)$, then find the η^* corresponding to the minimal $I(\eta^*)$. Specifically, note the SMI can be written as

$$I(\eta) = E_{\Theta,X^n}[-2\log(1+\exp\{-\sum_{i=1}^{n}(X_i-\theta)/\eta\})]$$

$$-E_{\Theta,X^n}[\log(m(X^n)] + \frac{1}{2}\log(2\pi). \tag{1.4.13}$$

We use 10^6 iterations for the Monte Carlo simulation. In each iteration, we generate θ from $N(0,1)$, then generate $x_1, ..., x_n$ *iid* from the logistic density $p(x|\theta, \eta)$ corresponding to this θ. We use the inverse distribution function method: generate a random samples $u_1, ..., u_n$

from a uniform(0,1) distribution, then get the logistic samples by $x_i = F^{-1}(u_i|\theta,\eta) = -\eta\log(1/u-1)+\theta$, where $F^{-1}(u_i|\theta,\eta)$ is the inverse cdf for the logistic density $p(x|\theta,\eta)$.

We calculated $I(\eta)$ for $\eta = 1,2,...,10$ and found a roughly decreasing pattern for the corresponding values of $I(\eta)$.

Example 1.4.7. In this example, we want to investigate the dependence in the MID. Consider $w(\cdot) \sim N(0,1)$ and $p(x_1,x_2|\theta,\eta) \sim N\Big(\begin{pmatrix}\theta\\ \theta\end{pmatrix}, \begin{pmatrix}1 & \eta\\ \eta & 1\end{pmatrix}\Big)$. Here the additional parameter η is the correlation coefficient between the two variables in the distribution. For a sample of size n, $(\mathbf{x}_1,\mathbf{x}_2)$ where $\mathbf{x}_1 = (x_{1,1},...,x_{1,n})$ and $\mathbf{x}_2 = (x_{2,1},...,x_{2,n})$, the joint density is

$$p(\mathbf{x}_1,\mathbf{x}_2|\theta,\eta) = \frac{1}{(2\pi)^n}\frac{1}{(\sqrt{1-\eta^2})^n}$$

$$\exp\Big\{-\frac{1}{2(1-\eta^2)}\Big(\sum_{i=1}^n(x_{1i}-\theta)^2 - 2\eta\sum_{i=1}^n(x_{1i}-\theta)(x_{2i}-\theta)+\sum_{i=1}^n(x_{2i}-\theta)^2\Big)\Big\}.$$

The marginal density is

$$m(\mathbf{x}_1,\mathbf{x}_2|\eta) = \frac{1}{(2\pi)^n}\frac{1}{(\sqrt{1-\eta^2})^n}\frac{1}{\sqrt{2\pi}}\int\exp\Big\{-\frac{1}{2(1-\eta^2)}$$

$$\Big(\sum_{i=1}^n(x_{1i}-\theta)^2 - 2\eta\sum_{i=1}^n(x_{1i}-\theta)(x_{2i}-\theta)+\sum_{i=1}^n(x_{2i}-\theta)^2\Big) - \frac{\theta^2}{2}\Big\}d\theta$$

$$= \frac{1}{(2\pi)^n}\frac{1}{(\sqrt{1-\eta^2})^n}\frac{1}{\sqrt{2\pi}}\int\exp\Big\{-\frac{1}{2(1-\eta^2)}$$

$$\Big(\sum_{i=1}^n(x_{1i}-\theta)^2 - 2\eta\sum_{i=1}^n(x_{1i}-\theta)(x_{2i}-\theta)+\sum_{i=1}^n(x_{2i}-\theta)^2 - (1-\eta^2)\theta^2\Big)\Big\}d\theta$$

$$= \frac{1}{(2\pi)^n}\frac{1}{(\sqrt{1-\eta^2})^n}\frac{1}{\sqrt{2\pi}}\int\exp\Big\{-\frac{2n+1+\eta}{2(1+\eta)}\Big(\theta-\frac{n(\overline{x}_1+\overline{x}_2)}{2n+1+\eta}\Big)^2\Big\}d\theta$$

$$\exp\Big\{-\frac{1}{2(1-\eta^2)}\Big(\sum_{i=1}^n x_{1i}^2 - 2\eta\sum_{i=1}^n x_{1i}x_{2i}+\sum_{i=1}^n x_{2i}^2 - \frac{n^2(1-\eta)(\overline{x}_1+\overline{x}_2)^2}{2n+1+\eta}\Big)\Big\}$$

$$= \frac{1}{(2\pi)^n}\frac{1}{(\sqrt{1-\eta^2})^n}\frac{1}{\sqrt{2\pi}}\exp\Big\{-\frac{1}{2(1-\eta^2)}$$

$$\Big(\sum_{i=1}^n x_{1i}^2 - 2\eta\sum_{i=1}^n x_{1i}x_{2i}+\sum_{i=1}^n x_{2i}^2 - \frac{n^2(1-\eta)(\overline{x}_1+\overline{x}_2)^2}{2n+1+\eta}\Big)\Big\},$$

so

$$\log\frac{p(\mathbf{x}_1,\mathbf{x}_2|\theta,\eta)}{m(\mathbf{x}_1,\mathbf{x}_2|\eta)} = \frac{1}{2}\log\frac{2n+1+\eta}{1+\eta} + \frac{1}{2(1-\eta^2)}\Big(\sum_{i=1}^n x_{1i}^2 - 2\eta\sum_{i=1}^n x_{1i}x_{2i}+\sum_{i=1}^n x_{2i}^2$$

$$-\frac{n^2(1-\eta)(\overline{x}_1+\overline{x}_2)^2}{2n+1+\eta}-\sum_{i=1}^{n}(x_{1i}-\theta)^2+2\eta\sum_{i=1}^{n}(x_{1i}-\theta)(x_{2i}-\theta)-\sum_{i=1}^{n}(x_{2i}-\theta)^2\Big),$$

and

$$I(\eta)=\int\int\int p(\mathbf{x}_1,\mathbf{x}_2|\theta,\eta)w(\theta)\log\frac{p(\mathbf{x}_1,\mathbf{x}_2|\theta,\eta)}{m(\mathbf{x}_1,\mathbf{x}_2|\eta)}d\mathbf{x}_1d\mathbf{x}_2d\theta$$

$$=\frac{1}{2}\log\frac{2n+1+\eta}{1+\eta}+\frac{1}{2(1-\eta^2)}\int\int\Big(\sum_{i=1}^{n}x_{1i}^2-2\eta\sum_{i=1}^{n}x_{1i}x_{2i}+\sum_{i=1}^{n}x_{2i}^2$$

$$-\frac{n^2(1-\eta)(\overline{x}_1+\overline{x}_2)^2}{2n+1+\eta}\Big)p(\mathbf{x}_1,\mathbf{x}_2|\theta,\eta)w(\theta)d\mathbf{x}_1d\mathbf{x}_2d\theta$$

$$-\frac{n}{2(1-\eta^2)}\int\int\Big((x_1-\theta)^2-2\eta(x_1-\theta)(x_2-\theta)+(x_2-\theta)^2\Big)p(x_1,x_2|\theta,\eta)dx_1dx_2$$

$$=\frac{1}{2}\log\frac{2n+1+\eta}{1+\eta}+\frac{n}{2(1-\eta^2)}\int\Big(1+\theta^2-2\eta(\eta+\theta^2)+1+\theta^2\Big)w(\theta)d\theta$$

$$-\frac{1-\eta}{2(1-\eta^2)(2n+1+\eta)}\int\int\Big(\sum_{i=1}^{n}x_{1i}+\sum_{i=1}^{n}x_{2i}\Big)^2p(\mathbf{x}_1,\mathbf{x}_2|\theta,\eta)w(\theta)d\mathbf{x}_1d\mathbf{x}_2d\theta$$

$$=\frac{1}{2}\log\frac{2n+1+\eta}{1+\eta}+n\frac{2-\eta(\eta+1)}{1-\eta^2}-n$$

$$-\frac{1}{2(1+\eta)(2n+1+\eta)}\int\int\Big(\sum_{i=1}^{n}x_{1i}^2+\sum_{i\neq j}x_{1i}x_{1j}+2\sum_{i=1}^{n}x_{1i}x_{2i}$$

$$+2\sum_{i\neq j}x_{1i}x_{2j}+\sum_{i=1}^{n}x_{2i}^2+\sum_{i\neq j}x_{2i}x_{2j}\Big)p(\mathbf{x}_1,\mathbf{x}_2|\theta,\eta)w(\theta)d\mathbf{x}_1d\mathbf{x}_2d\theta$$

$$=\frac{1}{2}\log(1+\frac{2n}{1+\eta})+n\frac{2-\eta(\eta+1)}{1-\eta^2}-n$$

$$-\frac{1}{2(1+\eta)(2n+1+\eta)}\Big(2n+n(n-1)+2n(1+\eta)+2n(n-1)+2n+n(n-1)\Big)$$

$$=\frac{1}{2}\log(1+\frac{2n}{1+\eta}),$$

which is minimized by $\eta^*=1$, with $I(\eta^*)=\frac{1}{2}\log n$. We see that the "optimal" correlation coefficient corresponding to the MID is just the highest dependence. In the calculations above, we have used the facts that, for $k=1,2$,

$$\int\int x_{k,i}^2p(\mathbf{x}_1,\mathbf{x}_2|\theta,\eta)w(\theta)d\mathbf{x}_1d\mathbf{x}_2d\theta=E_\Theta(E(X_{ki}^2))$$

$$=E_\Theta(Var(X_{ki})+E^2(X_{ki}))=E_\Theta(1+\theta^2)=2,$$

$$\int\int x_{k,i}x_{k,j}p(\mathbf{x}_1,\mathbf{x}_2|\theta,\eta)w(\theta)d\mathbf{x}_1d\mathbf{x}_2d\theta=E_\Theta(E(X_{k,i})E(X_{k,j}))$$

$$=E_\Theta(\theta^2)=1,$$

and

$$\int\int x_{1,i}x_{2,i}p(\mathbf{x}_1,\mathbf{x}_2|\theta,\eta)w(\theta)d\mathbf{x}_1d\mathbf{x}_2d\theta = E_\Theta(Cov(X_1,X_2))$$

$$= E_\Theta(\eta+\theta^2) = \eta+1.$$

The corresponding MID degenerates to a uni-dimensional $N(\theta,1)$. This likelihood updates the prior the least, since it basically produces one data point, and less data updates the prior less. Recall that large sample size will dominate the posterior and overwhelm the prior distribution. The corresponding posterior is

$$w(\cdot|\mathbf{x}_1,\mathbf{x}_2,\eta^*) \sim N\left(\frac{n(\overline{x}_1+\overline{x}_2)}{2n+1+\eta^*}, \frac{1+\eta^*}{2n+1+\eta^*}\right) = N\left(\frac{n(\overline{x}_1+\overline{x}_2)}{2(n+1)}, \frac{1}{n+1}\right).$$

If we add the constraint (1.2.1) with $L(x^n,\theta) = \frac{1}{n}\sum_{i=1}^n(x_{1i}+x_{2i}-2\theta)^2$, we have (take $0<l<4$)

$$l \geq \int\int\int \prod_{i=1}^n p(x_{1i},x_{2i}|\theta,\eta)\frac{1}{n}\sum_{i=1}^n[(x_{1i}-\theta)^2+2(x_{1i}-\theta)(x_{2i}-\theta)+(x_{2i}-\theta)^2]$$

$$w(\theta)d\mathbf{x}_1d\mathbf{x}_2d\theta = 2(1+\eta).$$

Now, $\eta^* = \frac{l}{2}-1$, $I(\eta^*) = \frac{1}{2}\log(1+\frac{4n}{l})$ and the corresponding MID is still a bivariate normal with mean θ, variance 1 and covariance $\frac{l}{2}-1$. This is the distribution in the class $\mathcal{P}$ which has the highest dependence between its two variables. In this way the effect of two data values will reduce to some extent to that of a single data value, and thus for the same reason as in the non-constraint case, updates the prior the least. The corresponding posterior is

$$w(\cdot|\mathbf{x}_1,\mathbf{x}_2,\eta^*) \sim N\left(\frac{2n(\overline{x}_1+\overline{x}_2)}{4n+l}, \frac{l}{4n+l}\right).$$

Since the other posterior in the class is $N\left(\frac{n(\overline{x}_1+\overline{x}_2)}{2n+1+\eta}, \frac{1+\eta}{2n+1+\eta}\right)$, with $\eta<\eta^*$, it has bigger Kullback-Leibler distance from N(0,1) than the MID does.

Example 1.4.8. Let $\mathcal{P} = \{t_\nu(\theta,\sigma) : \theta \in R^1,\ \eta \in R^+,\ \nu > 2\ \}$, where $t_\nu(\theta,\eta)$ is the t distribution with ν degree of freedom, location parameter θ and scale parameter η, that is $(X-\theta)/\eta \sim t_\nu$. In this example, the parameter to be optimized is the degree of freedom of the t-distribution and the dispersion, so we are seeking the t-distribution which, under the bounded Bayes risk constraint, updates the $N(0,1)$ prior the least. Since t_∞ is normal,

intuitively we expect the the MID is a normal distribution. Assume $p(x^n|\theta) = \prod_{i=1}^n p(x_i|\theta)$; the constraint (1.2.1.1) is

$$\int\int p(x^n|\theta)L_n(x^n,\theta)w(\theta)dx^n d\theta = n\eta^2\nu/(\nu-2).$$

The Fisher information for any member in $\mathcal{P}$ is the same

$$I(\theta|\eta,\nu) = E\Big(\frac{\partial \log p(x|\theta,\eta,\nu)}{\partial\theta}\Big)^2$$

$$= \frac{(\nu+1)^2}{\eta^2\nu}\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}\Big[\frac{\sqrt{(\nu+2)\pi}\Gamma(\frac{\nu+2}{2})}{\Gamma(\frac{\nu+3}{2})} - \frac{\sqrt{(\nu+1)\pi}\Gamma(\frac{\nu+4}{2})}{\Gamma(\frac{\nu+5}{2})}\Big]$$

$$\doteq \frac{\nu+1}{\eta^2}\frac{1}{\sqrt{\nu}}\Big[\sqrt{\nu+2} - \frac{\sqrt{\nu+4}(\nu+2)}{\nu+3}\Big],$$

which is independent of θ, so by the same reasoning as in the previous examples, minimizing the SMI subject to constraint (1.2.1.1) is equivalent to minimizing $I(\theta|\eta,\nu)$ subject to $n\eta^2\nu/(\nu-2) \le l$. Since $I(\theta|\eta,\nu)$ is decreasing in η^2, this leads to $\eta^2 = l(\nu-2)/(n\nu)$. Plugging this value in to $I(\theta|\eta,\nu)$, we are now to minimize

$$g(\nu) := \frac{\sqrt{\nu}(\nu+1)}{\nu-2}\Big(\sqrt{\nu+2} - \sqrt{\nu+4}\frac{\nu+2}{\nu+3}\Big),$$

which is positive for all finite ν, and is equivalent to, for large ν,

$$\frac{\sqrt{\nu(\nu+2)}(\nu+1)}{2(\nu-2)(\nu+1)^2} \to 0,$$

as $\nu \to \infty$. So, $g(\nu)$ is minimized as $\nu \to \infty$, or the minimizer for the SMI is $(\eta^*,\nu^*) = (\sqrt{l/n},\infty)$, this is in conformity with our intuition.

In the last two examples, we demonstrate how to use formula (1.2.2.1) efficiently to calculate the MID approximately for large sample size n.

Example 1.4.9. Let $Y = (X-\theta)/\eta \sim$ have a logistic distribution with density function $f(y) = e^{-y}/(1+e^{-y})^2$, $-\infty < y < \infty$. Assume $p(x^n|\theta) = \prod_{i=1}^n p(x_i|\theta)$, then constraint (1.2.1.1) is

$$\int\int p(x^n|\theta,\eta)L_n(x^n,\theta)(\theta)dx^n d\theta = Cn\eta^2 \le l,$$

for some constant C which is independent of n, η and θ. The Fisher information is

$$\gamma(\theta|\eta) = E\Big(-\frac{\partial^2 p(x|\theta,\eta)}{\partial\theta^2}\Big)$$

$$= \frac{2}{\eta^3} \int \frac{(\exp\{-\frac{x-\theta}{\eta}\})^2}{(1+\exp\{-\frac{x-\theta}{\eta}\})^4} dx$$
$$= \frac{2}{\eta^2} \int_0^\infty \frac{y}{(1+y)^4} dy = \frac{C}{\eta^2},$$

where C is a generic constant. Again for large n, by (1.2.2.1), minimizing the SMI over $\eta > 0$ subject to (1.2.1.1) is equivalent to minimizing $I(\theta|\eta)$ over $\eta > 0$ subject to $Cn\eta^2 \leq l$. This leads to the unique solution $\eta^{*2} = l/(nC)$ asymptotically.

Example 1.4.10. Let $\mathcal{P} = \{N(\theta, \eta^2) : \theta \in R^1, \sigma \in R^+\}$. Constraint (1.2.1.1) is

$$\int \int p(x^n|\theta) L_n(x^n, \theta) w(\theta) dx^n d\theta = n\eta^2,$$

so the corresponding Bayes risk bound l should be no smaller than $n\eta^2$. Minimizing the SMI over ρ for large n, using (1.2.2.1), is equivalent to minimizing $\gamma_{\eta^2}(\theta) = 1/\eta^2$, subject to $n\eta^2 \leq l$. The minimizer is asymptotically $\eta^2 = l/n$.

1.5 Dependence in the MIL

We see from the previous chapters that the n-dimensional MIL's are usually dependence models. It is natural to investigate the amount of dependence among the variables in the MIL. It is well known that for *iid* large data sets x^n, the posterior will be dominated by the data. Since the MIL is the likelihood which updates the prior the least, it is natural that the n-dimensional MIL will have high dependence among its variables to make the large data sets behave like a small data set. This is also suggested by Theorem 3.1.1 below. If this high dependence seems undesirable, we may model the multi-dimensional data by a product of uni-dimensional MILs. This may be appropriate in the data compression context (see Section 2.1.3).

To assess the dependence, we use a transformation of the Kullback-Leibler distance into the [0,1] scale proposed by Joe (1989). We calculate

$$\delta^* = [1 - \exp(-2\delta)]^2,$$

where

$$\delta = \int \ldots \int f(x_1, \ldots, x_n) \log \frac{f(x_1, \ldots, x_n)}{f_1(x_1) \ldots f_n(x_n)} dx_1 \ldots dx_n$$

is the relative entropy between a joint density $f(x_1, ..., x_n)$ and the product of its marginals. Since the joint MIL's we use are indexed by a parameter θ, we actually have a function $\delta_{MIL}(\theta)$. Integrating out θ to obtain an averaged measure δ_{MIL} of dependence amongst the variables in the joint MIL distribution gives

$$\delta_{MIL} = \int \delta_{MIL}(\theta) w(\theta) d\theta.$$

Indeed, sampling procedures often permit the assumption of independence, perhaps for a set of summary statistics. When this is possible, it simplifies computation. To get an independence MIL, we should add the constraint that the density for the data given the parameter factors, i.e.

$$p(x_1, ..., x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta). \tag{1.5.1}$$

Currently, we don't know how to perform the minimization of $I(X^n; \Theta)$ to get the desired independent $p^*(x^n|\theta)$ under constraint (1.5.1). Instead, as a simplification to understand the problem, what we may do is to select the loss functions and priors so that the Blahut-Arimoto algorithm will give MIL in the from of univariate products.

In Chapter 4, we will form two-dimensional independent MIL's by a product of two unidimensional MILs for the data analysis. This may be somewhat artificial, since it is not the result of the optimization procedure for the two-dimensional likelihood problem. Nevertheless, it provides a model which is not implausible and can be compared to other models we identify.

A compromise between the two methods above is to choose the independent likelihood which is closest, in the Kullback-Leibler distance, to the MIL $p^*(x^n|\theta)$ which is assumed not an independent model. That is, let $\mathcal{P}_0$ be the class of likelihoods which are independent among all their variables. We choose $\tilde{p}$ as our "independent minimally informative likelihood", i.e.

$$\tilde{p} = \arg\min_{p \in \mathcal{P}_0} D(p^*(\cdot, ..., \cdot|\theta)||p(\cdot|\theta)...p(\cdot|\theta)),$$

where

$$D(p^*(\cdot, ..., \cdot|\theta)||p(\cdot|\theta)...p(\cdot|\theta)) = \int p^*(x^n|\theta) \log \frac{p^*(x^n|\theta)}{\prod_{i=1}^{n} p(x_i|\theta)} dx^n.$$

We have

Proposition 1.5.1

$$\tilde{p}(x^n|\theta) = \prod_{i=1}^{n} p_1^*(x_i|\theta).$$

Proof: We see that

$$D(p^*(\cdot,...,\cdot|\theta)||p(\cdot|\theta)...p(\cdot|\theta)) =$$

$$\int p^*(x^n|\theta) \log \frac{p^*(x^n|\theta)}{\prod_{i=1}^n p^*(x_i|\theta)} dx^n + \int p^*(x^n|\theta) \log \frac{\prod_{i=1}^n p^*(x_i|\theta)}{\prod_{i=1}^n p(x_i|\theta)} dx^n,$$

and that the first term above does not involve $p(\cdot|\theta)$. The second term is minimized by setting $p(\cdot|\theta) = \tilde{p}(\cdot|\theta) = p^*(\cdot|\theta)$. So the "independent minimally informative likelihood" we seek is

$$\tilde{p}(x^n|\theta) = \prod_{i=1}^{n} p^*(x_i|\theta). \qquad \square$$

1.6 Computational Aspects

In cases no closed-form available for the MILs, we use the Blahut-Arimoto iterative procedure, as in Section 1.3, to evaluate the MILs numerically. Our current C-program for the MIL is effective for one-dimensional data and one-dimensional parameter case as a demonstration. The structure of the Blahut-Arimoto iterative procedure, as in (1.3.3) and (1.3.4), makes it difficult to use the current C-routines for integration. Instead, we used summation of 100 to 500 grid points to approximate integration. The convergence of the procedure depends on the choices of priors and the loss functions. Roughly, in the one-dimensional case, it needs 10 to 10^2 iterations to reach an uniformly absolute accuracy of the order 10^{-4} to 10^{-6}. The computational limit for MILs with multi-dimensional data/multi-dimensional parameter(s) is routine and machine dependent.

In the MIL we need to choose the λ so that equality in the constraint (1.2.1.1) is satisfied for the given Bayes risk bound l. For this, we can use (iv), (v) and (vi) of Theorem 3.3.2 as a guide to search for the corresponding value of λ for given l, which states roughly that l is a decreasing function of λ. This suggests the bisection rule in the choice of λ corresponds to a given l. Indeed, (Blahut 1972a) shows that the minimum in the rate distortion function is achieved for this λ.

It is expected that as the dimension increases, the amount of computation will increase exponentially, as the amount of computation involved in the integration does. The number of iterations may increase linearly, as the number of comparisons for accuracy does.

We also wrote the C-program for the corresponding posteriors for Models I and II, as in Section 4.2.2. They are for 4-dimensional data and require more CPU time than the MILs, since one must produce the MIL first, then get the corresponding posterior. It can be extended to higher dimensional cases, and will cause similar increases in the amount of computations.

To assess convergence of $p_{k,\lambda}(x|\theta)$ to the limit $p^*_\lambda(x|\theta)$ we used the supremum norm. The computation terminates when $\sup_x |p_{k,\lambda}(x|\theta) - p_{k-1,\lambda}(x|\theta)| < \epsilon$ for a given value of θ and prespecified $\epsilon > 0$. We note that the sequence $p_{k,\lambda}(x|\theta)$ tends to the limit $p^*_\lambda(x|\theta)$ independently of the initial density $m_0(\cdot)$ chosen. Indeed, one can verify in closed form that if L is squared error, w is a standard normal, $\theta = 0$ and $\lambda = 1$ then $p^*_\lambda(x|\theta = 0)$ is a standard normal. Our program gave this and $p_{k,\lambda}(x|\theta)$ was observed to converge numerically to a standard normal for a wide range of choices of m_0. Thus, the program matched what we knew had to be the case from manual calculation.

A second test of the program was to replicate numerically the results of Theorem 3.3.1, when λ tends to infinity. Specifically, we verified expressions (i) and (ii) of Theorem 3.3.1 computationally. They show that as λ increases the MIL converges to unit mass at a parameter value and the posterior from the MIL converges to unit mass at a data point. Figure 1.a shows this for the MIL: it is seen that as λ increases, p^* concentrates at the parameter value. Figure 1.b shows that as λ increases, w_{p^*} concentrates at the data value. L is squared error loss.

A third test of the program was to replicate numerically the results of Theorem 3.3.2, when λ tends to zero. Consider part (vii) of Theorem 3.3.2, write $x_0 = arg\inf_x \int L(x,\theta)w(\theta)d\theta$ assuming it is well defined, i.e $x_0 < \infty$. Then, Theorem 3.3.2 gives conditions under which $p^*_\lambda(x|\theta)$ will concentrate at x_0 independently of θ as λ tends to zero. For squared error loss and priors with finite variances, x_0 is just the prior mean. So, if w is $N(0,1)$ we find $x_0 = 0$ and that $p^*_\lambda(x|\theta)$ concentrates at zero. For w proportional to $\exp(-(x+10)), (x \geq -10)$ we found $x_0 = -9$ and $p^*_\lambda(x|\theta)$ concentrates at -9. For w proportional to $\exp(x-15), (x \leq 15)$, $x_0 = 14$ and $p^*_\lambda(x|\theta)$ concentrates at 14. In all these cases, the concentration was pronounced

by the time λ had decreased to .01 and was independent of θ, see Figure 2.

This confirms (vii) of Theorem 3.3.2. One can verify the other conclusion of Theorem 3.3.2 computationally as well, i.e., we observed the posterior formed from $p^*_\lambda(\cdot|\cdot)$ converges to the prior as λ decreases to zero; we omit showing figures for this case since the meaning is clear from Theorem 3.3.2.

Figure 1

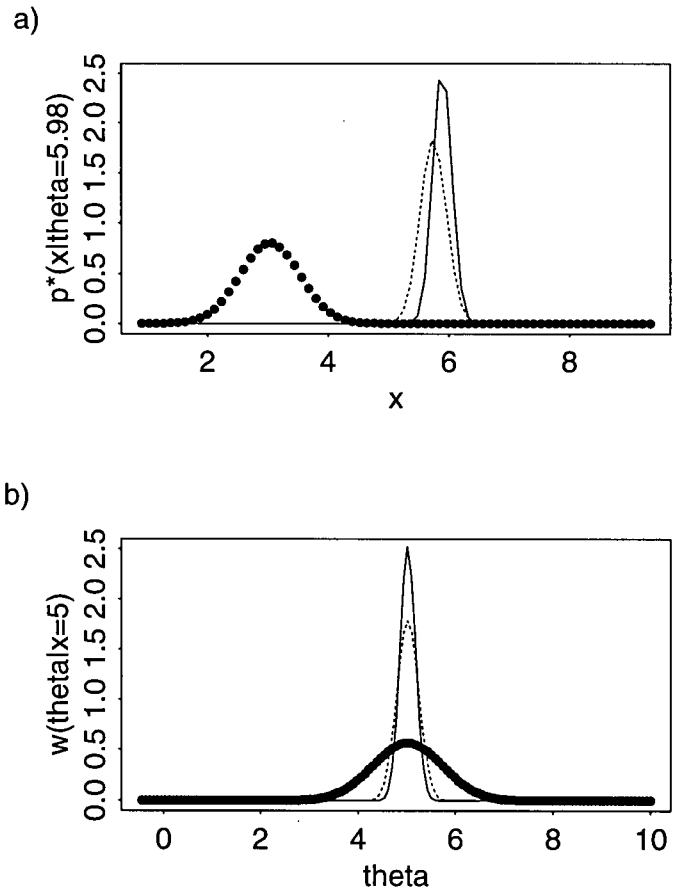

Figure 1: Effect of Increasing λ on the MIL and Posterior Density. Figure 1.a shows how $p^*_\lambda(x|\theta)$ concentrates as λ increases. Plotted are the MIL's for $\lambda = 1$ (dots), $\lambda = 10$ (dashes) and $\lambda = 20$ (solid) when w is $N(0,1)$, $\theta = 5.98$. Figure 1.b shows how the posterior based on a single observation changes as λ increases. Plotted are the posterior's for $\lambda = 1$ (dots), $\lambda = 10$ (dashes) and $\lambda = 20$ (solid) when w is $N(0,1)$ and $x = 5$.

Figure 2

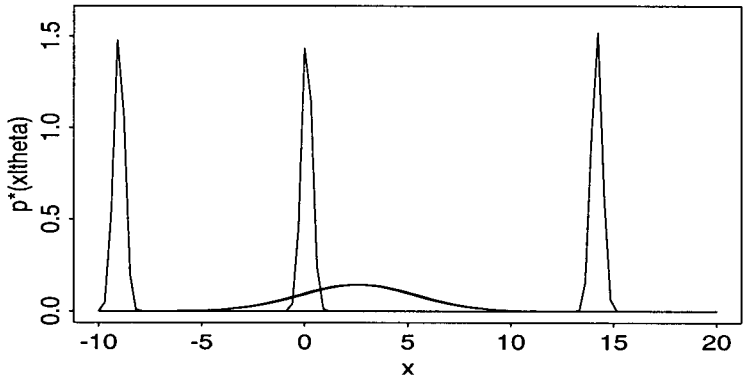

Figure 2: Effect of Decreasing λ on the MIL. Graphs of $p_\lambda(x|\theta)$ for $\lambda = .01$. The three strongly peaked curves correspond to different priors, $N(0,1)$ with $x_0 = 0$, a prior proportional to $\exp(-(x+10))$ with $x_0 = -9$ and a prior proportional to $\exp(-(x-15))$ with $x_0 = 14$. The values of θ used were $\theta = 1, 5, 10$ respectively, but the convergence to x_0 is independent of θ. The more dispersed density is $p_\lambda(x|\theta)$ for $\lambda = .001$, w given by $U(-10, 15)$, and $\theta = 5, 10$. In this case x_0 does not exist and p^* does not concentrate. L is squared error loss.

Chapter 2

Information Theory and Other Background

The likelihood is the link between what we observed and what we seek to know. Consequently, the information tacitly assumed by choosing a likelihood largely determines the results of the analysis.

Nevertheless, in practice, sometimes researcher chooses a likelihood for convenience. Sometimes a diagnostic check is used to assess the adequacy of a model. Alternatively, the statistician may choose the likelihood according to one of a large number of model selection principles or use nonparametric techniques. However once the model is obtained, it is a means to doing statistical inference. It represents the statistician's understanding of the linkage between the data observed and the values of the parameter that might specify the data generating mechanism. Here we focus on parametric families but recognize that nonparametrics and model selection provide at least in large sample cases alternatives to the technique we propose here.

Since information theoretic considerations underlie most of the key results we have to present, we now turn to the relevant background in information theory. It will be seen that the above summary of the statistical problem translates into the information theoretical setting.

2.1 Information Theory

2.1.1 Entropy, Relative Entropy and Source Coding

The concept of entropy was developed by Shannon in 1948. In his attempt to quantify the uncertainty of a random variable Θ satisfying a set of reasonable axioms, he showed

that the unique functional of the probability density $w(\cdot)$ for a discrete random variable satisfies these axioms is

$$H(\Theta) = -\sum_i w(\theta_i) \log w(\theta_i).$$

Because this quantity is similar to the entropy in thermodynamics, the name "entropy" was adopted.

Here the base of the log is e, and the unit of the entropy is measured in "nats". If another base for the logarithm is chosen, for instance, $b \neq e$, we write the entropy as $H_b(\Theta)$. If one chooses $b = 2$, and the corresponding entropy is measured in "bits". For a continuous random variable Θ with density $w(\cdot)$, its entropy is defined as

$$H(\Theta) = -\int w(\theta) \log w(\theta) d\theta.$$

For finite discrete distribution $w(\cdot)$, $H(\Theta)$ is non-negative, but not so in general.

Later, Kullback and Leibler (1951) extended the definition of entropy to measure the discrepancy between two density functions p and q as the relative entropy (or the Kullback-Leibler distance) $D(p||q)$

$$D(p||q) = E_p \log \frac{p(\Theta)}{q(\Theta)} = E_q \frac{p(\Theta)}{q(\Theta)} \log \frac{p(\Theta)}{q(\Theta)}.$$

It is not a metric, but has some metric like properties, such as non-negativity, the Pythagorean relationship, and is zero if and only if $p \equiv q$. It is stronger then the L_1 distance, see Csiszar (1975).

Similarly, the conditional entropy of Θ given X is

$$H(\Theta|X) = \int m(x) H(\Theta|X = x) dx,$$

where $m(\cdot)$ is the marginal density for X and

$$H(\Theta|X = x) = -\int w(\theta|X = x) \log w(\theta|X = x) d\theta.$$

Entropy characterizes some natural phenomena. Consider a discrete random variable Θ with mass function $w(\cdot)$. Suppose messages are drawn form $w(\cdot)$ and sent to a receiver. Before they are sent, these messages are coded into a b-ary alphabet $\mathcal{B}$ codeword (usually binary i.e $b = 2$). There are many coding methods. A code is said to be non-singular if different messages correspond to different codewords; a code is called instantaneous if it is

not a prefix of any other codeword. An instantaneous code is preferred because any given codeword can be decoded without reference to any other codeword. For a value θ, let $l(\theta)$ be the code length of θ. For any instantaneous code, we want a small average code length $El(\Theta) = \sum l(\theta)w(\theta)$ to describe a given source. Instantaneous codes are not unique for a random variable, but the set of codeword lengths for instantaneous codes is limited by the following result:

Kraft inequality: For any instantaneous code over an alphabet $\mathcal{B}$, the code lengths $l_1, ..., l_m$ must satisfy the inequality

$$\sum_i B^{-l_i} \leq 1.$$

Conversely, given a set of codeword lengths that satisfy this inequality, there exists an instantaneous code with these lengths.

An instantaneous code is said to be optimal, if it satisfies the Kraft inequality, and its expected code length is the smallest among all such codes.

The following theorem states the implication of entropy for the average length of the shortest description of a random variable.

Theorem 2.1.1.1 (Cover and Thomas, 1991). Let $l^*(\theta)$ be the optimal code length assignment to the source $w(\cdot)$ and B-ary alphabet, then

$$H_B(\Theta) \leq El^*(\Theta) \leq H_B(\Theta) + 1.$$

So, roughly speaking, entropy is the average length of the shortest description of a random variable.

The Shannon code-length assignment for the random variable Θ is $\hat{l}(\theta) = \lceil \log \frac{1}{w(\theta)} \rceil$ (here $\lceil r \rceil$ denote the smallest integer greater or equal to r). This is not necessarily optimal, but the Shannon code is operationally simple and is within 1 bit of optimal. Indeed, the following result is well known.

Theorem 2.1.1.2 (Cover and Thomas, 1991).

$$0 \leq E\hat{l}(\Theta) - H_B(\Theta) \leq 1.$$

For any source $w(\cdot)$, the optimal code length assignment can be obtained by the Huffman coding, see Cover and Thomas (1991).

Suppose we use the Shannon code, with code length assignment $\hat{l}(\theta) = \lceil \log \frac{1}{v(\theta)} \rceil$ based on the mass function $v(\cdot)$, while the true mass function for Θ is $w(\cdot)$. Then we will not achieve the optimal expected length $H_B(\Theta)$. The following theorem verifies that the increase in description length due to using the wrong mass function is the relative entropy $D(w||v)$.

Theorem 2.1.1.3 (Cover and Thomas, 1991). The expected length under $w(\cdot)$ by using the code length assignment $\hat{l}(\theta) = \lceil \log \frac{1}{v(\theta)} \rceil$ satisfies

$$D(w||v) \leq E\hat{l}(\Theta) - H(\Theta) \leq D(w||v) + 1.$$

This theorem provides the main information theoretic interpretations of the relative entropy. It is the average number of extra bits of information that one would have to send in general, but that one wouldn't have to send if one knew the true density. In other words, the relative entropy is the redundancy of a coding scheme. Thus, we shall see that Bernardo's reference prior (see Section 2.2) is the source distribution which yields the worst case Bayes redundancy, and what we have called an MIL is the parametric family of densities which produces, for a given prior, the least Bayes redundancy within a class good parametric families.

2.1.2 Channel Capacity

Consider a random variable Θ with distribution $w(\cdot)$. In the present context Θ is often called a "source" because it is preserved to supply a string of data we want to encode for some purpose. For simplicity we assume Θ is a random draw from $\{1, 2, ..., M\}$. A sender wants to send a message θ (a realization of Θ). Before sending, the message is coded into a b-ary alphabet with code-length n. This is called an (M, n) code. When the codeword reaches the receiver, it is translated back into a message. Due to background noise, there may be transmission error, so the codeword that reaches the receiver may be corrupted. The conditional probability $p(x|\theta)$ that the message sent is θ and the received message is x is called a channel. It describes the distribution of messages that might be received given the sent message. One wants $p(x|\theta)$ be high for those x's near θ, *i.e.*, relative to x, θ is a

location parameter.

Usually, larger values of M require larger code-lengths n to guarantee the distinguishability of different codewords, but this costs more. The proportion $R = M/n$ is called the rate of a code. A rate R is said to be achievable, if there exists a sequence of (M, n) codes with arbitrary small probability of error, i.e., the maximum probability of error $e(n)$ for each code word tends to zero as n tends to infinity, where $e(n) = \max_{i\in\{1,2,\ldots,B\}} e_i$, and $e_i = P(X \neq i|\Theta = i)$. That is, the received message is not i when the sent message is i. A basic question in data transmission is: what is the maximum number of bits per unit time we can send, or equivalently bits per transmission through a given channel with arbitrarily small probability of error? The capacity of a channel is defined as the maximum of all achievable rates for this channel.

Shannon's channel coding theorem establishes that the channel capacity is the supremum of the Shannon mutual information over all the input marginals, i.e.,

$$\sup_{w(\cdot)} I(X;\Theta); \tag{2.1.2.1}$$

see Cover and Thomas (1991). So, roughly speaking, we can send at most $\approx 2^{nI(X;\Theta)}$ distinguishable sequences of length n across the given channel in a single transmission. The $w^*(\cdot)$ which achieves the supremum in (2.1.2.1) is the source distribution which permits the fastest data transmission over the given channel.

We have seen the definition of the Kullback-Leibler distance (or relative entropy) between two density functions $p(\cdot)$ and $q(\cdot)$. The mutual information of two random variables X and Y with a joint density function $p(x, y)$ and marginal density $p_1(x)$ and $p_2(y)$ is defined as the relative entropy between the joint distribution and the product of the marginals, i.e.,

$$\begin{aligned} I(X;Y) &= \int\int p(x,y)\log\frac{p(x,y)}{p_1(x)p_2(y)}dxdy \\ &= E_{p_1} D(p(\cdot|X)||p_2(\cdot)). \end{aligned} \tag{2.1.2.2}$$

A Jensen's inequality argument shows that $I(X;Y) \geq 0$. $I(X;Y)$ is a measure of dependence and so arises naturally in channel coding as rate because the message required should depend strongly on the message sent. Regarding $I(X;Y)$ as a measure of dependence we see that another interpretation of Bernardo's reference prior (Section 2.2) is that it is the distribution of the parameter which depends most, asymptotically, on the data distribution

in that it changes the most upon receipt of the data. It is immediate from the definition that the larger $I(X;Y)$ is, the more dependence there is between X and Y; Indeed, $I(X;Y) = 0$ if and only if X and Y are independent. Thus, minimizing $I(X;\Theta)$ will yield a trivial result with no constraint, since this will result in a "likelihood" independent of parameter as in our first consideration for the minimization.

2.1.3 Data Compression and the Rate Distortion Function

This concept is motivated by the discretization of continuous random variables into discrete ones or data compression. Since outcomes of a continuous random variable require infinitely many bits to represent, the only practical way to represent is to "compress" it by representing it with finitely many bits. Consider dividing the support of a one-dimensional continuous random variable X into 2^{nR} intervals. Let $\hat{X}(X)$ be a discrete random variable assuming values 1, 2, ..., 2^{nR}, depending on which cell X lands in, where R is the code rate described in section 1.3.2. A distortion function $L(x, \hat{x})$ is a measure of the loss representing x by $\hat{x}$. For instance $\hat{x}$ may be the midpoint of the interval x lies in and $L(x, \hat{x}) = |x - \hat{x}|$. We want the compressed data $\hat{X}(X)$ to represent the true data X with small "distortion", i.e. the expected loss or Bayes risk

$$EL(X, \hat{X}(X)) = \int \int p(x)p(\hat{x}|x)L(\hat{x}, x)dxd\hat{x}.$$

Intuitively, the larger the rate R is, the more accurate the representation, and the smaller the distortion, but the higher the cost in operation. In the present case, the more accurate our representation will be. However, we want to use as few intervals as possible. Assigning more x values into larger intervals means we are throwing out information. If we fix a level of distortion we are willing to tolerate, we are led to minimizing R since we want to compress as much as possible, i.e., throw out information by permitting less accurate representations of x, subject to the distortion constraint.

We are interested in: For a given source, what is the minimum rate to achieve a distortion no greater than a given tolerable distortion l? And what is the corresponding channel?

For a given positive number l, the rate distortion function $R(l)$ is defined as the minimum rate to achieve the distortion l. It is the minimum amount of information needed for representing the source with average loss bounded by l. The rate distortion theorem

establishes that the rate distortion function valued at l is the minimum of the SMI over conditional densities with distortion (Bayes risk) bounded by l:

$$R(l) = \min_{p(\hat{x}|x) \in \mathcal{P}_l} I(X, \hat{X}),$$

where $\mathcal{P}_l = \{p(\hat{x}|x) : \int \int p(x)p(\hat{x}|x)L(\hat{x}, x)dxd\hat{x} \leq l\}$ is the class of channels with distortion no greater than l. This is the quantity we investigated in choosing the MIL.

Note that the above concepts can be expressed in terms of channels as well. The conditional density achieving the rate $R(l)$ is the channel with the slowest transmission for the given source, with tolerable distortion l. It is what we have called the MIL. It is the conditional density providing optimal data compression, in the sense that it provides the greatest compression within the allowed distortion.

In practice, one uses $R(l)$ as a theoretical lower bound, seeking discretizations of X into regions which provide optimal compression. Usually, one wants as few regions as possible provided they do not cause excessive distortion. Current work on this problem is often called vector quantization.

2.1.4 Comparison with the ME Formulation

The principal of the maximum entropy and our method are similar, since minimizing the SMI is equivalent to maximizing the conditional entropy. However, there are some differences also.

The maximum entropy, ME, method is used for selecting an optimal likelihood based on incomplete information about the likelihood. The information available is incorporated into a set of known constraint(s), and the least informative likelihood subject to these constraints is found in an entropy sense.

The ME likelihoods are in the exponential family

$$p(x|\theta) = a(\beta) \exp[\beta_1 T_1(x) + \cdots + \beta_k T_k(x)],$$

where the parameter $\beta = (\beta_1, ..., \beta_k)'$ is chosen so that the likelihood satisfies the constraint(s). Its exponent part has a fixed form, the corresponding sufficient statistics $T_1(x), ..., T_k(x)$ are determined by the form of the constraints.

Our method is aimed at selecting an optimal likelihood in a Bayesian setting with a known prior and some incomplete information about the likelihood: It has bounded Bayesian risk. Here we assume less than the ME method and many other known methods, and we have incorporated the prior information about the parameter.

The MIL is a function of an exponential family and the marginal of the data

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

where $m(x) = \int p^*(x|\theta)w(\theta)d\theta$ is the marginal of the data, and λ is chosen so that the equality in the Bayesian risk constraint is satisfied. Its exponent structure is determined by the loss function $L(\cdot,\cdot)$ rather than by moments as is the ME likelihood. Our method produces a likelihood, i.e., a parametric family as a functional of the prior information. The parametric family can be used in frequentist techniques, or (with a different prior even) in Bayesian techniques. In general the MIL is not an exponential family, but we conjecture the set of MIL's contains the collection of exponential families. The MIL defines the channel which transmits information as slowly as possible subject to a distortion constraint that ensures data transmission actually occurs.

2.1.5 Interpretation of the MIL

From the information theory point of view, the MIL is the conditional density which achieves the rate distortion function lower bound. The rate distortion function plays a central role in data compression and has an interpretation in transmitting data across a channel. Usually in data transmission, large amount of source message is compressed into a relatively smaller number of representatives for practical purposes. For example, continuous variables must be represented by finitely many representatives for transmission for economical or operational reasons. There are many ways to do so. For a given source, the possible representatives constitute a codebook. We want a code that is optimal in that it has the fewest codewords (x^n) (fewest representatives or greatest compression), where n is the code length. This is to be accomplished subject to not losing too much information about the original source (θ). The loss is quantified by the distortion $L(x^n,\theta)$, the Bayes risk bound l in (1.2.1.1) constrains the average distortion. So, the MIL is just the conditional distribution of the optimal code given the source, subject to average distortion bound l. This causes high dependence be-

tween x^n and θ and amongst the entries of x^n. For more details about data compression, see Blahut (1987), Cover and Thomas (1991). Both provide information-theoretic justification based on data compression for calling p^*_λ minimally informative. Here, we only describe the channel-based interpretation which we argue is more appropriate to the statistical context.

An information-theoretic channel is a conditional distribution which specifies the distribution of the output received given the input sent. The input is a coded version of the message. The output received has a probabilistic description because even though we transmit a specific message it may be corrupted by background noise. For instance, a conditional density such as $p(x|\theta)$ defines a channel: If θ is the input then the channel gives output x with probability $p(x|\theta)$. If θ is drawn from a source distribution with density w then the SMI can be interpreted as a rate of transmission in bits per unit time. Therefore, minimizing the SMI over a constrained set of channels defined by conditional densities yields the channel in that set with the slowest rate of transmission which we have called minimally informative.

Here, the set we have used is the collection of densities for which the Bayes risk of estimating θ with X is bounded by a number l. Information theoretically, this means that the average discrepancy, or distortion, between the output X and the input θ is bounded. That is, all the channels we are considering must transmit at least some information related to the input θ. The minimal SMI is the slowest rate for this transmission and we have numerically found the conditional density achieving this rate. We regard it as an "optional" likelihood with the desired Bayesian loss based on the incomplete information, and propose to use it as a default likelihood in certain settings we identify in Chapter 4.

Now we look at the parameter λ in the MIL. The inverse of the parameter λ in $p^*_\lambda(x|\theta)$ behaves like a dispersion parameter. Under reasonable conditions, it is a decreasing function of l which controls the amount of risk (distortion). For the $p^*_\lambda(\cdot|\theta)$ in Example 1.4.3, 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 (iii) of Theorem 3.3.1. More generally, our computations show that $p^*_\lambda(\cdot|\theta)$ spreads out as λ shrinks and concentrates at θ as λ grows.

Alternatively, one can regard λ as a smoothing parameter ensuring that a minimally informative density does not just concentrate at the data points.

2.2 Relation to Reference Priors

Since the method we used for selecting the optimal likelihood has some connection with that used for noninformative priors, we also review some background on prior selection. In a Bayesian setting, the pre-experimental knowledge about the parameters of interests is incorporated into the prior distribution. When such experience is available, the Bayesian is expected to be more efficient in statistical inference about the parameter than the non-Bayesian, in the sense that the class of all Bayes rules is a complete class, or non-Bayes is inadmissible (see Wald, 1950). However, when such pre-experimental information is far from enough to establish a prior distribution, how to choose a prior for inference remains an important issue for a Bayesian. Much work has been done on this. For example, conjugate priors are often based on mathematical convenience. They require that the posterior and prior be in the same distribution family. Another criterion was invariance. Jeffreys' non-informative prior was originally proposed to satisfy an invariance principle. In 1979, Bernardo proposed the reference prior, which is based on an information theoretic optimality criterion: One selects the prior for which the posterior is updated the most, asymptotically in the expected Kullback-Leibler measure, i.e. it is $\arg\max_w \lim_{n\to\infty} E_m D(w(\cdot|X^n)||w(\cdot))$, so it is the prior that permits the posterior to change from it the most upon receipt of the data, on average, in an asymptotic sense.

In a fully Bayes setting, one has some pre-experimental beliefs about the parameter encapsulated in a prior density. Often, in practice, we do not have as many data points as desirable for many known methods, and the reasoning for the basic assumptions behind is unclear. In this case, choosing any known likelihood to model the data seems inappropriate, and how to model the data reasonably becomes a basic and practical problem. In the fully Bayes setting, the prior represents partial information in parallel to that specified by the constraints of the ME method. We want a likelihood which is "unbiased" in that it is reasonable based on this partial information.

The MIL method here is, in some sense, the reverse of the reference prior method of Bernardo (1979). Our task is to choose a likelihood given the prior, while Bernardo identified a prior given the likelihood. Specifically, Bernardo found a way to choose a prior in the

absence of information about the parameter. He used the Shannon mutual information (SMI), or the expected Kullback-Leibler distance between the posterior and the prior

$$I(\Theta, X^n) = E_m D(w(\cdot|X^n)||w(\cdot))$$

$$= \inf_{q \in \mathcal{Q}} \int w(\theta) p(x^n|\theta) \log \frac{p(x^n|\theta)}{q(x^n)} d\theta dx^n,$$

where $m(x^n) = \int p(x^n|\theta) w(\theta) d\theta$ is the marginal density of the data X^n, $\mathcal{Q}$ is the class of all the n dimensional densities. It measures, on average, the discrepancy between the posterior and the prior. He maximized, asymptotically, the SMI over all priors. Recognizing that m is the Bayes estimator for $p(\cdot|\theta)$, Bernardo examined

$$\max_{w \in \mathcal{W}} \inf_{q \in \mathcal{Q}} \int w(\theta) p(x^n|\theta) \log \frac{p(x^n|\theta)}{q(x^n)} d\theta dx^n. \tag{2.2.1}$$

The asymptotic maximizer $w^*(\cdot)$ is his reference prior. It differs most, on average, from the posterior in an asymptotic sense. It is the prior that contains the least information about the parameter, since the posterior based on it is furthest away from the prior.

Under some regularity conditions, Jeffreys' non-informative prior is a special case of Bernardo's reference prior, see Clarke and Barron (1994).

We see that prior selection forces the prior to be far from the posterior, but if we are selecting a likelihood, we want the posterior differ not too much from the prior.

Thus, we have minimized the SMI under a constraint to get an optimal likelihood, while Bernardo maximized the SMI (asymptotically) to get an optimal prior. Operationally, Bernardo's method is a max-min procedure, while our method is a double minimization

$$\min_{q \in B} \min_{p \in A} D(p||q),$$

where A is the set of all the likelihoods that satisfy the Bayes risk constraint, B is the set of product distributions $w(\theta)r(x^n)$ with arbitrary densities $w(\theta)$ and $r(x^n)$, see Cover and Thomas (1991). Likelihoods are less informative when the posterior is close to the prior. Priors are less informative when they give a posterior far from the prior.

Our initial efforts to find a minimally informative likelihood and reverse Bernardo's approach originally led us to consider minimizing the expected Kullback-Leibler distance between the posterior and the "contaminated" prior over likelihoods. That is, we used $(1-\alpha)w(\theta) + \alpha\phi(x,\theta)/m(x)$ in the SMI, where $0 \leq \alpha \leq 1$ is fixed and $\phi(x,\theta)$ is a given

non-negative function. In the case of a single outcome, our functional can be written as

$$E_m D\Big(w(\cdot|X)||(1-\alpha)w(\cdot)+\alpha\frac{\phi(X,\cdot)}{m(X)}\Big),$$

where $D(p||q)=\int p(x)\log\frac{p(x)}{q(x)}dx$ is the relative entropy between two densities $p(\cdot)$ and $q(\cdot)$. The standard method of calculus of variations gives a type of Fredholm equation, see for example, Kondo (1991),

$$p(x,\theta)=\alpha c(\theta)\phi(x,\theta)+(1-\alpha)c(\theta)w(\theta)\int p(x,\xi)d\xi,$$

which results in

$$p(x|\theta)=\alpha\frac{c(\theta)}{w(\theta)}\phi(x,\theta)+\frac{\alpha(1-\alpha)\sqrt{c(\theta)/w(\theta)}\int c(\xi)\phi(x,\xi)d\xi}{1-(1-\alpha)\int c(\xi)w(\xi)d\xi},$$

where $c(\theta)$ is the normalizing constant for each θ. We verified minimality for this solution, and for some choices of α and ϕ this $p(\cdot|\theta)$ is non-negative. However, it is not clear that this solution admits any physical interpretation. Moreover, it appeared mathematically intractable.

Later, instead of modifying the functional to be optimized, we tried restricting the class of likelihoods over which we conducted the optimization. When we sought meaningful quantities to optimize over large classes of likelihoods, it seemed natural to start with the SMI. Finally, Decision theory led us to considered minimizing the SMI over likelihoods in the class with a bounded Bayes risk, and gave a constraint of the form

$$\int\int w(\theta)p(x|\theta)L(x,\theta)dxd\theta\le l. \tag{2.2.2}$$

From this one can recognize that the optimization is the same as that in the definition of the rate distortion function in an information theory context. Statistically, the MIL is the likelihood which gives a posterior updated from the prior the least on average. In the next Chapter, we will see formally that the Bayes risk bound l behaves like a dispersion parameter in the MIL.

In practice, the Bayes risk bound l may be chosen subjectively according to the experimenters tolerance for the risk. However, how to choose l in a general setting is still a question for further work. We address this heuristically in the application of Chapter 4.

2.3 Other Background

There are numerous methods in literature regards likelihood selection. Here we give a partial review of those have some relevance with our methods.

Many authors have used and developed the maximum entropy (ME) method for choosing a likelihood based on incomplete information. Usually one assumes the "partial information" may be incorporated into a set of moment constraint(s)

$$E[T_k(X)] = \theta_k, \qquad k = 0, 1, ..., m$$

It can be shown, see Jaynes (1957), that the maximum entropy distribution under these constraints is of the form

$$p(x|\theta) = a(\beta) \exp[\beta_1 T_1(x) + \cdots + \beta_k T_k(x)], \tag{2.3.1}$$

where the β's are chosen so that the moment constraints are satisfied. The family (2.3.1) is the "least informative" distribution in the absence of the adequate knowledge about the data generating mechanism.

In practice, usually little is known about the data generating mechanism, so the ME method plays an important role in data modeling in these situations.

As the other data modeling strategies, the ME is not a perfect principle. The concern is how closely the ME distribution approximates the data generating distribution for a given data. It is reasonable to ask if the data generating distribution is well approximated by the information specified constraints. If this is the case, then the entropy of the data distribution is expected to be somewhat close to the maximum entropy H_{max}. "When the constraints do not reflect the information content of the underlying random mechanism of data generating process, then a non-parametric estimate of the entropy solely based on the data would generally yield an unacceptable lower value than H_{max} estimated by the data. In such a case, the use of the ME distribution would be inadequate because it will fail to predict the future outcomes correctly", see Soofi (1994).

The minimum complexity or minimum description length criterion developed by Kolmogorov (1965) is another information theoretic modeling selection criterion. Assume

$X_1, ..., X_n$ are *iid* random variables with common density $p(x)$. Denote $p(x^n) = \prod_{i=1}^n p(x_i)$. Let Γ_n be a countable collection of density functions. For each $p(\cdot) \in \Gamma_n$, there is a non-negative number $L_n(p)$ which is the description length of p. In Kolmogorov's original formulation it is the length of the shortest computer program that can calculate p. Barron and Cover (1989) modified this idea so as to interpret L_n as a code length from a codebook which provides a code for each member of Γ_n. In this case, the minimum complexity or minimum description length criterion is to choose the $\hat{p}_n \in \Gamma_n$ which minimizes the complexity of the data X^n relative to L_n and Γ_n, *i.e.* the $\hat{p}_n \in \Gamma_n$ defined by

$$\hat{p}_n = \arg\min_{p \in \Gamma_n} B(X^n) = \arg\min_{p \in \Gamma_n} \Big(L_n(p) - \log p(X^n) \Big).$$

In information theory, the terms $L_n(p)$ and $-\log p(X^n)$ are, respectively the description length of p and the Shannon code length of X^n based on p.

This minimum complexity estimator has many useful properties, see Barron and Cover (1989). They also provided a Bayesian interpretation based on using the Kraft inequality to regard $L_n(p)$ as a prior.

The information criteria Akaike Information Criteria, AIC and the Bayes Information Criteria, BIC are also well known methods for model selection (see, Akaike, 1977). Let $\mathcal{P}$ be a class of *iid* likelihoods with a k-dimensional parameter θ, let $\hat{\theta}$ be the maximum likelihood estimate of θ based on a sample of size n, and assume the density for the data is in $\mathcal{P}$. The AIC criterion is to choose the model in $\mathcal{P}$ which minimizes

$$AIC(k) = -2\log(p(x^n|\hat{\theta})) + 2k.$$

It is argued that the AIC has a maximum entropy interpretation.

An alternative to the AIC is the BIC. It chooses the optimal k for the dimensionality of the parameter. The BIC is

$$BIC(k) = AIC + k(\log(n) - 1) + \log(Q(k)/k),$$

where $Q(k)$ is the projection of the n-dimensional observed data into a k-dimensional space, its functional form depending on the method of estimation. The BIC method is to choose the k which minimizes the BIC (see, Akaike, 1977).

These usual methods are only asymptotically optimal under regularity conditions. By contrast, the method we propose has some small sample optimality and provides a flexible

class of models for consideration. We note that the AIC is rarely consistent and The BIC rests on Bayes testing for its optimality. See Schwartz (1978) and Haughton (1988).

Chapter 3

Main Results on The MILs

In this chapter we establish our main results on the MIL's.

Let $\underline{r} = \inf_x \int w(\theta)L(x,\theta)d\theta$. This value $\underline{r}$ is achieved at the point x which is closest to the center of the distribution of Θ. Our first result is that the parametric family we identified is unique.

Proposition 3.1. (i) For each $l \in [0, \underline{r})$, $R(l)$ has an unique minimizer $p^*(\cdot|\cdot)$ in $\mathcal{P}_l$.

(ii) For $l \geq \underline{r}$, $R(l) = 0$ and it is achieved by any $p(\cdot)$ which is independent of θ.

(iii) Assume the parameter θ is a permutation symmetric functional of the distribution of X^n, *i.e.* let $F_{X_1,...X_n}$ be the joint distribution of $(X_1, ..., X_n)$, there is a functional $G(\cdot)$ such that for any permutation $(i_1, ..., i_n)$ of $(1, ..., n)$

$$\theta = G(F_{X_1,...,X_n}) = G(F_{X_{i_1},...,X_{i_n}})$$

and $L_n(x^n, \theta)$ is permutation symmetric in $x_1, ..., x_n$, then $p^*(x^n|\cdot)$ is permutation symmetric in $x_1, ..., x_n$.

Proof: (i) First note that any $p(\cdot)$ which is independent of θ is excluded form $\mathcal{P}_l$. In fact, for any $p(\cdot)$,

$$\int\int p(x)w(\theta)L(x,\theta)dxd\theta \geq \int p(x) \inf_t \int w(\theta)L(t,\theta)d\theta dx$$
$$= \int p(x)\underline{r}dx = \underline{r},$$

so, $p(\cdot)$ is not in $\mathcal{P}_l$.

Next note that $\mathcal{P}_l$ is a convex set of probability densities. In fact $\forall p_1(\cdot|\cdot) \in \mathcal{P}_l$, $p_2(\cdot|\cdot) \in \mathcal{P}_l$ and $0 \leq \alpha \leq 1$,

$$\int\int \Big(\alpha p_1(x|\theta) + (1-\alpha)p_2(x|\theta)\Big)L(x,\theta)w(\theta)dxd\theta$$

$$= \alpha \int\int p_1(x|\theta)L(x,\theta)w(\theta)dxd\theta + (1-\alpha)\int\int p_2(x|\theta)L(x,\theta)w(\theta)dxd\theta$$
$$\leq \alpha l + (1-\alpha)l = l,$$

that is $\alpha p_1(\cdot|\cdot) + (1-\alpha)p_2(\cdot|\cdot) \in \mathcal{P}_l$.

Now it is enough to show that $I(\Theta, X)$ is strictly convex on $\mathcal{P}_l$ as a functional of $p(\cdot|\cdot)$. Write $I(\Theta, X)$ as $I(p, \Theta)$ to indicate its relationship with $p(\cdot|\cdot)$. Now $\forall\ 0 \leq \lambda \leq 1$, $p_1(\cdot|\cdot), p_2(\cdot|\cdot) \in \mathcal{P}_l$, with $p_1(\cdot|\cdot) \neq p_2(\cdot|\cdot)$, we have

$$I(\lambda p_1 + (1-\lambda)p_2, \Theta)$$
$$= \int\int w(\theta)[\lambda p_1(x|\theta) + (1-\lambda)p_2(x|\theta)] \log \frac{\lambda p_1(x|\theta) + (1-\lambda)p_2(x|\theta)}{\int w(\xi)[\lambda p_1(x|\xi) + (1-\lambda)p_2(x|\xi)]d\xi} d\theta dx$$
$$= \int\int w(\theta)[\lambda p_1(x|\theta) + (1-\lambda)p_2(x|\theta)] \log \frac{\lambda p_1(x|\theta) + (1-\lambda)p_2(x|\theta)}{\lambda m_1(x) + (1-\lambda)m_2(x)} d\theta dx.$$

The log-sum inequality states that for any integer n and any non-negative numbers $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$,

$$(\sum_{i=1}^{n} a_i) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \leq \sum_{i=1}^{n} a_i \log \frac{a_i}{b_i},$$

with equality if and only if a_i/b_i is a constant over all i. Now, $I(\lambda p_1 + (1-\lambda)p_2, \Theta)$ is bounded from above by

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

(ii) Let $x_0 = \arg\inf_x \int w(\theta)L(x,\theta)d\theta$, $p_0(\cdot)$ be the density which is independent of θ and is concentrated in a small neighbourhood of x_0, then apparently, $p_0(\cdot) \in \mathcal{P}_l$, and $I(p_0||\Theta) = 0$, since $p_0(\cdot)$ is independent of θ.

(iii) By the Blahut-Arimoto iterative procedure in Section 1.6.1, we can choose $m_0(\cdot, ..., \cdot)$ to be a permutation symmetric density, so in each step k of iteration, $p_k(\cdot, ... \cdot|\theta)$ is permutation symmetric. Thus $\forall i, j$

$$p^*(x_1, ..., x_i, ..., x_j, ..., x_n|\theta) = \lim_{k\to\infty} p_k(x_1, ..., x_i, ..., x_j, ..., x_n|\theta) =$$

$$\lim_{k\to\infty} p_k(x_1, ..., x_j, ..., x_i, ..., x_n|\theta) = p^*(x_1, ..., x_j, ..., x_i, ..., x_n|\theta),$$

that is, $p^*(x_1, ..., x_n|\theta)$ is permutation symmetric in its argument. □

We comment that MIL's can be used to form a posterior density or can be used to obtain frequentist estimators. The prior has thus far only been used to get a likelihood. One need not use it again to form a posterior. This is a frequentist usage (getting a point estimator) of a Bayesian quantity (a prior).

3.1 Large Sample Properties of the MIL

Consider the collection of parametric families of the same form as $\mathcal{P}_l$, but for a random variable X^n in place of the univariate X. That is, let

$$\mathcal{P}_n = \{p_n(x^n|\theta) : \int\int p_n(x^n|\theta)w(\theta)L_n(x^n,\theta)dx^n d\theta \leq l_n\}. \tag{3.1.1}$$

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

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

Similar to the univariate case handled in Blahut (1972a), one can obtain a form for the MIL based on the loss function L. For given prior w, this is

$$p_{MIL}(x^n|\theta) = \frac{m_n^*(x^n)e^{-\lambda_n L_n(x^n,\theta)}}{\int m_n^*(y^n)e^{-\lambda_n L_n(y^n,\theta)}dy^n}, \tag{3.1.2}$$

where $m_n^*(x^n)$ is determined by

$$\int \frac{e^{-\lambda_n L_n(x^n,\theta)}w(\theta)}{\int m_n^*(y^n)e^{-\lambda_n L_n(y^n,\theta)}dy^n}d\theta \leq 1 \tag{3.1.3}$$

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 (3.1.2) and the prior w is asymptotically the same as w in a relative entropy sense when convergence is assessed in the mixture distribution. That is, the data update w trivially. In addition, we will see that use of p_{MIL}, or $p_\lambda^*(x|\theta)$ to denote its dependence on λ explicitly, gives the weakest inferences possible amongst the elements of $\mathcal{P}_n$

To establish the asymptotic equivalence of w and the posterior based on w and (3.1.2) we note that P_{MIL} is typically a dependence model, in which the dependence structure depends on n and λ_n. Because P_{MIL} typically cannot be given in closed form, the proof of our first theorem requires a carefully chosen independence density $p_n(\cdot|\theta)$ in $\mathcal{P}_n$. This

$p_n(\cdot|\theta)$ is chosen so that the expected relative entropy between the posterior based on p_{MIL} and w and the prior is bounded by the relative entropy between the posterior based on $p_n(\cdot|\theta)$ and w and the prior. Then we prove the latter tends to zero as n goes to infinity.

In the definition of $\mathcal{P}_n$, if we take $L_n(x^n,\theta) = a_n \sum_{i=1}^n L(x_i,\theta)$ for given $L(\cdot,\cdot)$, then we can absorb the l_n into a_n and assume $l_n = 1$, for all n. Thus our set $\mathcal{P}_n$ is the same as used in the usual formulation of the rate distortion problem, see Cover and Thomas (1991).

To state the theorem, we define the average loss for fixed x as $r(x) = \int w(\theta)L(x,\theta)d\theta$, and denote its supremum and infimum by

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

Now we show that a posterior based on $p_{MIL}(x^n|\theta)$ and the prior w which generated it updates w trivially, in an asymptotically average sense.

Theorem 3.1.1. Assume that

$$\forall x, \quad r(x) = \int w(\theta)L(x,\theta)d\theta < \infty,$$

and that, for all n, $l_n = 1$. Let $L_n(x^n,\theta) = a_n \sum_{i=1}^n L(x_i,\theta)$ where $L(\cdot,\cdot)$ is continuous in both arguments and assume that the limit of na_n exists and is s. Now, if $\underline{r}s < 1$ we have that

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

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

$$p_n(x^n|\theta) = \frac{e^{-L_n(x^n,\theta)} \prod_{i=1}^n q(x_i)}{\int e^{-L_n(y^n,\theta)} \prod_{i=1}^n q(y_i)dy^n} \tag{3.1.4}$$

is an element of $\mathcal{P}_n$ for n large enough.

Indeed, take a constant $\underline{r} < \tilde{r} \leq \bar{r}$ and a constant $b \in (\underline{r}, \tilde{r})$ such that $bs < 1$. Choose a probability density $q(\cdot)$ such that for all θ

$$\int q(x)L(x,\theta)dx < \infty,$$

and

$$\int w(\theta)q(x)L(x,\theta)dxd\theta = b.$$

That is, we have $\int q(x)(r(x)-b)dx = 0$. Such a probability density $q(\cdot)$ exists because $\underline{r} < b < \tilde{r}$.

By the symmetry of $p_n(x^n|\theta)$, the sum of integrals from L_n can be reduced to a univariate integral. Now, the density $p_n(x^n|\theta)$ satisfies

$$\int\int p_n(x^n|\theta)w(\theta)L_n(x^n,\theta)dx^nd\theta$$

$$= na_n \int\int \frac{w(\theta)\prod_{i=1}^n q(x_i)e^{-a_n\sum_{i=1}^n L(x_i,\theta)}L(x_1,\theta)}{\int\prod_{i=1}^n q(y_i)e^{-a_n\sum_{i=1}^n L(y_i,\theta)}dy^n}dx^nd\theta$$

$$= na_n \int\int \frac{w(\theta)q(x)e^{-a_nL(x,\theta)}L(x,\theta)}{\int q(y)e^{-a_nL(y,\theta)}dy}dxd\theta. \tag{3.1.5}$$

Denote the double integral in (3.1.5) by $I(a_n)$. Since $na_n \to s$, and $sb < 1$, since in the definition of $\mathcal{P}_n$, $l_n = 1$ 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 standard inequalities we have that,

$$|I(a_n) - b|$$

$$= \left|\int\int \frac{w(\theta)q(x)L(x,\theta)e^{-a_nL(x,\theta)}}{\int q(y)e^{-a_nL(y,\theta)}dydx}d\theta - \int\int w(\theta)q(x)L(x,\theta)dxd\theta\right|$$

$$\leq \int\int w(\theta)q(x)L(x,\theta)\frac{|e^{-a_nL(x,\theta)} - \int q(y)e^{-a_nL(y,\theta)}dy|}{\int q(t)e^{-a_nL(t,\theta)}dt}dxd\theta$$

$$\leq \int w(\theta)\frac{\int\int q(x)q(y)L(x,\theta)|e^{-a_nL(x,\theta)} - e^{-a_nL(y,\theta)}|dxdy}{\int q(t)e^{-a_nL(t,\theta)}dt}d\theta. \tag{3.1.6}$$

Let $A_\theta = \{(x,y)|L(x,\theta) \geq L(y,\theta)\}$. Since

$$\frac{\int\int q(x)q(y)L(x,\theta)|e^{-a_nL(x,\theta)} - e^{-a_nL(y,\theta)}|dxdy}{\int q(t)e^{-a_nL(t,\theta)}dt}$$

$$\leq \frac{\int\int_{A_\theta} q(x)q(y)L(x,\theta)|e^{-a_nL(x,\theta)} - e^{-a_nL(y,\theta)}|dxdy}{\int q(t)e^{-a_nL(t,\theta)}dt}$$

$$+\frac{\int\int_{A_\theta^c} q(x)q(y)L(y,\theta)|e^{-a_nL(x,\theta)} - e^{-a_nL(y,\theta)}|dxdy}{\int q(t)e^{-a_nL(t,\theta)}dt}$$

$$= \frac{\int\int_{A_\theta} q(x)L(x,\theta)q(y)e^{-a_nL(y,\theta)}|e^{-a_n(L(x,\theta)-L(y,\theta))} - 1|dxdy}{\int q(t)e^{-a_nL(t,\theta)}dt}$$

$$+\frac{\int\int_{A_\theta^c} q(y)L(y,\theta)q(x)e^{-a_nL(x,\theta)}|1-e^{-a_n(L(y,\theta)-L(x,\theta))}|dxdy}{\int q(t)e^{-a_nL(t,\theta)}dt}$$

$$\leq \frac{\int\int_{A_\theta} q(x)L(x,\theta)q(y)e^{-a_nL(y,\theta)}dxdy}{\int q(t)e^{-a_nL(t,\theta)}dt} + \frac{\int\int_{A_\theta^c} q(y)L(y,\theta)q(x)e^{-a_nL(x,\theta)}dxdy}{\int q(t)e^{-a_nL(t,\theta)}dt}$$

$$\leq \frac{\int\int q(x)L(x,\theta)q(y)e^{-a_nL(y,\theta)}dxdy}{\int q(t)e^{-a_nL(t,\theta)}dt} + \frac{\int\int q(y)L(y,\theta)q(x)e^{-a_nL(x,\theta)}dxdy}{\int q(t)e^{-a_nL(t,\theta)}dt}$$

$$= 2\int q(x)L(x,\theta)dx,$$

which is integrable w.r.t. $W(\theta)$, since

$$2\int w(\theta)\int q(x)L(x,\theta)dxd\theta = 2\int w(\theta)q(x)L(x,\theta)dxd\theta$$

$$= 2b < \infty.$$

By the Dominated Convergence Theorem, the limit of the left hand side of (3.1.6) is bounded by

$$\int w(\theta)\lim_n\left(\frac{\int\int q(x)q(y)L(x,\theta)|e^{-a_nL(x,\theta)}-e^{-a_nL(y,\theta)}|dxdy}{\int q(t)e^{-a_nL(t,\theta)}dt}\right)d\theta, \tag{3.1.7}$$

and for any fixed θ, the limit in (3.1.7) is

$$\frac{\lim_n \int\int q(x)q(y)L(x,\theta)|e^{-a_nL(x,\theta)}-e^{-a_nL(y,\theta)}|dxdy}{\lim_n \int q(t)e^{-a_nL(t,\theta)}dt} \tag{3.1.8}$$

provided the numerator of (3.1.8) exists and the denominator of (3.1.8) exists and is non-zero. In the numerator, the integrand is bounded by $2q(x)q(y)L(x,\theta)$, which is integrable w.r.t. x, y for any fixed θ by our choice of $q(\cdot)$. So by Dominated Convergence, the numerator of (3.1.8) is

$$\int\int q(x)q(y)L(x,\theta)\lim_n|e^{-a_nL(x,\theta)}-e^{-a_nL(y,\theta)}|dxdy = 0, \tag{3.1.9}$$

since $a_n \to 0$, so for fixed x and θ, $\lim_n |e^{-a_nL(x,\theta)}-e^{-a_nL(y,\theta)}| = 0$. For the denominator in (3.1.8), the integrand is upper bounded by $q(t)$, which is integrable w.r.t. t, so by Dominated Convergence again,

$$\lim_n \int q(t)e^{-a_nL(t,\theta)}dt = \int \lim_n q(t)e^{-a_nL(t,\theta)}dt = 1. \tag{3.1.10}$$

Now, by (3.1.7), (3.1.8), (3.1.9) and (3.1.10), the limit of the left hand side of (3.1.7) is zero, i.e. $p_n(x^n|\theta) \in \mathcal{P}_n$ for all large n.

Step 2: Now we prove the assertion of the theorem. Let

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

be the mixture of $p_n(x^n|\theta)$ with respect to the prior $w(\theta)$ and write $q(x^n) = \prod_{i=1}^n q(x_i)$.

By the definition of p_n^*, its posterior is the closest to $w(\theta)$ in the expected Kullback-Leibler distance among all the posteriors based on any other probability densities in $\mathcal{P}_n$. We have

$$0 \le E_{m_{p_n^*}} D(w_{p_n^*}(\cdot|X^n)||w(\cdot)) \le E_{m_{p_n}} D(w_{p_n}(\cdot|X^n)||w(\cdot)) \tag{3.1.11}$$

$$= \int\int w(\theta)p_n(x^n|\theta)\log\left(\frac{p_n(x^n|\theta)}{m_{p_n}(x^n)}\right)dx^n d\theta$$

$$= \int\int w(\theta)p_n(x^n|\theta)\log\left(\frac{q(x^n)e^{-L_n(x^n,\theta)}}{\int q(y^n)e^{-L_n(y^n,\theta)}dy^n}\right)dx^n d\theta$$

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

$$= -\int\int w(\theta)p_n(x^n|\theta)L_n(x^n,\theta)dx^n d\theta \tag{3.1.12}$$

$$-\int w(\theta)\log\left(\int q(y^n)e^{-L_n(y^n,\theta)}dy^n\right)d\theta \tag{3.1.13}$$

$$-\int\int w(\theta)p_n(x^n|\theta)\log\left(\int \frac{e^{-L_n(x^n,\xi)}w(\xi)}{\int q(y^n)e^{-L_n(y^n,\xi)}dy^n}d\xi\right)dx^n d\theta. \tag{3.1.14}$$

Term (3.1.12) is $-na_nI(a_n) \to -sb$, and we shall show (3.1.13) $\to sb$ and (3.1.14) $\to 0$.

For (3.1.13), since $-\log(\cdot)$ is convex , we have that for any θ and n,

$$0 < -\log\left(\int q(y^n)e^{-L_n(y^n,\theta)}dy^n\right) = -\log[E_q(e^{-L_n(Y^n,\theta)})]$$

$$\le E_q[-\log(e^{-L_n(Y^n,\theta)})] = E_q(L_n(Y^n,\theta))$$

$$= na_n\int q(y)L(y,\theta)dy < \infty. \tag{3.1.15}$$

Denote the integral in the right hand side of (3.1.15) by $a(\theta)$. Now, $a(\theta)$ is integrable w.r.t. $W(\cdot)$. Indeed, $\int a(\theta)W(d\theta) = \int\int w(\theta)q(y)L(y,\theta)dyd\theta = b < \infty$. By the strong law of large numbers, we have that for all θ, $L_n(Y^n,\theta) \to sa(\theta)$, almost surely with respect to q. So, for any fixed θ, $\epsilon > 0$, when n is large enough, we have that

$$e^{-(sa(\theta)+\epsilon)} \le e^{-L_n(Y^n,\theta)} \le e^{-(sa(\theta)-\epsilon)} \tag{3.1.16}$$

with high $q(\cdot)$ probability.

Let U be the set of Y^n's such that (3.1.16) holds. For n large, we have $E_q\chi_{U^c} \leq \epsilon$, and

$$E_q(e^{-L_n(Y^n,\theta)}) = E_q(e^{-L_n(Y^n,\theta)}\chi_U) + E_q(e^{-L_n(Y^n,\theta)}\chi_{U^c}).$$

Because $e^{-L_n(Y^n,\theta)}$ is bounded, we now have

$$e^{-(sa(\theta)+\epsilon)} - \epsilon \leq E_q[e^{-L_n(Y^n,\theta)}] \leq e^{-(sa(\theta)-\epsilon)} + \epsilon.$$

Since $\epsilon > 0$ can be arbitrarily small, we get

$$\int q(y^n)e^{-L_n(y^n,\theta)}dy^n \to e^{-sa(\theta)}, \tag{3.1.17}$$

for each θ. Hence by (3.1.17) and the Dominated Convergence Theorem, expression (3.1.13) converges to

$$-\int \log(e^{-sa(\theta)})W(d\theta) = sb. \tag{3.1.18}$$

So, as n goes to infinity, (3.1.12) and (3.1.13) will cancel each other.

To complete the proof we only need to prove (3.1.14) tends to zero as n goes to infinity, and by the non-negativity of the Kullback-Leibler distance, we only need to show that (3.1.14) is non-positive.

Recall

$$-\log x \leq x^{-1} - 1 \tag{3.1.19}$$

and that expectations can be written w.r.t. $q(\cdot)$ rather than $p_n(\cdot|\theta)$. We have that, for all n, (3.1.14) is

$$\begin{aligned}
&-\int w(\theta)E_q\Big\{\frac{e^{-L_n(X^n,\theta)}}{E_q[e^{-L_n(Y^n,\theta)}]}\log\Big(\int\frac{w(\xi)e^{-L_n(X^n,\xi)}}{E_q[e^{-L_n(Y^n,\xi)}]}d\xi\Big)\Big\}d\theta \\
&\leq \int w(\theta)E_q\Big\{\frac{e^{-L_n(X^n,\theta)}}{E_q[e^{-L_n(Y^n,\theta)}]}[\Big(\int\frac{w(\xi)e^{-L_n(X^n,\xi)}}{E_q[e^{-L_n(Y^n,\xi)}]}d\xi\Big)^{-1} - 1]\Big\}d\theta \\
&= E_q\Big\{\int\frac{w(\theta)e^{-L_n(X^n,\theta)}}{E_q[e^{-L_n(Y^n,\theta)}]}d\theta\Big(\int\frac{w(\xi)e^{-L_n(X^n,\xi)}}{E_q[e^{-L_n(Y^n,\xi)}]}d\xi\Big)^{-1}\Big\} - 1 = 0.
\end{aligned} \tag{3.1.20}$$

Thus we have

$$0 \leq \overline{\lim}E_{m_{p_n^*}}D(w_{p_n^*}(\cdot|X^n)||w(\cdot)) \leq \overline{\lim}E_{m_{p_n}}D(w_{p_n}(\cdot|X^n)||w(\cdot)) \leq 0. \quad \square$$

Comment: This theorem shows that, asymptotically, the n-dimensional MIL does not in

fact updates the prior at all. In this sense the MIL is minimally informative. Also one may use the product of n-fold 1-dimensional MILs and do regular Bayesian updating given the data for independent observations. We have seen in Proposition 1.5.1 that the product of marginals is the product density closest in Kullback-Leibler distance to a joint density. Accordingly, we may use a product of unidimensional MIL's when the dependence is believed to be slight or absent. If the dependence cannot be ignored we have nevertheless done the best possible subject to dependence.

In cases where the data may be assumed *iid*, we expect to get consistency results for the MIL parallel to those for usual likelihoods. Here we only consider these questions heuristically from a Bayesian standpoint and may investigate them in detail in our future studies. Specifically, assume the data are *iid* and model their common distribution by the MIL, *i.e.* choose

$$p(x^n|\theta) = \prod_{i=1}^{n} C(\theta) m(x_i) e^{-\lambda L(x_i,\theta)},$$

where

$$C(\theta) = \left(\int m(x) e^{-\lambda L(x,\theta)} dx \right)^{-1}$$

is the normalizing constant. Now, the log likelihood is

$$G(\theta|\mathbf{x}) = n \log C(\theta) - \lambda \sum_{i=1}^{n} L(x_i,\theta) + \sum_{i=1}^{n} m(x_i).$$

So the *m.l.e.* $\hat{\theta}_n$ of θ based on the MIL can be obtained by solving the equation

$$0 = \Psi(\mathbf{x}, \hat{\theta}) \equiv \frac{\partial G(\theta|\mathbf{x})}{\partial \theta}\Big|_{\theta=\hat{\theta}_n}.$$

Denote the MIL given θ by p^*_θ and the true density of X by p_{θ_0}, here we assume the same parametrization for both p^*_θ and p_{θ_0}. Essentially, we are using the likelihood equation from the MIL as an estimating equation whose solution is the wrong model *m.l.e.* (under p^*_θ). Since such estimators are typically consistent and asymptotically normal even if their asymptotic variance is higher than the Fisher information. Let

$$\xi_{\theta_0}(\theta) = E_{p^*_\theta}\left(\frac{\partial}{\partial \theta} L(X,\theta)\right) - E_{p_{\theta_0}}\left(\frac{\partial}{\partial \theta} L(X,\theta)\right).$$

By modifying the proof of Jørgensen and Labouriau (1994), we have the following consistency result of the *m.l.e.* based on the MIL.

Theorem 3.1.2. Assume $C(\theta)$ and $\frac{\partial}{\partial\theta}L(X,\theta)$ exist and are continuous in θ almost everywhere with respect to P_{θ_0}, and that there exists a $\delta_0 > 0$ such that for all $\theta \in (\theta_0 - \delta_0, \theta_0)$, $\xi_{\theta_0}(\theta) > 0$, and for all $\theta \in (\theta_0, \theta_0 + \delta_0)$, $\xi_{\theta_0}(\theta) < 0$. Then, there exists a sequence of roots $\{\hat{\theta}_n\}$ of $\Psi(\mathbf{x}, \hat{\theta})$ such that

$$\hat{\theta}_n \overset{P_{\theta_0}}{\to} \theta_0, \qquad as \quad n \to \infty.$$

Proof: First note

$$\frac{\partial}{\partial\theta}\log C(\theta) = -\lambda\frac{\int \frac{\partial}{\partial\theta}L(x,\theta)m(x)e^{-\lambda L(x,\theta)}dx}{C(\theta)} = -\lambda E_{p^*_\theta}(\frac{\partial}{\partial\theta}L(X,\theta)),$$

so

$$\begin{aligned} E_{p_{\theta_0}}\Psi(\mathbf{X},\theta) &= -\frac{\partial}{\partial\theta}\log C(\theta) - \lambda\int\frac{\partial}{\partial\theta}L(x,\theta)p(x|\theta_0)dx \\ &= \lambda\xi_{\theta_0}(\theta). \end{aligned}$$

Thus, take $\delta \in (0, \delta_0)$, by the strong law of large numbers

$$\frac{1}{n}\Psi(\mathbf{X}, \theta_0 - \delta) \overset{P_{\theta_0}}{\to} \lambda\xi_{\theta_0}(\theta_0 - \delta) > 0,$$

and

$$\frac{1}{n}\Psi(\mathbf{X}, \theta_0 + \delta) \overset{P_{\theta_0}}{\to} \lambda\xi_{\theta_0}(\theta_0 + \delta) < 0,$$

as $n \to \infty$. Hence for large n we have

$$\Psi(\mathbf{X}, \theta_0 - \delta) > 0 \qquad \text{and} \qquad \Psi(\mathbf{X}, \theta_0 + \delta) < 0.$$

By the continuity of $\Psi(\mathbf{X}, \theta)$, there exists a root $\hat{\theta}_n(\delta)$ of $\Psi(\mathbf{X}, \theta) = 0$ in the interval $(\theta_0 - \delta, \theta_0 + \delta)$ such that

$$P_{\theta_0}\Big(|\hat{\theta}_n(\delta) - \theta_0| < \delta\Big) \to 1,$$

as $n \to \infty$. Now, instead of $\hat{\theta}_n(\delta)$, we take the root $\hat{\theta}_n$ which is closest to θ_0, this root does not depend on δ and also satisfy

$$P_{\theta_0}\Big(|\hat{\theta}_n - \theta_0| < \delta\Big) \to 1. \qquad \square$$

Next we state a well known result for the asymptotic normality of the solution of an

estimating equation. We first recall the definition of *regular inference function.* An inference function $\Psi(\mathbf{X}, \theta)$ is regular if and only if for all θ

i) $E_{p_\theta}\Psi(\mathbf{X}, \theta) = 0$;

ii) $\partial\Psi(\mathbf{X}, \theta)/\partial\theta$ exists for μ-almost all x, where μ is the common σ-finite dominating measure for the likelihood: $p(\cdot|\theta) = dP(\cdot|\theta)/d\mu$;

iii) The order of the integration and differentiation may be changed:

$$\frac{d}{d\theta}\int \Psi(x, \theta)p(x|\theta)\mu(dx) = \int \frac{\partial}{\partial\theta}[\Psi(x, \theta)p(x|\theta)]\mu(dx);$$

iv) $0 < E_{p_\theta}\{\Psi^2(X, \theta)\} < \infty$;

v) $0 < E_{p_\theta}\{\partial\Psi^2(X, \theta)/\partial\theta\} < \infty$.

Within the context of *Estimating Equations* (see, for example, Godambe, 1960 or Jørgensen and Labouriau, 1994), we know that under the above regularity conditions

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{L} N(0, \sigma^2(\theta)) \quad \text{as } n \to \infty,$$

where the asymptotic variance is given by the Godambe information

$$\sigma^{-2}(\theta) = \frac{E^2_{P_\theta}\frac{\partial}{\partial\theta}\Psi(\mathbf{x}, \hat{\theta})}{E_{P_\theta}\Psi^2(\mathbf{x}, \hat{\theta})}.$$

Now, we have used an MIL to generate an estimator, its *m.l.e.* is consistent and asymptotically normal. In principle, we can examine the optimality of this estimator in terms of the Godambe information. However, for the present, we note that the above results on consistent and asymptotical normality suggest, but do not prove, that the posterior density formed from an MIL concentrates asymptotically at the true value of the parameter in a mode of convergence defined by the true model, i.e.,

$$w(\theta|X^n) \xrightarrow{P_{\theta_0}} \theta_0.$$

This conjecture is supported by Strasser (1981) who demonstrate that Bayes posterior consistency is weaker than frequentist *m.l.e.* consistency. Indeed, Strasser showed that any set of conditions ensuring the *m.l.e.* also ensures posterior concentration. In the present context, we would want to use Laplace's method of integration on $m(x^n)$ at $\hat{\theta}_n$ to extend

Walker's proof. Indeed, a modification of Walker's (1969) proof should give the desired consistency and asymptotic normality of the posterior formed from an MIL.

3.2 Small Sample Properties of MIL

Now, we turn to a non-asymptotic sense in which the MIL as we have defined is minimally informative. Let $p_n^*(x^n|\theta)$ be the MIL from $\mathcal{P}_n$ based on w and let $w_{p_n^*}(\theta|X^n)$ be the posterior formed from $w(\theta)$ and $p_n^*(x^n|\theta)$. Following Csiszar (1975), the tangent hyperplane determined by $w(\theta)$ and $w_{p_n^*}(\theta|x^n)$ is given by

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

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

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

The two tangent hyperplane divide the whole space of priors into subspaces, one of them which we denote by $S(x^n, w, w_{p_n^*}, w_p)$ is

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

$$\int w'(\theta) \log \frac{w_p(\theta|x^n)}{w(\theta)} d\theta \geq D(w_p(\cdot|x^n)||w(\cdot))\Big\}.$$

Let $S_n(w, w_{p_n^*}, w_p) = \cap_{x^n} S(x^n, w, w_{p_n^*}, w_p)$, which is a subspace in the prior space independent of data. Let w_0 be a member of $S_n(w, w_{p_n^*}, w_p)$. We show that, on average, using the MIL likelihood $p_n^*(x^n|\theta)$ to update $w(\cdot)$ gives a posterior $w_{p^*}(\theta|x^n)$ further from w_0 in Kullback-Leibler distance than any other likelihood $p(x^n|\theta)$ in $\mathcal{P}$ does. i.e., w_{p^*} is further away from any untrue w_0 than any other w_p.

To get a pointwise result in the above sense, let

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

Since $E_{m_{p_n^*}} D(w_{p_n^*}(\cdot|X^n)||w(\cdot)) \leq E_{m_{p_n}} D(w_p(\cdot|X^n)||w(\cdot))$, it is likely that for some x^n, $U(w, w_{p_n^*}, w_p) \neq \phi$.

Theorem 3.2.1.

(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

$$D(w_0(\cdot)||w_{p_n^*}(\cdot|x^n)) \geq D(w_0(\cdot)||w_p(\cdot|x^n)). \tag{3.2.1}$$

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

$$E_{m_{p_n^*}} D(w_0(\cdot)||w_{p_n^*}(\cdot|X^n)) \geq E_{m_p} D(w_0(\cdot)||w_p(\cdot|X^n)). \tag{3.2.2}$$

Proof: (i) Since

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

and

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

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

$$D(w_0(\cdot)||w(\cdot)) \leq D(w_0(\cdot)||w_{p_n^*}(\cdot|x^n)) + D(w_{p_n^*}(\cdot|x^n)||w(\cdot)),$$

and that

$$D(w_0(\cdot)||w(\cdot)) \geq D(w_0(\cdot)||w_p(\cdot|x^n)) + D(w_p(\cdot|x^n)||w(\cdot)).$$

Since $D(w_{p_n^*}(\cdot|x^n)||w(\cdot)) \leq D(w_p(\cdot|x^n)||w(\cdot))$, so for $x^n \in U(w, w_{p_n^*}, w_p)$, by the above two inequalities we have

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

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

$$D(w_0(\cdot)||w(\cdot)) \leq D(w_0(\cdot)||w_{p_n^*}(\cdot|X^n)) + D(w_{p_n^*}(\cdot|X^n)||w(\cdot)),$$

and that

$$D(w_0(\cdot)||w(\cdot)) \geq D(w_0(\cdot)||w_p(\cdot|X^n)) + D(w_p(\cdot|X^n)||w(\cdot)).$$

Taking expectations we have

$$D(w_0(\cdot)||w(\cdot)) \leq E_{m_{p_n^*}} D(w_0(\cdot)||w_{p_n^*}(\cdot|X^n)) + E_{m_{p_n^*}} D(w_{p_n^*}(\cdot|X^n)||w(\cdot)),$$

and

$$D(w_0(\cdot)||w(\cdot)) \geq E_{m_p} D(w_0(\cdot)||w_p(\cdot|X^n)) + E_{m_p} D(w_p(\cdot|X^n)||w(\cdot)).$$

By definition of $p_n^*(x^n|\theta)$ we have

$$E_{m_{p_n^*}} D(w_{p_n^*}(\cdot|X^n)||w(\cdot)) \leq E_{m_p} D(w_p(\cdot|X^n)||w(\cdot)),$$

so we have

$$E_{m_{p_n^*}} D(w_0(\cdot)||w_{p_n^*}(\cdot|X^n)) \geq E_{m_p} D(w_0(\cdot)||w_p(\cdot|X^n)). \quad \square$$

3.3 Behavior of the MIL for large and small values of λ

Clearly, the MIL depends on the choice of λ (or equivalently l) used to define $\mathcal{P}$. In this section, we prove two theorems that show how the size of λ affects the behavior of the MIL. To emphasize the dependence of the MIL on λ, we write $p_\lambda^*(x|\theta)$ for the MIL, and we denote the corresponding marginal density by $m_\lambda^*(x)$, and the corresponding posterior density by $w_\lambda^*(\theta|x)$. Let $\zeta(\theta)$ be the degenerate probability mass function at θ, $\xrightarrow{D}$ denote convergence in distribution, and $\mu(\cdot)$ be the Lebesgue measure on R^1. First, we characterize the behavior of the MIL for λ large. For simplicity, we only prove the results for one dimensional data case, the proofs are also valid for n-dimensional data case.

Theorem 3.3.1. (i) The marginal density for X from $p_\lambda^*(x|\theta)$ is $m_\lambda^*(x)$, i.e. ,

$$m_\lambda^*(x) = \int p_\lambda^*(x|\theta) w(\theta) d\theta.$$

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 $L(x,\theta) = \underline{r}(|x-\theta|)$ is strictly increasing in $|x-\theta|$, with $\underline{r}(0) = 0$, and $\underline{r}(s+t) \geq \underline{r}(s) + \underline{r}(t)$, for all $s \geq 0, t \geq 0$. Then as $\lambda \to \infty$, we have the following

(ii) The marginal density for the data satisfies

$$m_\lambda^*(x) \to \begin{cases} w(x), & \text{if} \quad x \in S \cap C \\ 0, & \text{for} \quad a.e. \ \mu(\cdot) \ \ x \in S^c, \end{cases} \tag{3.3.1}$$

(iii) the MIL densities satisfy

$$p_\lambda^*(x|\theta) \xrightarrow{D} \zeta(\theta), \quad \forall \theta \in S^0 \cap C, \tag{3.3.2}$$

$$\text{and} \quad w_\lambda^*(\theta|x) \xrightarrow{D} \zeta(x), \quad \forall x \in S^0 \cap C. \tag{3.3.3}$$

Proof: (i) Since

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

if $p^*_\lambda(x|\theta) > 0$, then $m^*_\lambda(x) > 0$, 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 (1.2.1.5). If $p^*_\lambda(x|\theta) = 0$, then $m^*_\lambda(x) = 0$, we still have

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

(ii) To prove the result, recall that $m^*_\lambda(x)$ is determined by

$$\int \frac{e^{-\lambda L(x,\theta)}w(\theta)}{\int m^*_\lambda(y)e^{-\lambda L(y,\theta)}dy}d\theta \leq 1, \tag{3.3.4}$$

where equality holds for $x \in S_\lambda$, where S_λ is the support of $m^*_\lambda(\cdot)$. Since $w(\cdot)$ and $m^*_\lambda(\cdot)$ integrable, they are continuous almost everywhere, without loss of generality we restrict to the continuity points of $w(\cdot)$. Take $\delta > 0$ small, then $\forall x \in S \cap C$, by (3.3.4) we have

$$1 \geq \int_{[x-\delta,x+\delta]} \frac{e^{-\lambda L(x,\theta)}w(\theta)}{\int_{[\theta-\delta,\theta+\delta]} m^*_\lambda(y)e^{-\lambda L(y,\theta)}dy + (1+h(\lambda,\theta))}d\theta$$

$$+ \int_{[x-\delta,x+\delta]^c} \frac{e^{-\lambda L(x,\theta)}w(\theta)}{\int m^*_\lambda(y)e^{-\lambda L(y,\theta)}dy}d\theta, \tag{3.3.5}$$

where

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

We show that the second term on the right hand side of (3.3.5) tends to zero as λ tends to infinity, and $h(\lambda,\theta)$ is negligiblely small for large λ, so the remaining part of (3.3.5) gives a ratio which is approximately "$w(x)/m^*_\infty(x)$", and equals 1. There are six steps in the proof.

Step 1: Show that the second term in (3.3.5) goes to zero as λ increases to infinity, i.e.

$$\int_{[x-\delta,x+\delta]^c} \frac{e^{-\lambda L(x,\theta)}w(\theta)}{\int m^*_\lambda(y)e^{-\lambda L(y,\theta)}dy}d\theta \quad \to 0, \quad \text{as} \quad \lambda \to \infty.$$

Indeed, the second term in (3.3.5) equals

$$\int_{-\infty}^{x-\delta} \frac{e^{-\lambda \underline{r}(x-\theta)}w(\theta)}{\int m^*_\lambda(y)e^{-\lambda L(y,\theta)}dy}d\theta + \int_{x+\delta}^{\infty} \frac{e^{-\lambda \underline{r}(\theta-x)}w(\theta)}{\int m^*_\lambda(y)e^{-\lambda L(y,\theta)}dy}d\theta.$$

Since $\forall\theta \in (-\infty, x-\delta], \underline{r}(x-\theta) = \underline{r}(\delta+x-\delta-\theta) \geq \underline{r}(\delta)+\underline{r}(x-\delta-\theta) = \underline{r}(\delta)+L(x-\delta,\theta)$, and $\forall\theta \in [x+\delta,\infty), L(\theta-x) = L(\delta+\theta-(x+\delta)) \geq L(\delta)+L(\theta-(x+\delta)) = L(\delta)+L(x+\delta,\theta)$, so the second term in (3.3.5) is bounded from above by

$$e^{-\lambda L(\delta)}\Big(\int_{-\infty}^{x-\delta} \frac{e^{-\lambda L(x-\delta,\theta)}w(\theta)}{\int m_\lambda^*(y)e^{-\lambda L(y,\theta)}dy}d\theta + \int_{x+\delta}^{\infty} \frac{e^{-\lambda L(x+\delta,\theta)}w(\theta)}{\int m_\lambda^*(y)e^{-\lambda L(y,\theta)}dy}d\theta\Big)$$

$$\leq e^{-\lambda L(\delta)}\Big(\int \frac{e^{-\lambda L(x-\delta,\theta)}w(\theta)}{\int m_\lambda^*(y)e^{-\lambda L(y,\theta)}dy}d\theta + \int \frac{e^{-\lambda L(x+\delta,\theta)}w(\theta)}{\int m_\lambda^*(y)e^{-\lambda L(y,\theta)}dy}d\theta\Big)$$

$$\leq 2e^{-\lambda \underline{r}(\delta)} \to 0, \quad \text{as} \quad \lambda \to \infty,$$

since by (3.3.4)

$$\int \frac{e^{-\lambda L(x\pm\delta,\theta)}w(\theta)}{\int m_\lambda^*(y)e^{-\lambda L(y,\theta)}dy}d\theta \leq 1,$$

this completes the proof of Step 1.

Now, from Step 1 and (3.3.5), we have

$$1 \geq \int_{[x-\delta,x+\delta]} \frac{e^{-\lambda L(x,\theta)}w(\theta)}{\int_{[\theta-\delta,\theta+\delta]} m_\lambda^*(y)e^{-\lambda L(y,\theta)}dy(1+h(\lambda,\theta))}d\theta + o(1), \tag{3.3.6}$$

where o(1) goes to zero as $\lambda \to \infty$. For fixed θ, it is easy to show $h(\lambda,\theta) \to 0$, but this may not hold uniformly for $\theta \in [x-\delta, x+\delta]$ the domain of integration in (3.3.6). So, we split $[x-\delta, x+\delta]$ into a "good" set on which $h(\lambda,\theta)$ is uniformly small, and a "bad" set on which $h(\lambda,\theta)$ is not small. We show the "bad" set is negligible in Lebesgue measure for large λ. Formally, let $\epsilon > 0$, and let $A_\lambda = \{\theta \in [x-\delta, x+\delta] \mid h(\lambda,\theta) \geq \epsilon\}$. If A_λ is contained in a sub-interval of $[x-\delta, x+\delta]$ which excludes x, we can reduce δ and there is nothing to prove, otherwise the Lebesgue measure of A_λ is controlled as follows.

Step 2: We show that as $\lambda \to \infty$,

$$\mu(A_\lambda \cap [x-\delta/2, x+\delta/2]) = O(e^{-\lambda(L(\delta)-L(\delta/2))}).$$

By reducing the domain of integration in (3.3.5) we have

$$1 \geq \int_{[x-\delta,x+\delta]\cap A_\lambda} \frac{e^{-\lambda L(x,\theta)}w(\theta)}{\int_{[\theta-\delta,\theta+\delta]} m_\lambda^*(y)e^{-\lambda L(y,\theta)}dy + \int_{[\theta-\delta,\theta+\delta]^c} m_\lambda^*(y)e^{-\lambda L(y,\theta)}dy}d\theta$$

$$\geq \int_{[x-\delta/2,x+\delta/2]\cap A_\lambda} \frac{e^{-\lambda L(x,\theta)}w(\theta)}{\epsilon^{-1}+1)\int_{[\theta-\delta,\theta+\delta]^c} m_\lambda^*(y)e^{-\lambda L(y,\theta)}dy}d\theta, \tag{3.3.7}$$

since $h(\lambda,\theta) \leq \epsilon$ on A_λ, i.e. we have

$$\int_{[\theta-\delta,\theta+\delta]} m_\lambda^*(y)e^{-\lambda L(y,\theta)}dy \leq \frac{1}{\epsilon}\int_{[\theta-\delta,\theta+\delta]^c} m_\lambda^*(y)e^{-\lambda L(y,\theta)}dy.$$

We can bound the $e^{-\lambda L(x,\delta)}$ in the numerator of (3.3.7) from below by $e^{-\lambda L(\delta/2)}$, and bound the $e^{-\lambda L(y,\delta)}$ in the denominator of (3.3.7) from above by $e^{-\lambda L(\delta)}$. This means that (3.3.7) is bounded below by

$$\begin{aligned}
&\frac{e^{-\lambda \underline{r}(\delta/2)}}{(\epsilon^{-1}+1)e^{-\lambda \underline{r}(\delta)}}\int_{[x-\delta/2,x+\delta/2]\cap A_\lambda}\frac{w(\theta)}{\int_{[\theta-\delta,\theta+\delta]^c} m_\lambda^*(y)dy}d\theta \\
&\geq \frac{e^{\lambda(\underline{r}(\delta)-\underline{r}(\delta/2))}}{\epsilon^{-1}+1}\int_{[x-\delta/2,x+\delta/2]\cap A_\lambda} w(\theta)d\theta \\
&\geq \frac{e^{\lambda(\underline{r}(\delta)-\underline{r}(\delta/2))}}{\epsilon^{-1}+1}\frac{w(x)}{2}\mu([x-\delta/2,x+\delta/2]\cap A_\lambda),
\end{aligned}$$

since the continuity of w at x guarantees that for δ small we have $w(\theta) \geq w(x)/2$, when $\theta \in [x-\delta/2, x+\delta/2]$. Step 2 now follows.

Now by Step 1, and the definition of A_λ, we get

$$\begin{aligned}
1 \geq &\int_{[x-\delta/2,x+\delta/2]\cap A_\lambda^c}\frac{e^{-\lambda L(x,\theta)}w(\theta)}{\int_{[\theta-\delta,\theta+\delta]} m_\lambda^*(y)e^{-\lambda L(y,\theta)}dy(1+o(1))}d\theta \\
&+\int_{[x-\delta/2,x+\delta/2]\cap A_\lambda}\frac{e^{-\lambda L(x,\theta)}w(\theta)}{\int m_\lambda^*(y)e^{-\lambda L(y,\theta)}dy}d\theta + o(1), \qquad (3.3.8)
\end{aligned}$$

as λ tends to infinity. We will see that the second term in the right hand side of (3.3.8) tends to zero as λ tends to infinity. Also, by the mean value theorem for integrals, we will see that the first term of the right hand side of (3.3.8) becomes $w(\zeta)$ over $m_\lambda^*(\eta)$ times an integral 1 to 1.

Step 3: As $\lambda \to \infty$,

$$\int_{[x-\delta/2,x+\delta/2]\cap A_\lambda^c}\frac{e^{-\lambda L(x,\theta)}}{\int_{[\theta-\delta,\theta+\delta]} e^{-\lambda L(y,\theta)}dy}d\theta \to 1.$$

We start by showing that as $\lambda \to \infty$,

$$\int_{[x-\delta/2,x+\delta/2]\cap A_\lambda}\frac{e^{-\lambda L(x,\theta)}}{\int_{[\theta-\delta,\theta+\delta]} e^{-\lambda L(y,\theta)}dy}d\theta \to 0. \qquad (3.3.9)$$

Indeed, let $0 < \delta' < \delta$ satisfy $L(\delta') < L(\delta) - L(\delta/2)$. Now, the left hand side of (3.3.9) is bounded from above by

$$\int_{[x-\delta/2,x+\delta/2]\cap A_\lambda}\frac{e^{-\lambda L(x,\theta)}}{\int_{[\theta-\delta',\theta+\delta']} e^{-\lambda L(y,\theta)}dy}d\theta.$$

Since the numerator is bounded above by 1, and for $y \in [\theta-\delta', \theta+\delta']$, we have $L(y,\theta) \leq L(\delta')$, the last expression is bounded above by

$$\frac{\mu([x-\delta/2, x+\delta/2] \cap A_\lambda)}{2\delta' e^{-\lambda \underline{r}(\delta')}} \leq \frac{K}{2\delta'} e^{-\lambda[\underline{r}(\delta)-\underline{r}(\delta/2)-\underline{r}(\delta')]} \to 0,$$

as $\lambda \to \infty$, for some constant K.

Now by adding and subtracting the left hand side of (3.3.9) to the left hand side of Step 3, the integral in Step 3 becomes

$$\int_{[x-\delta/2, x+\delta/2]} \frac{e^{-\lambda L(x,\theta)}}{\int_{[\theta-\delta,\theta+\delta]} e^{-\lambda L(y,\theta)} dy} d\theta + o(1)$$

$$= \int_{[x-\delta, x+\delta]} \frac{e^{-\lambda L(x,\theta)}}{\int_{[\theta-\delta,\theta+\delta]} e^{-\lambda L(y,\theta)} dy} d\theta - \int_{\frac{\delta}{2}<|x-\theta|\leq\delta} \frac{e^{-\lambda L(x,\theta)}}{\int_{[\theta-\delta,\theta+\delta]} e^{-\lambda L(y,\theta)} dy} d\theta + o(1)$$

$$= \int_0^\delta \frac{e^{-\lambda \underline{r}(t)}}{\int_0^\delta e^{-\lambda \underline{r}(s)} ds} dt - \int_{\delta/2}^\delta \frac{e^{-\lambda \underline{r}(t)}}{\int_0^\delta e^{-\lambda \underline{r}(s)} ds} dt + o(1)$$

$$= 1 - \int_0^{\delta/2} \frac{e^{-\lambda \underline{r}(t+\delta/2)}}{\int_0^\delta e^{-\lambda \underline{r}(s)} ds} dt + o(1).$$

Since the absolute value of the second term is not greater than

$$e^{-\lambda \underline{r}(\delta/2)} \frac{\int_0^{\delta/2} e^{-\lambda \underline{r}(t)} dt}{\int_0^\delta e^{-\lambda \underline{r}(s)} ds} \leq e^{-\lambda \underline{r}(\delta/2)} \to 0, \quad \text{as} \quad \lambda \to \infty,$$

Step 3 is complete.

To ensure that the equality is achieved in (3.3.5), we first verify that $m_\lambda^*(x)$ is positive in a stronger sense.

Step 4: We show that $\forall x \in S^0$,

$$\underline{\lim}_{\lambda\to\infty} m_\lambda^*(x) > 0. \tag{3.3.10}$$

To prove Step 4, note that by (3.3.8) we have (we will show later that the second term in (3.3.8) tends to zero)

$$1 \geq \int_{[x-\delta/2, x+\delta/2] \cap A_\lambda^c} \frac{e^{-\lambda L(x,\theta)} w(\theta)}{\int_{[\theta-\delta,\theta+\delta]} m_\lambda^*(y) e^{-\lambda L(y,\theta)} dy (1+o(1))} d\theta + o(1)$$

$$= \frac{w(\zeta)}{m_\lambda^*(\eta)} \int_{[x-\delta/2, x+\delta/2] \cap A_\lambda^c} \frac{e^{-\lambda L(x,\theta)}}{\int_{[\theta-\delta,\theta+\delta]} e^{-\lambda L(y,\theta)} dy (1+o(1))} d\theta$$

$$= \frac{w(\zeta)}{m_\lambda^*(\eta)} \frac{1}{(1+o(1))} \int_{[x-\delta/2, x+\delta/2] \cap A_\lambda^c} \frac{e^{-\lambda L(x,\theta)}}{\int_{[\theta-\delta,\theta+\delta]} e^{-\lambda L(y,\theta)} dy} d\theta + o(1),$$

by using the median point theorem of integration twice, where $\zeta \in [x-\delta/2, x+\delta/2] \cap A_\lambda^c$, and $\eta \in [\theta-\delta, \theta+\delta] \subset [x-\frac{3}{2}\delta, x+\frac{3}{2}\delta]$, since both $w(\cdot)$ and $m_\lambda^*(\cdot)$ are continuous at x. By Step 3, the last expression is

$$\frac{w(\zeta)}{m_\lambda^*(\eta)(1+o(1))} + o(1). \tag{3.3.11}$$

Now, Step 4 follows by way of contradiction: Suppose $m_\lambda^*(x) \to 0$, as $\lambda \to \infty$. Then, there exists $\delta > 0$, and λ so large that $w(\zeta)/m_\lambda^*(\eta) > 1$, which is impossible by the above inequality. This means we must have $\liminf_{\lambda\to\infty} m_\lambda^*(x) > 0$, i.e. Step 4 is completed. Note that the result of Step 4 applies to each $y \in [x-\delta, x+\delta] \subset S$.

Now we prove the second term in the right hand side of (3.3.8) is small as λ increases.

Step 5: As $\lambda \to \infty$, we have that

$$\int_{[x-\delta/2,x+\delta/2]\cap A_\lambda} \frac{e^{-\lambda L(x,\theta)} w(\theta)}{\int m_\lambda^*(y) e^{-\lambda L(y,\theta)} dy} d\theta \to 0. \tag{3.3.12}$$

To see this, let $\epsilon > 0$ and δ' be as in Step 3, and let

$$B_{\lambda,\epsilon} = \{y \in [x-\delta/2-\delta', x+\delta/2+\delta'] \mid m_\lambda^*(y) > \epsilon\}.$$

By choosing ϵ small enough and λ large enough, $B_{\lambda,\epsilon}$ can be made as close to $[x-\delta/2-\delta', x+\delta/2+\delta']$ in $\mu(\cdot)$ measure as we want, i.e. for small ϵ and large λ we have

$$\mu([x-\delta/2-\delta', x+\delta/2+\delta']) - \mu(B_{\lambda,\epsilon}) < \delta'/2.$$

By intersecting the interval with $B_{\lambda,\epsilon}$ and $B_{\lambda,\epsilon}^c$ we get

$$\mu(B_{\lambda,\epsilon} \cap [\theta-\delta', \theta+\delta']) > \frac{\delta'}{2}, \quad \forall \theta \in [x-\delta/2, x+\delta/2].$$

Hence the left hand side of (3.3.12) is bounded above by

$$\int_{[x-\delta/2,x+\delta/2]\cap A_\lambda} \frac{e^{-\lambda L(x,\theta)} w(\theta)}{\int_{[\theta-\delta',\theta+\delta']\cap B_{\lambda,\epsilon}} m_\lambda^*(y) e^{-\lambda L(y,\theta)} dy} d\theta$$

$$\leq \frac{2w(x)}{\epsilon} \int_{[x-\delta/2,x+\delta/2]\cap A_\lambda} \frac{e^{-\lambda L(x,\theta)}}{\int_{[\theta-\delta',\theta+\delta']\cap B_{\lambda,\epsilon}} e^{-\lambda L(y,\theta)} dy} d\theta,$$

since $w(\theta) \leq 2w(x), \forall \theta \in [x-\delta/2, x+\delta/2]$, and $m_\lambda^*(y) \geq \epsilon, \forall y \in B_{\lambda,\epsilon}$. Noting that $e^{-\lambda L(x,\theta)} \leq 1$, and $e^{-\lambda L(y,\theta)} \geq e^{-\lambda \underline{r}(\delta')}$ for $y \in [\theta-\delta', \theta+\delta'] \cap B_{\lambda,\epsilon}$, the last expression can be bounded above by

$$\frac{2w(x)}{\epsilon} \frac{\mu([x-\delta/2, x+\delta/2] \cap A_\lambda)}{e^{-\lambda \underline{r}(\delta')} \inf_{\theta\in[x-\delta/2,x+\delta/2]} \mu(B_\lambda \cap [\theta-\delta', \theta+\delta'])}$$

$$\leq \frac{4w(x)}{\epsilon}\frac{e^{-\lambda[\underline{r}(\delta)-\underline{r}(\delta/2)-\underline{r}(\delta')]}}{\delta'} \to 0, \quad \text{as} \quad \lambda \to \infty,$$

by Step 4. Now, Step 5 is complete.

Now let ζ and η as in (3.3.11), write $w(\zeta) = w(x) + o(1)$, and $m_\lambda^*(\eta) = m_\lambda^*(x) + o(1)$. By definition of $m_\lambda^*(\cdot)$, equality is achieved in (3.3.5) for $x \in S_\lambda$, thus (3.3.11) achieves the upper bound 1, i.e.

$$1 = \frac{w(\zeta)}{m_\lambda^*(\eta)(1+o(1))} + o(1).$$

Letting $\delta \to 0$, we get

$$1 = \frac{w(x)}{\lim_{\lambda\to\infty} m_\lambda^*(x)},$$

proving (3.3.1) for $x \in S \cap C$.

For the final step in the proof, let $x \in S^c$. If for such an x we have $\forall\lambda, m_\lambda^*(x) = 0$, then the conclusion follows. Otherwise, we must have $m_{\lambda_k}^*(x) > 0$ for some sequence $\{\lambda_k\}$ which satisfies $\lim_{k\to\infty}\lambda_k = \infty$ and $\lim_{k\to\infty}\inf m_{\lambda_k}^*(x) > 0$. For this case, it is enough to show $\lim_{k\to\infty}\inf m_{\lambda_k}^*(\cdot) = 0$ a.e. on a small neighborhood of x. Without loss of generality, assume x lies on the left of S.

Step 6: There is a δ such that $0 < \delta < d$, where $d = \inf_{y\in S}|x-y|$ is the distance between x and S so that $m_\lambda^*(\cdot) = 0, a.e.\ \mu(\cdot)$ on $(x, x+\delta]$.

We prove this by way of contradiction. If this is not true, there is a least δ satisfying such that $0 < \delta < d$, and $\lim_{k\to\infty}\inf m_{\lambda_k}^*(x) > 0$ a.e. on $(x, x+\delta]$. Now, let $a > 0$ and write $D_{\lambda_k} = \{y \in [x+\delta/2, x+\delta] \mid m_{\lambda_k}^*(y) \geq a\}$. Choose a so small and k so large that $\mu(D_{\lambda_k}) > \delta/4$.

Since $m_{\lambda_k}^*(x) > 0$, (3.3.4) implies

$$1 = \int_S \frac{e^{-\lambda_k L(x,\theta)}w(\theta)}{\int m_{\lambda_k}^*(y)e^{-\lambda_k L(y,\theta)}dy}d\theta.$$

Let $M > 0$, and let S_M be a closed set satisfying $\sup_{\theta\in S_M} w(\theta) \leq M < \infty$, and

$$\frac{1}{2} \leq \int_{S_M} \frac{e^{-\lambda_k L(x,\theta)}w(\theta)}{\int m_{\lambda_k}^*(y)e^{-\lambda_k L(y,\theta)}dy}d\theta.$$

This gives that

$$\frac{1}{2} \leq \int_{S_M} \frac{e^{-\lambda_k L(x,\theta)}w(\theta)}{\int_{D_{\lambda_k}} m_{\lambda_k}^*(y)e^{-\lambda_k L(y,\theta)}dy}d\theta$$

$$\leq \frac{M}{a}\int_{S_M} \frac{e^{-\lambda_k L(x,\theta)}}{\int_{D_{\lambda_k}} e^{-\lambda_k L(y,\theta)}dy}d\theta.$$

Since $L(x,\theta) \geq L(y,\theta) + \underline{r}(\delta/2), \forall y \in D_{\lambda_k}, \forall \theta \in S_M$, the last expression can be bounded by

$$\frac{M}{a} e^{-\lambda_k \underline{r}(\delta/2)} \int_{S_M} \frac{e^{-\lambda_k L(y,\theta)}}{\int_{D_{\lambda_k}} e^{-\lambda_k L(y,\theta)} dy} d\theta. \tag{3.3.13}$$

By an argument similar to that used in Step 3, we have that

$$\int_{S_M} \frac{e^{-\lambda_k L(y,\theta)}}{\int_{D_{\lambda_k}} e^{-\lambda_k L(y,\theta)} dy} d\theta \to 1, \quad \text{as} \quad k \to \infty.$$

In particular, for large k, the last expression is bounded from above by 2, and the right hand side of (3.3.13) is bounded by

$$\frac{4M}{a} e^{-\lambda_k \underline{r}(\delta/2)} \to 0,$$

which is a contradiction thereby establishing Step 6. This completes the proof of part (i) because we have shown that $\forall x \in S^c$, if $\liminf_{\lambda \to \infty} m^*_\lambda(x) > 0$, then $\lim_{\lambda \to \infty} m^*_\lambda(\cdot) = 0$, a.e. on $(x, x+\delta]$, for some $0 < \delta < d$. i.e.

$$\lim_{\lambda \to \infty} m^*_\lambda(x) = 0, \quad a.e. \quad \mu(\cdot) \quad \text{on} \quad S^c.$$

This last statement is equivalent to (3.3.1) for $x \in S^c$.

(iii) To prove (3.3.2), let $\phi_\lambda(\cdot)$ be the characteristic function of $p^*_\lambda(\cdot|\theta)$. We have

$$\phi_\lambda(t) = \frac{\int m^*_\lambda(x) e^{-\lambda L(x,\theta)} e^{ixt} dx}{\int m^*_\lambda(y) e^{-\lambda L(y,\theta)} dy}$$

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

To simplify the expression, we first prove

$$\frac{\int_{[\theta-\delta,\theta+\delta]^c} m^*_\lambda(y) e^{-\lambda L(y,\theta)} dy}{\int_{[\theta-\delta,\theta+\delta]} m^*_\lambda(y) e^{-\lambda L(y,\theta)} dy} \to 0,$$

as $\lambda \to \infty$. In fact

$$\frac{\int_{[\theta-\delta,\theta+\delta]^c} m^*_\lambda(y) e^{-\lambda L(y,\theta)} dy}{\int_{[\theta-\delta,\theta+\delta]} m^*_\lambda(y) e^{-\lambda L(y,\theta)} dy} \leq \frac{\int_{[\theta-\delta,\theta+\delta]^c} m^*_\lambda(y) e^{-\lambda L(y,\theta)} dy}{\int_{[\theta-\delta/2,\theta+\delta/2]} m^*_\lambda(y) e^{-\lambda L(y,\theta)} dy}.$$

Since $\theta \in S^0$, Step 4 gives that $\lim_{\lambda \to \infty} \inf m^*_\lambda(y) > 0$, pointwise for $y \in [\theta-\delta/2, \theta+\delta/2] \cap S^0$. For $a > 0$, let $D_\lambda = \{y \in [\theta - \delta/2, \theta + \delta/2] \mid m^*_\lambda(y) \geq a\}$. Now, for a small enough

and λ large enough we have $\mu(D_\lambda) > \frac{\delta}{2}$. Since $\forall y \in [\theta-\delta, \theta+\delta]^c$, $e^{-\lambda L(y,\theta)} \leq e^{-\lambda \underline{r}(\delta)}$, $\forall y \in D_\lambda$, $e^{-\lambda L(y,\theta)} \geq e^{-\lambda \underline{r}(\delta/2)}$, the right hand side in the above inequality is bounded by

$$\frac{\int_{[\theta-\delta,\theta+\delta]^c} m_\lambda^*(y) e^{-\lambda L(y,\theta)} dy}{\int_{D_\lambda} m_\lambda^*(y) e^{-\lambda L(y,\theta)} dy}$$

$$\leq \frac{e^{-\lambda \underline{r}(\delta)}}{e^{-\lambda \underline{r}(\delta/2)}} \frac{\int_{[\theta-\delta,\theta+\delta]^c} m_\lambda^*(y) dy}{\int_{D_\lambda} m_\lambda^*(y) dy}$$

$$\leq \frac{2e^{-\lambda[\underline{r}(\delta)-\underline{r}(\delta/2)]}}{a\delta} \to 0, \quad \text{as} \quad \lambda \to \infty.$$

Now by this inequality we have

$$\frac{|\int_{[\theta-\delta,\theta+\delta]^c} m_\lambda^*(x) e^{-\lambda L(x,\theta)} e^{ixt} dx|}{\int_{[\theta-\delta,\theta+\delta]} m_\lambda^*(y) e^{-\lambda L(y,\theta)} dy + \int_{[\theta-\delta,\theta+\delta]^c} m_\lambda^*(y) e^{-\lambda L(y,\theta)} dy}$$

$$\leq \frac{\int_{[\theta-\delta,\theta+\delta]^c} m_\lambda^*(y) e^{-\lambda L(y,\theta)} dy}{\int_{[\theta-\delta,\theta+\delta]} m_\lambda^*(y) e^{-\lambda L(y,\theta)} dy} \to 0, \quad \text{as} \quad \lambda \to \infty.$$

thus we have

$$\phi_\lambda(t) = \frac{\int_{[\theta-\delta,\theta+\delta]} m_\lambda^*(x) e^{-\lambda L(x,\theta)} e^{ixt} dx}{(\int_{[\theta-\delta,\theta+\delta]} m_\lambda^*(y) e^{-\lambda L(y,\theta)} dy)(1+o(1))} + o(1)$$

$$= \frac{\int_{[\theta-\delta,\theta+\delta]} m_\lambda^*(x) e^{-\lambda L(x,\theta)} e^{ixt} dx}{\int_{[\theta-\delta,\theta+\delta]} m_\lambda^*(y) e^{-\lambda L(y,\theta)} dy} + o(1)$$

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

$$\equiv J_1(\lambda,t) + iJ_2(\lambda,t) + o(1).$$

Obviously, for all λ

$$\inf_{x\in[\theta-\delta,\theta+\delta]} \cos(xt) \leq J_1(\lambda,t) \leq \sup_{x\in[\theta-\delta,\theta+\delta]} \cos(xt)$$

and

$$\lim_{\delta\to 0} \inf_{x\in[\theta-\delta,\theta+\delta]} \cos(xt) = \lim_{\delta\to 0} \sup_{x\in[\theta-\delta,\theta+\delta]} \cos(xt) = \cos(\theta t),$$

so, $\lim_{\delta\to 0} J_1(\lambda,t) = \cos(\theta t)$ holds for all λ. Similarly, for all λ, $\lim_{\delta\to 0} J_2(\lambda,t) = \sin(\theta t)$. Thus, if we first let $\delta \to 0$ and then let $\lambda \to \infty$, we get

$$\lim_{\lambda\to\infty} \phi_\lambda(t) = e^{i\theta t},$$

which is the characteristic function of $\zeta(\theta)$.

To prove (3.3.3), let $\psi_\lambda(\cdot)$ be the characteristic function of $w^*_\lambda(\cdot|x)$, then

$$\psi_\lambda(t) = \int \frac{p^*_\lambda(x|\theta)w(\theta)e^{i\theta t}}{m^*_\lambda(x)}d\theta$$

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

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

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

The absolute value of the second term in the last expression is bounded by

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

which tends to zero as λ tends to infinity, by Step 1. By the transformation $r = \theta - x$, for large λ we have

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

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, ζ and η tend to x as δ tends to zero. Therefore, the ratio $w(\zeta)/m^*_\lambda(\eta)$ tends to 1 by (3.3.1). Thus we have

$$\lim_{\lambda\to\infty} \psi_\lambda(t) = e^{ixt}, \quad a.e.\ \ \mu(\cdot)\ \ \text{on}\ \ x \in S,$$

and part (iii) is proved. □

Our next result characterizes the behavior of the MIL for the range of the parameter λ between zero and infinity and relates λ to the l used to define $\mathcal{P}_l$. Let $\underline{r} = \inf_x \int w(\theta)L(x,\theta)d\theta$, $x_0 = \arg\inf_x \int w(\theta)L(x,\theta)d\theta$. 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. The following theorem is proved for one dimensional data and parameter, the results and proof should be the same for random vectors and multi-dimensional parameters.

Theorem 3.3.2. Assume $L(\cdot,\theta)$ is not constant. Then we have:

(i) For $l \in (0, \underline{r}), p^*_\lambda(x|\theta)$ 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 $l \in (0, \underline{r})$, λ and l determine each other uniquely. We can therefore write $l = l(\lambda)$, or $\lambda = \lambda(l)$.

(iii) For $l \in [\underline{r}, \infty]$, $\inf_{p(\cdot|\theta)\in\mathcal{P}_l} I_p(\Theta, X) = 0$, and the infimum is achieved by any $p(x) \in \mathcal{P}_l$ which is independent of θ.

(iv) Assume $D(m^*_{\lambda_2}||m^*_{\lambda_1}) + D(m^*_{\lambda_1}||m^*_{\lambda_2}) < \infty$ for $0 < \lambda_1 \leq \lambda_2$, then $l(\lambda_2) \leq l(\lambda_1)$. i.e. $l(\cdot)$ is a decreasing function.

Under conditions of Theorem 3.3.1, we have the following:

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

(vi) $l(\lambda) \to \underline{r}$, 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

$$M^*_\lambda(\cdot) \xrightarrow{D} M_0(\cdot) \tag{3.3.14}$$

for some probability measure $M_0(\cdot)$ as $l \to \underline{r}$ (or $\lambda \to 0$), then

$$P^*_\lambda(\cdot|\theta) \xrightarrow{D} M_0(\cdot) \tag{3.3.15}$$

$$W^*_\lambda(\cdot|x) \xrightarrow{D} W(\cdot) \tag{3.3.16}$$

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

We comment that $l(\cdot)$ is usually continuous in examples, but we have not established a general result showing this.

Proof: (i) By Proposition 3.1, for given $l \in (0, \underline{r})$, $p^*_\lambda(\cdot|\theta)$ exists uniquely. From Theorem 6.3.2 of Blahut (1987), we know that for $l \in (0, \underline{r})$, the rate distortion function $\inf_{p(\cdot|\theta)\in\mathcal{P}_l} I_p(\Theta, X)$ is strictly positive and is a convex (hence continuous) decreasing function of l.

(ii) By (i), l is determined uniquely by λ.

On the other hand, for $l \in (0, \underline{r})$, we know that there is a unique $p^*_\lambda(\cdot|\theta)$. If there is another $\lambda' \neq \lambda$ such that $p^*_{\lambda'}(\cdot|\theta) = p^*_\lambda(\cdot|\theta)$, then by (i) of Theorem 3.3.1, $m^*_{\lambda'}(\cdot) = m^*_\lambda(\cdot)$,

so by (1.2.1.3) we have, $\forall x, \theta$

$$\frac{e^{-\lambda' L(x,\theta)}}{\int m_\lambda^*(y) e^{-\lambda L(y,\theta)} dy} = \frac{e^{-\lambda L(x,\theta)}}{\int m_\lambda^*(y) e^{-\lambda L(y,\theta)} dy},$$

or

$$e^{-(\lambda'-\lambda)L(x,\theta)} = \int m_\lambda^*(y) e^{-\lambda' L(y,\theta)} dy / \int m_\lambda^*(y) e^{-\lambda L(y,\theta)} dy,$$

which is impossible, since the left hand side of the above is a function of x and the right hand side is independent of x.

(iii) In the proof of Theorem 6.3.2 of Blahut (1987), it was shown that

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

By (i), it is decreasing in l.

Clearly, if $p(\cdot)$ is independent of θ, then $I_p(\Theta, X) = 0$.

(iv) Consider

$$p_\lambda^*(x|\theta) = \frac{m_\lambda^*(x) e^{-\lambda L(x,\theta)}}{\int m_\lambda^*(y) e^{-\lambda L(y,\theta)} dy}$$

for two values λ_1 and λ_2. For fixed θ, we have

$$D(p_{\lambda_2}^* || p_{\lambda_1}^*)(\theta) = \int p_{\lambda_2}^*(x|\theta) \log \frac{p_{\lambda_2}^*(x|\theta)}{p_{\lambda_1}^*(x|\theta)} dx$$

$$= \int p_{\lambda_2}^*(x|\theta) \log \left(\frac{\int m_{\lambda_1}^*(y) e^{-\lambda_1 L(y,\theta)} dy}{\int m_{\lambda_2}^*(y) e^{-\lambda_2 L(y,\theta)} dy} e^{(\lambda_1-\lambda_2)L(x,\theta)} \frac{m_{\lambda_2}^*(x)}{m_{\lambda_1}^*(x)} \right) dx$$

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

$$+ \int p_{\lambda_2}^*(x|\theta) \log \frac{m_{\lambda_2}^*(x)}{m_{\lambda_1}^*(x)} dx. \tag{3.3.17}$$

Similarly,

$$D(p_{\lambda_1}^* || p_{\lambda_2}^*)(\theta) = \log \left(\frac{\int m_{\lambda_2}^*(y) e^{-\lambda_2 L(y,\theta)} dy}{\int m_{\lambda_1}^*(y) e^{-\lambda_1 L(y,\theta)} dy} \right)$$

$$+(\lambda_2 - \lambda_1) \int L(x,\theta) p_{\lambda_1}^*(x|\theta) dx + \int p_{\lambda_1}^*(x|\theta) \log \frac{m_{\lambda_1}^*(x)}{m_{\lambda_2}^*(x)} dx. \tag{3.3.18}$$

Adding (3.3.17) and (3.3.18) gives

$$D(p_{\lambda_2}^* || p_{\lambda_1}^*)(\theta) + D(p_{\lambda_1}^* || p_{\lambda_2}^*)(\theta)$$

$$= (\lambda_1 - \lambda_2) \left(\int L(x,\theta) p_{\lambda_2}^*(x|\theta) dx - \int L(x,\theta) p_{\lambda_1}^*(x|\theta) dx \right)$$

$$+\int p^*_{\lambda_2}(x|\theta)\log\frac{m^*_{\lambda_2}(x)}{m^*_{\lambda_1}(x)}dx+\int p^*_{\lambda_1}(x|\theta)\log\frac{m^*_{\lambda_1}(x)}{m^*_{\lambda_2}(x)}dx. \tag{3.3.19}$$

Averaging over θ in (3.3.19), we get

$$\begin{aligned}0 &\le E_w D(p^*_{\lambda_2}||p^*_{\lambda_1})(\theta)+E_w D(p^*_{\lambda_1}||p^*_{\lambda_2})(\theta)\\ &=\int (D(p^*_{\lambda_2}||p^*_{\lambda_1})(\theta)+D(p^*_{\lambda_1}||p^*_{\lambda_2})(\theta))w(\theta)d\theta\\ &=(\lambda_1-\lambda_2)(l(\lambda_2)-l(\lambda_1))+D(m^*_{\lambda_2}||m^*_{\lambda_1})+D(m^*_{\lambda_1}||m^*_{\lambda_2}).\end{aligned} \tag{3.3.20}$$

By the same technique as in the proof of Proposition 3.1, we can prove that $D(p_2||p_1)$ is convex in p_2, p_1:

$$D(\alpha p'_2+(1-\alpha)p''_2||\alpha p'_1+(1-\alpha)p''_1)\le \alpha D(p'_2||p'_1)+(1-\alpha)D(p''_2||p''_1), \tag{3.3.21}$$

with equality holds if and only if $p'_2(\cdot)=p'_1(\cdot)$ and $p''_2(\cdot)=p''_1(\cdot)$. So, we have

$$E_w D(p^*_{\lambda_2}||p^*_{\lambda_1})\ge D(E_w p^*_{\lambda_2}||E_w p^*_{\lambda_1}),\quad E_w D(p^*_{\lambda_1}||p^*_{\lambda_2})\ge D(E_w p^*_{\lambda_1}||E_w p^*_{\lambda_2}). \tag{3.3.22}$$

With equality holds iff $p^*_{\lambda_2}(\cdot\theta)=p^*_{\lambda_1}(\cdot|\theta)$ $a.s.\theta$. Thus, from (3.3.20) and (3.3.22) we have $l(\lambda_2)\le l(\lambda_1)$, for $\lambda_1\le\lambda_2$.

(v) For all λ we have

$$l(\lambda)=\int\int p^*_\lambda(x|\theta)L(x,\theta)w(\theta)dxd\theta\ge 0.$$

Taking the limit superior gives

$$\begin{aligned}\limsup_{\lambda\to\infty} l(\lambda) &\le \int w(\theta)\limsup_{\lambda\to\infty}\int p^*_\lambda(x|\theta)L(x,\theta)dxd\theta\\ &=\int_{S^0\cap C} w(\theta)\limsup_{\lambda\to\infty}\int p^*_\lambda(x|\theta)L(x,\theta)dxd\theta.\end{aligned}$$

So, it is enough to show

$$\limsup_{\lambda\to\infty}\int p^*_\lambda(x|\theta)L(x,\theta)dx=0,\quad \forall\theta\in S^0\cap C. \tag{3.3.23}$$

Indeed, by continuity we have that $\forall\epsilon>0,\exists\delta>0,\ni L(\delta)<\epsilon$. This gives

$$\begin{aligned}\int p^*_\lambda(x|\theta)L(x,\theta)dx &=\int_{|x-\theta|\le\delta} p^*_\lambda(x|\theta)L(x,\theta)dx+\int_{|x-\theta|>\delta} p^*_\lambda(x|\theta)L(x,\theta)dx\\ &\le \epsilon+\frac{\int_{[x-\theta,x+\delta]^c} m^*_\lambda(x)e^{-\lambda L(x,\theta)}L(x,\theta)dx}{\int m^*_\lambda(y)e^{-\lambda L(y,\theta)}dy}.\end{aligned} \tag{3.3.24}$$

Since $\exists\ T$, for $t > T, e^{-t}t < e^{-t/2}$, let λ be so large that $\lambda \underline{r}(\delta) > T$. Now for such $\lambda(> 1)$,

$$e^{-\lambda L(x,\theta)} L(x,\theta) \leq e^{-\lambda L(x,\theta)} \lambda L(x,\theta) \leq e^{-\frac{\lambda}{2} L(x,\theta)}, \quad \forall x \in [x-\delta, x+\delta]^c.$$

Let $\delta' > 0$, satisfy $\underline{r}(\delta') < \underline{r}(\delta)/2$, and $[\theta - \delta', \theta + \delta'] \subset S^0$. For such δ', the second term on the right hand side of (3.3.24) is bounded by

$$\frac{\int_{[x-\delta,x+\delta]^c} m_\lambda^*(x) e^{-\frac{\lambda}{2} L(x,\theta)} dx}{\int_{[\theta-\delta',\theta+\delta']} m_\lambda^*(y) e^{-\lambda L(y,\theta)} dy} \leq \frac{e^{-\frac{\lambda}{2}\underline{r}(\delta)}}{e^{-\lambda \underline{r}(\delta')}} \frac{\int_{[x-\theta,x+\delta]^c} m_\lambda^*(x) dx}{\int_{[\theta-\delta',\theta+\delta']} m_\lambda^*(y) dy}$$

$$\leq e^{-\lambda(\frac{\underline{r}(\delta)}{2} - \underline{r}(\delta'))} \frac{1}{\int_{[\theta-\delta',\theta+\delta']} m_\lambda^*(y) dy}.$$

Let $b = \inf_{y \in [\theta-\delta',\theta+\delta']} w(y)$, and $B_\lambda = \{y \in [\theta - \delta', \theta + \delta'] \mid m_\lambda^*(y) \geq \frac{b}{2}\}$. Now by (ii) of Theorem 3.3.1, we know that for large λ, we have $\mu(B_\lambda) > \delta'$. Now, the last upper bound is bounded by

$$e^{-\lambda(\frac{\underline{r}(\delta)}{2} - \underline{r}(\delta'))} \frac{1}{\frac{b}{2}\mu(B_\lambda)} \leq \frac{2}{b\delta'} e^{-\lambda(\frac{\underline{r}(\delta)}{2} - \underline{r}(\delta'))} \to 0, \quad \text{as} \quad \lambda \to \infty,$$

establishing (v).

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

(vii) For (3.3.15), it is enough to prove that for all compact $A \in \mathcal{B}$, the Borel algebra on R^1, as $\lambda \to 0$

$$P_\lambda^*(A|\theta) \to M_0(A).$$

Indeed, since

$$p_\lambda^*(x|\theta) = \frac{m_\lambda^*(x) e^{-\lambda L(x,\theta)}}{\int m_\lambda^*(y) e^{-\lambda L(y,\theta)} dy},$$

we have

$$P_\lambda^*(A|\theta) = \int_A \frac{e^{-\lambda L(x,\theta)}}{\int e^{-\lambda L(y,\theta)} M_\lambda^*(dy)} M_\lambda^*(dx). \tag{3.3.25}$$

For any pre-assigned $\epsilon > 0$, we can choose a, b, λ_0 such that $-\infty < a < b < \infty, 0 < \lambda_0 < \infty$ and

$$\int_{[a,b]} e^{-\lambda_0 L(y,\theta)} M_0(dy) \geq 1 - \frac{\epsilon}{2}.$$

Since $e^{-\lambda L(x,\theta)} \leq 1$, we have that, for $\lambda \leq \lambda_0$,

$$P_\lambda^*(A|\theta) \leq \frac{M_\lambda^*(A)}{\int_{[a,b]} e^{-\lambda_0 L(y,\theta)} M_\lambda^*(dy)}. \tag{3.3.26}$$

Note that a, b, λ_0 are independent of λ. Since $e^{-\lambda_0 L(y,\theta)}$ is bounded and continuous on $[a,b]$, and $M^*_\lambda(\cdot) \xrightarrow{D} M_0(\cdot)$, we have

$$\int_{[a,b]} e^{-\lambda_0 L(y,\theta)} M^*_\lambda(dy) \to \int_{[a,b]} e^{-\lambda_0 L(y,\theta)} M_0(dy),$$

as $\lambda \to 0$. That is, for λ small enough,

$$\int_{[a,b]} e^{-\lambda_0 L(y,\theta)} M^*_\lambda(dy) \geq 1 - \epsilon,$$

and $M^*_\lambda(A) \leq M_0(A) + \epsilon$. Now (3.3.26) is bounded from above by

$$\frac{M_0(A) + \epsilon}{1 - \epsilon}.$$

Since ϵ was arbitrary, we have

$$\limsup_{\lambda \to 0} P^*_\lambda(A|\theta) \leq M_0(A).$$

On the other hand, since the denominator in (3.3.25) is no greater than 1, we have

$$P^*_\lambda(A|\theta) \geq \int_A e^{-\lambda L(x,\theta)} M^*_\lambda(dx)$$

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

where a, b and λ_0 are chosen so that

$$\int_{A \cap [a,b]} e^{-\lambda_0 L(x,\theta)} M_0(dx) \geq M_0(A) - \epsilon.$$

So, for λ small, we have

$$P^*_\lambda(A|\theta) \geq M_0(A) - \epsilon.$$

Thus

$$\liminf_{\lambda \to 0} P^*_\lambda(A|\theta) \geq M_0(A),$$

establishing (3.3.14).

For (3.3.16), note

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

So, for small λ

$$W^*_\lambda(A|x) = \int_A \frac{e^{-\lambda L(x,\theta)} W(d\theta)}{\int e^{-\lambda L(y,\theta)} M^*_\lambda(dy)}$$

$$\leq \int_A \frac{W(d\theta)}{\int_{[a,b]} e^{-\lambda_0 L(y,\theta)} M_\lambda^*(dy)}. \tag{3.3.27}$$

As before, choose finite numbers a, b so large and λ_0 so small that uniformly for $\theta \in A$ (recall that A is compact)

$$\int_{[a,b]} e^{-\lambda_0 L(y,\theta)} M_\lambda^*(dy) \geq 1 - \epsilon.$$

Now, (3.3.27) is bounded above by $W(A)/(1-\epsilon)$.

On the other hand, for small λ

$$W_\lambda^*(A|x) \geq \int_A e^{-\lambda L(x,\theta)} W(d\theta)$$

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

As before, choose a, b and $\lambda_0 > 0$ small enough that

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

Since $\epsilon > 0$ was arbitrary, we get (2.4.16).

(viii) Now we prove $M_0(\cdot) = \zeta(x_0)$. By way of contradiction, suppose $M_0(\cdot) \neq \zeta(x_0)$. By the constraint (1.2.1) we have

$$l(\lambda) = \int\int L(x,\theta) P_\lambda^*(dx|\theta) W(d\theta)$$

$$\geq \int_{[a,b]} \int_{[c,d]} L(x,\theta) P_\lambda^*(dx|\theta) W(d\theta). \tag{3.3.28}$$

Since $M_0(\cdot)$ does not concentrate at $x_0 = \arg\inf_x \int L(x,\theta) W(d\theta)$, there is $\epsilon_2 = \epsilon_2(\epsilon_1) > 0$ such that

$$\int\int L(x,\theta) M_0(dx) W(d\theta) > \underline{r} + \epsilon, \tag{3.3.29}$$

for some $\epsilon > 0$. Strictness of the inequality follows from $M_0(\cdot) \neq \zeta(x_0)$, since the inequality implies $M_0(\cdot)$ assigns positive mass away from x_0. We can choose a, b, c and d large and independent of λ so that

$$\int_{[a,b]} \int_{[c,d]} L(x,\theta) M_0(dx) W(d\theta) \geq \int\int L(x,\theta) M_0(dx) W(d\theta) - \frac{\epsilon}{2}. \tag{3.3.30}$$

Now, using (3.3.28), (3.3.19), (3.3.30) and (3.3.14), and the fact that $L(x,\theta)$ is bounded and continuous in x on $[a,b]$, and the fact that $l(\lambda) \to \underline{r}$ as $\lambda \to 0$, we get

$$\lim_{\lambda\to 0} l(\lambda) \geq \lim_{\lambda\to 0} \int_{[a,b]} \int_{[c,d]} L(x,\theta) P_\lambda^*(dx|\theta) W(d\theta)$$

$$= \int_{[a,b]} \int_{[c,d]} L(x,\theta) M_0(dx) W(d\theta) > \underline{r} + \frac{\epsilon}{2},$$

which is impossible. Thus, (3.3.16) follows. □

3.4 Hypothesis testing Using MILs

In this subsection, we demonstrate a parallel result for the exponential rate of the type II error when the product form of the MIL are used for testing the simple versus simple hypothesis.

Specifically, assume the independence model for the MIL, ie, for $x^n = (x_1, ..., x_n)$, $p^*_\lambda(x^n|\theta) = \prod_{i=1}^n p^*_\lambda(x_i|\theta)$. Now for fixed θ, the family now is parameterized by λ, we may interested in testing the hypotheses $H_1 : \lambda = \lambda_1$ vs $H_2 : \lambda = \lambda_2$. For simplicity, we denote $p^*_{\lambda_1}(\cdot|\theta)$ and $p^*_{\lambda_2}(\cdot|\theta)$ just by $p_1(\cdot)$ and $p_2(\cdot)$ respectively. If we use the Neyman-Pearson level α test with acceptance region A_n based on an *iid* sample $X_1, ..., X_n$. Let β_n denote the type 2 error, ie. $\beta_n = P_2(A_n)$, then we identify the exponential rate of β_n as in the following:

Theorem 3.4.1.

$$\lim_{n\to\infty} \frac{1}{n} \log \beta_n = -\frac{1}{2} \frac{[l(\lambda_2) - l(\lambda_1)]^2}{\mathrm{Var}_{p_2}[L(X,\theta)]} \tag{3.4.1}$$

Proof: We first prove

$$\lim_{n\to\infty} \frac{1}{n} \log \beta_n = -\frac{1}{2} \frac{[D(p_2||p_1) + D(p_1||p_2)]^2}{\int p_2 (\log \frac{p_2}{p_1})^2 - D^2(p_2||p_1)}. \tag{3.4.2}$$

Since A_n has the form

$$A_n = \Big\{ X^n : \frac{p_2(X^n)}{p_1(X^n)} \geq c_n \Big\},$$

where c_n is determined by $P_1(A_n) = 1 - \alpha$. So we have

$$1 - \alpha = P_1\Big(\frac{1}{\sqrt{n}} \sum_{i=1}^n \log \frac{p_1(X_i)}{p_2(X_i)} \geq \frac{1}{\sqrt{n}} \log c_n \Big)$$

$$= P_1\Big(\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\log\frac{p_1(X_i)}{p_2(X_i)} - D(p_1||p_2)]}{\sqrt{\int p_1(\log\frac{p_1}{p_2})^2 - D^2(p_1||p_2)}} \geq \frac{\frac{1}{\sqrt{n}}\log c_n - \sqrt{n}D(p_1||p_2)}{\sqrt{\int p_1(\log\frac{p_1}{p_2})^2 - D^2(p_1||p_2)}}\Big)$$

$$\sim 1 - \Phi\Big(\frac{\frac{1}{\sqrt{n}}\log c_n - \sqrt{n}D(p_1||p_2)}{\sqrt{\int p_1(\log\frac{p_1}{p_2})^2 - D^2(p_1||p_2)}}\Big),$$

$$\Rightarrow \frac{1}{\sqrt{n}}\log c_n \sim \sqrt{\int p_2(\log\frac{p_2}{p_1})^2 - D^2(p_2||p_1)}\,\Phi^{-1}(\alpha) + \sqrt{n}D^2(p_1||p_2).$$

And

$$\beta_n = P_2(A_n) = P_2\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\log\frac{p_2(X_i)}{p_1(X_i)} \leq -\frac{1}{\sqrt{n}}\log c_n\Big)$$

$$= P_2\Big(\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\log\frac{p_2(X_i)}{p_1(X_i)} - D(p_2||p_1)]}{\sqrt{\int p_2(\log\frac{p_2}{p_1})^2 - D^2(p_2||p_1)}} \leq -\frac{\frac{1}{\sqrt{n}}\log c_n + \sqrt{n}D(p_2||p_1)}{\sqrt{\int p_2(\log\frac{p_2}{p_1})^2 - D^2(p_2||p_1)}}\Big)$$

$$\sim \Phi\Big(-\frac{\frac{1}{\sqrt{n}}\log c_n + \sqrt{n}D(p_2||p_1)}{\sqrt{\int p_2(\log\frac{p_2}{p_1})^2 - D^2(p_2||p_1)}}\Big)$$

$$\sim \Phi\Big(-\frac{\Phi^{-1}(\alpha)\sqrt{\int p_1(\log\frac{p_1}{p_2})^2 - D^2(p_1||p_2)} + \sqrt{n}D(p_1||p_2) + \sqrt{n}D(p_2||p_1)}{\sqrt{\int p_2(\log\frac{p_2}{p_1})^2 - D^2(p_2||p_1)}}\Big)$$

$$\equiv \Phi(a_n),$$

where $a_n \to -\infty$, as $n \to \infty$, so by using the L'Hospital rule we get

$$\lim_{n\to\infty}\frac{1}{n}\log\beta_n = \lim_{n\to\infty}\frac{1}{n}\log\Phi(a_n) = \lim_{n\to\infty}\frac{1}{n}\log(\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{a_n} e^{-\frac{x^2}{2}}dx)$$

$$= \lim_{n\to\infty}\frac{e^{-\frac{a_n^2}{2}}(-\frac{1}{2\sqrt{n}}\frac{D(p_1||p_2)+D(p_2||p_1)}{\sqrt{\int p_2(\log\frac{p_2}{p_1})^2 - D^2(p_2||p_1)}})}{\int_{-\infty}^{a_n} e^{-\frac{x^2}{2}}dx}$$

$$= \lim_{n\to\infty}[\frac{e^{-\frac{a_n^2}{2}}(-\frac{1}{2\sqrt{n}}\frac{D(p_1||p_2)+D(p_2||p_1)}{\sqrt{\int p_2(\log\frac{p_2}{p_1})^2 - D^2(p_2||p_1)}})^2(-a_n)}{e^{-\frac{a_n^2}{2}}(-\frac{1}{2\sqrt{n}}\frac{D(p_1||p_2)+D(p_2||p_1)}{\sqrt{\int p_2(\log\frac{p_2}{p_1})^2 - D^2(p_2||p_1)}})}$$

$$+\frac{e^{-\frac{a_n^2}{2}}(-\frac{1}{4\sqrt{n^3}}\frac{D(p_1||p_2)+D(p_2||p_1)}{\sqrt{\int p_2(\log\frac{p_2}{p_1})^2 - D^2(p_2||p_1)}})}{e^{-\frac{a_n^2}{2}}(-\frac{1}{2\sqrt{n}}\frac{D(p_1||p_2)+D(p_2||p_1)}{\sqrt{\int p_2(\log\frac{p_2}{p_1})^2 - D^2(p_2||p_1)}})}]$$

$$= -\frac{1}{2}\frac{(D(p_1||p_2)+D(p_2||p_1))^2}{\int p_2(\log\frac{p_2}{p_1})^2 - D^2(p_2||p_1)},$$

ie. (3.4.2) is true.

Now since

$$D(p^*_{\lambda_2}||p^*_{\lambda_1})(\theta)+D(p^*_{\lambda_1}||p^*_{\lambda_2})(\theta)$$
$$=(\lambda_1-\lambda_2)(\int L(x,\theta)p^*_{\lambda_2}(x|\theta)dx-\int L(x,\theta)p^*_{\lambda_1}(x|\theta)dx),$$

and note θ has a degenerate distribution $w(\cdot)$, so for $i=1,2$

$$\int L(x,\theta)p_i(x|\theta)dx=\int\int L(x,\theta)p_i(x|\theta)w(\theta)dxd\theta=l(\lambda_i),$$

and hence

$$D(p_2||p_1)+D(p_1||p_2)=(\lambda_2-\lambda_1)(l(\lambda_1)-l(\lambda_2)). \tag{3.4.3}$$

$$\int p_2(x|\theta)(\log\frac{p_2(x|\theta)}{p_1(x|\theta)})^2dx=\int p_2(x|\theta)\Big(\log\Big(\frac{\int m^*_2(y)e^{-\lambda_2L(y,\theta)}dy}{\int m^*_1(y)e^{-\lambda_1L(y,\theta)}dy}e^{(\lambda_1-\lambda_2)L(x,\theta)}\Big)\Big)^2dx$$
$$=\Big(\log\Big(\frac{\int m^*_2(y)e^{-\lambda_2L(y,\theta)}dy}{\int m^*_1(y)e^{-\lambda_1L(y,\theta)}dy}\Big)\Big)^2+2(\lambda_1-\lambda_2)l(\lambda_2)\log\Big(\frac{\int m^*_2(y)e^{-\lambda_2L(y,\theta)}dy}{\int m^*_1(y)e^{-\lambda_1L(y,\theta)}dy}\Big)$$
$$+(\lambda_1-\lambda_2)^2\int L(x,\theta)^2p_2(x|\theta)dx. \tag{3.4.4}$$

Since

$$D(p^*_{\lambda_1}||p^*_{\lambda_2})(\theta)=\log\Big(\frac{\int m^*_{\lambda_2}(y)e^{-\lambda_2L(y,\theta)}dy}{\int m^*_{\lambda_1}(y)e^{-\lambda_1L(y,\theta)}dy}\Big)+(\lambda_2-\lambda_1)\int L(x,\theta)p^*_{\lambda_1}(x|\theta)dx. \tag{3.4.5}$$

we have

$$D^2(p_2||p_1)=\Big(\log\Big(\frac{\int m^*_2(y)e^{-\lambda_2L(y,\theta)}dy}{\int m^*_1(y)e^{-\lambda_1L(y,\theta)}dy}\Big)\Big)^2+2(\lambda_1-\lambda_2)l(\lambda_2)\log\Big(\frac{\int m^*_2(y)e^{-\lambda_2L(y,\theta)}dy}{\int m^*_1(y)e^{-\lambda_1L(y,\theta)}dy}\Big)$$
$$+(\lambda_1-\lambda_2)^2l^2(\lambda_2). \tag{3.4.6}$$

Now by (3.4.3), (3.4.4) and (3.4.6), the RHS of (3.4.2) is

$$-\frac{1}{2}\frac{[l(\lambda_2-l(\lambda_1]^2}{\mathrm{Var}_{p_2}(L(X,\theta))},$$

thus complete the proof. □

3.5 Remarks

The result of Theorem 3.3.1 is in striking contrast to the fact that for *iid* data distributed according to a density $p(x|\theta)$ where θ is a d-dimensional parameter, we have

$$E_m D(w(\cdot|X^n)||w(\cdot)) = \frac{d}{2}\ln n + O(1).$$

Asymptotically maximizing this latter expression over priors w as is done to find a reference prior, see Bernardo (1979), leads to Jeffreys prior, Clarke and Barron (1994).

Let $x_0 = \arg\inf_x \int w(\theta)L(x,\theta)d\theta$ as in the condition of Theorem 3.3.1. It is necessary for $w(\cdot)$ to have a finite second moment if x_0 is to be finite for $L(x,\theta) = (x-\theta)^2$. 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 $Exp(\alpha,\mu)$, 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 $x_0 = \mu + 1/\alpha$.

In general, $x_0 =$ mean (θ), if $L(x,\theta) = (x-\theta)^2$.

Chapter 4

Application

4.1 Introduction

Here we give a practical example in which MIL's can be used to provide answers to questions of interest that do not seem amenable to other techniques. In the present case, it appears that MIL's do better than a conventional analysis because they can be applied to summary statistics.

In general we suggest that MIL's may prove useful in settings satisfying the following criteria. First, a true parametric family cannot be proposed. That is, a general form for the relationship between outcomes and a parameter is not apparent. Second, the unknown parameter must be location like, that is, it is not necessarily a location parameter in the strict sense of the term but does nevertheless track the typical range for the outcomes. The interpretation for θ is pre-set, the prior must be formulated according to this. Here, we use a Bayes hypothesis test, Theorem 3.1.2 to suggest some point estimate problem from a frequentist perspective way also be feasible.

We assume that a finite dimensional parametrization for the parameter has been chosen and that it permits estimation of the parameter of interest, possibly as a function of the coordinates of the parameter. Let $\theta = (\theta_1, ..., \theta_k)$ denote the finite dimensional parametrization. In general, the data can be written as a random vector $X^n = (X_1, \ldots, X_n)$ with outcomes denoted $x^n = (x_1, \ldots, x_n)$. Assume that it is possible to associate to the distribution of X^n a parametric family that has a density that can be written as $p(x^n|\theta)$ for $\theta \in R^k$ but that the form of $p(\cdot|\theta)$ is unknown. In particular, suppose that there is no basis for any assumptions about $p(\cdot|\theta)$, i.e., we know nothing about how values of the parameter affect the probabilities of the outcomes, only that the two are related in some way.

Even in such cases where little is known, an experimenter may have some idea about

what values of θ are more surprising than others. Assume these preconceptions have been formulated into a prior density $w(\theta)$.

Now, from either a Bayesian or frequentist perspective, the key quantity remaining to be identified is the likelihood. Assuming there is no practical basis for choosing $p(\cdot|\theta)$, we adopt a minimal information criterion. That is, we seek a likelihood which requires relatively relaxed assumptions in a precise sense. Although the data compression interpretation is valid, we present our method from a data transmission viewpoint since this is easier to describe compactly. The likelihood can be used in initial data analysis where one cannot make detailed modeling assumptions and one must rely chiefly on arguments from robustness: If one has robustness against modeling strategy in the sense that the same results obtain for several different modeling strategies (none of which make strong assumptions i.e., are minimally informative), and the results are insensitive to the choice of prior, loss function and bound on the Bayes risk then one has more confidence in the validity of the conclusions.

4.2 Application to A Real Data Set

Here we demonstrate the use of MIL's by re-analyzing data from Nader and Reboussin (1994). To model the data, we used the MILs in two different ways. That is, the models we use make weak assumptions in an information theoretic sense. However, for large sample sizes, Theorem 3.3.1 tells us that the n-dimensional MIL may unsuitable for practice in some situations, especially for large data sets. Other models and assumptions may be more appropriate for this data set, we are not aiming to search through for the best model for this data, just simply want a demonstration of possible uses with the MIL approach. The particular model by the MILs may also not appropriate, however, as a study of our method for an initial use to a real data set, we made our best effort in the modeling and data assumptions, and welcome any criticism for our future improvement.

In the formulation of the MILs, we used the Bayes risk bound l, but in most of our examples and applications, we used the parameter λ to determine the MILs. The two parameters are equivalent as described in (ii) of Theorem 3.3.2. Use of λ is direct and more convenient in inferences.

4.2.1 Description of the Data

The experimental data studied in Nader and Reboussin (1994) was collected to investigate whether two different training methods will produce different effects on the behavior of the monkeys. In this experiment, eight monkeys were initially trained to respond under a fixed interval 5-minute schedule of intravenous cocaine presentation, FI5. In this training, the first time that a monkey pulled a lever after having waited at least five minutes produced a cocaine injection. The injection lasted ten seconds. (Responses during the injection were not counted.)

Prior to the second phase of training, the monkeys were rated from one to eight based on response rates under the FI5 schedule. Based on this rating, the monkeys were paired so that two monkeys who formed a pair would have similar average FI5 response rates before two different cocaine self-administration reinforcement schedules were applied. The two highest ratings gave the first block; the third and fourth highest gave the next and so on. Within each pair, members were randomly assigned to one of two schedules. Thus, four monkeys were trained under an FR50 ("fixed response 50") 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 ("inter-response time 30 seconds") schedule; the monkeys were reinforced by a cocaine dose for lever presses at least thirty seconds apart. A lever press before 30 seconds elapsed reset the IRT30 timer. For all monkeys, each cocaine injection was followed by a two minute timeout and a sixty minute timeout followed the tenth and twentieth cocaine injections.

Following the 65th session under FR50 or IRT30, availability of cocaine was again scheduled under FI5 for sixty consecutive sessions. For each of these sixty sessions three variables were measured. The primary variable of interest was the response rate which was the total number of responses during the session divided by the session length in minutes. Here, session length was the actual session length less the timeouts. Secondary variables were cocaine intake (in mg/kg per session) and average quarter life. Quarter life values are the proportion of the fixed interval elapsed when 25% of the responses in that interval had occurred. The average is taken over the five minute intervals that occurred during a session. The intake measures how quickly the monkey made a response after the five minute

intervals in a session. Figure 3 shows plots of the response rate data over the sixty sessions for the eight monkeys. The left 4 plots are the data from the 4 monkeys of the FR50 group, the right 4 plots are the data from the 4 monkeys of the IRT30 group. The monkeys are paired row wise.

First, label the monkeys in pairs as (1,2), (3,4), (5,6) and (7,8), (here the odd numbers and even numbers correspond to, respectively, the FR50 and IRT30, or the left column and the right column of plots in Figure 3) where the odd labels mean that the first in each pair was trained under FR50 and the even labels mean that the second in each pair was trained under IRT30. From Figure 3, we see a few data points in the early sessions of monkeys I and III that are obviously larger than the rest of the corresponding observations. We treat these as outliers and delete the first three observations from all the eight monkeys in our data analysis. For simplicity of notation, we just relabel the remaining observations for each money as 1 to 57. Let y_{ij} be the datum on rate for the i-th monkey on the j-th day where $i = 1, ..., 8$ and $j = 1, ..., 57$. For each i, let $\bar{y}_i = \frac{1}{57}\sum_{i=1}^{57} y_{ij}$ be the sample mean of the rate data from the i-th monkey. Now, take differences of the means within each pair. We write these differences 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 $x_4 = \bar{y}_7 - \bar{y}_8$. From the paired data we can obtain a few simple descriptive statistics. For the first pair, (1,2) we can find the mean, variance and lag-1 auto correlation for the vector $(y_{1,1} - y_{1,2}, ..., y_{1,60} - y_{2,60})$ of sessional differences. Doing this for the other 3 pairs, and for the vector of sessional differences with the first 3 entries deleted gives the summary statistics in Table 1. Note that following the rule of thumb which gives $2/\sqrt{n} \approx 0.26$ as a threshold for assessing the presence of serial correlation (see Farnum and Stanton, 1989, P.78) leads us to suspect that most of the vector of differences does exhibit dependence. In addition, we see that all of the means are positive, consistent with IRT30 and FR50 having different effects. The range of the sample variances is too large to permit meaningful assertions. Deleting the outliers does not appear to affect the summary statistics uniformly, apart from reducing the variances.

Table 1. Data Summary for Observed Differences in the Two Groups

	Full Data			1st 3 Obs. Deleted		
Diff.	Mean	Var.	Auto Corr.	Mean	Var.	Auto Corr.
1-2	0.80	16.31	0.25	0.59	1.03	0.31
3-4	4.10	10.46	0.35	3.65	5.25	0.22
5-6	9.42	19.15	0.43	9.27	19.69	0.44
7-8	1.79	3.59	0.44	2.27	2.33	0.40

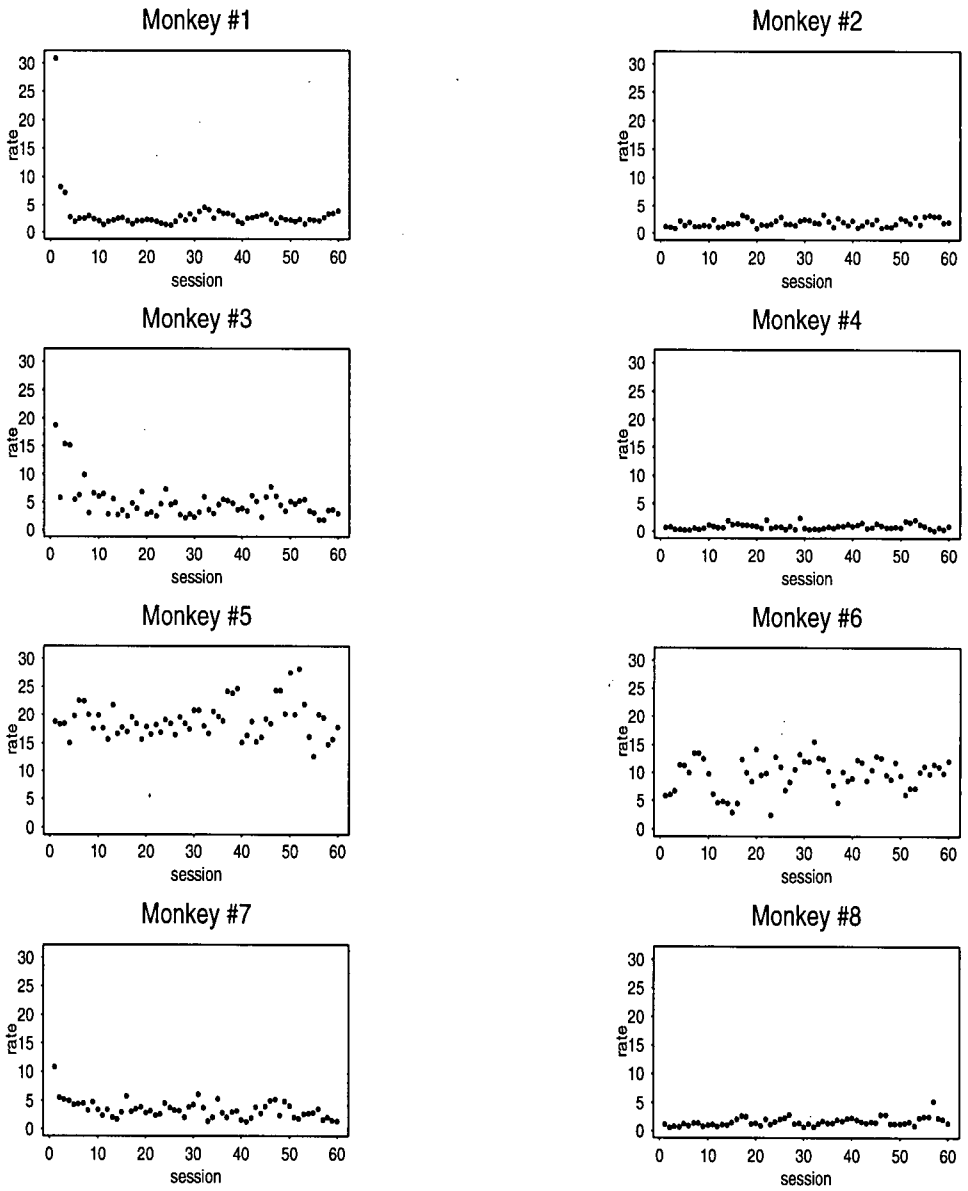

Figure 3: Plots of Rate Over the Sixty Sessions. This sequence of figures shows a plot of rate over the sessions for each of the eight monkeys. Each row corresponds to a pair from matching baseline lever pressing rates; each column corresponds to a treatment either FR-50 (left) or IRT-30 (right).

Here, the data vector for monkey I is $(y_{1,1}, ..., y_{1,60})$ with mean, variance and lag-1 auto correlation defined as before for the vector of differences. The corresponding summary statistics are displayed in Table 2. Note that, again, deleting of the outliers does not affect the summary statistics uniformly, and the rule of thumb gives that each Monkey's data vector exhibits dependence, although for Monkey III with the 3 outliers deleted one may be skeptical.

Table 2. Data Summary for the Eight Monkeys

	Full Data			1st 3 Obs. Deleted		
Monkey	Mean	Var.	Auto Corr.	Mean	Var.	Auto Corr.
1	2.65	14.76	0.23	2.47	0.53	0.31
2	1.85	0.46	0.27	1.88	0.44	0.24
3	5.00	9.73	0.33	4.57	4.76	0.18
4	0.90	0.26	0.24	0.91	0.24	0.22
5	19.03	9.54	0.31	19.06	10.03	0.31
6	9.61	8.89	0.45	9.78	8.71	0.48
7	3.39	2.49	0.34	3.19	1.45	0.31
8	1.59	0.57	0.47	1.63	0.57	0.43

Remark: The summary statistics x^4 and y^4 do not reflect the role of the sample size n. To address this criticism, one could use the standardized summary statistics. Here, for example one might use $\sqrt{n}x^4$ in place of x^4 and the same for y^4. If for the pair of monkeys, the numbers of observations are different, say n_1 and n_2 respectively for the two monkeys, we may use the weighted standard summary statistics, for example use $\sqrt{n_1}x^4$ in place of x^4 and similarly for y^4.

There are other ways to address this criticism also. In view of Theorem 3.1.1, we do not want to lump together too many data points. So, we might replace a string $x_1, ..., x_n$ by a sequence of summary statistics, the number of summary statistics to be taken as independent being chosen by the experimenter to reflect the number of independent data points his data is equivalent to.

One may also interested in how sample size affects the width of the HPD region based on the MIL model. At the two extremes, the MILs can be used to form a product of i.i.d. densities or to get a single n-fold dependent density. In the former case, the result is the same as for i.i.d distributions. As sample size increases, the accuracy of inference increases, so the width of the HPD region decreases roughly as $\sqrt{n}$ times root inverse of the Fisher

information. For the latter case, since the n-fold MIL has high dependency, in general, among its arguments, the sample size effect on the width of the HPD regions is not as evident as in the former case, since dependence causes large data sets to look like smaller ones.

4.2.2 Models for the Data and Results

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) looked for 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 is 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. The other variables intake and quarter life were analyzed separately and gave conclusions compatible with the rate analysis.

The main virtue of this modeling approach is its simplicity. However, 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 fit two models and consider a third. We model the data differences for the paired monkeys in the two groups and look for training differences in the mean response rates. We find a significant difference in the mean rates. We only analyze the rate data since it is regarded as the most important index. Our methods can be applied to either the average quarter life or the cocaine intake data as well.

Model I

The data on rate for the eight monkeys over the 60 sessions are plotted in Figure 3 . Neither the plots in these plots, nor the form of the experiment suggests an obvious parametric family. Moreover, there is not enough data to perform a nonparametric analysis.

A key problem is that for each monkey one cannot assume the data from the sixty sessions are independent, even conditionally on the monkey. In addition, it is not clear how the differences from pair to pair can be modeled. As a consequence of the absence of strong modeling assumptions, it is unreasonable to use any standard likelihood to model the data. However, the MIL method gives a likelihood which makes relatively weak assumptions on the data distribution. We use it to extract initial conclusions.

We begin our analysis by using the MILs. There are other modeling strategies that are feasible with MILs but they are more elaborate and require much more programming. In the first modeling strategy, we take differences within each pair in the data and average over the fifty seven sessions. (Recall we deleted the first three observations as outliers from each monkey. Also, one may consider different models, for example, use the CLT for a normal approximation of the average. However, if the number of observations is small, the CLT and other models may not be practical to use. Here our only intent is to demonstrate the MIL approach.) Thus, we estimate a single parameter using, effectively, four data points which are the mean differences. In the second model, we again average over the 57 data points for each monkey but do not take differences in the data. Note that the eight monkeys are assumed to be independent of each other, although for each fixed monkey, the 57 data points may not be independent. Nevertheless, the eight mean data points are independent, though they may not be identically distributed. Instead, we use two parameters, one for the IRT30 group and one for the FR50 group and then obtain a posterior for the difference in the parameters.

For Model I, we suppose that the expected values of the x_i's are the same and treat this as the parameter of interest θ, with the same units as the observations. Now, the problem reduces to finding a posterior for this parameter given the data. If the posterior assigns most of its mass around a positive value we infer that the expectation of x_i is strictly positive and therefore the IRT30 rate is lower than the FR50 rate.

For a given prior $w(\theta)$, a given loss function L, and a given value of λ, we can get an MIL $p^*(x|\theta)$ by the procedure in Chapter 1. For simplicity, we treat the four pair differences $x_1, ..., x_4$ as approximately independent and identically distributed. Although dependence may present between blocks, we may assume they are canceled in the difference, leaving

only the effects of the training. Now, we can form the posterior density

$$w^*(\theta|x_1,x_2,x_3,x_4) = \frac{p^*(x_1|\theta)p^*(x_2|\theta)p^*(x_3|\theta)p^*(x_4|\theta)w(\theta)}{\int p^*(x_1|\xi)p^*(x_2|\xi)p^*(x_3|\xi)p^*(x_4|\xi)w(\xi)d\xi}. \qquad (4.2.2.1)$$

Given $\alpha > 0$, one can see whether the $(1-\alpha)$ highest posterior density (H.P.D.) region for θ contains 0.

We obtained graphs of the posterior in (4.2.2.1) for a range of values of λ, several choices of prior and two choices of L. In practice, the priors are chosen according to pre-experimental knowledge about the parameter distribution. Here we choose a few of them for convenience. In particular, we chose $w(\cdot)$ to be $U(-15,20)$ or any of $N(-2,1)$, $N(0,1)$, $N(2,1)$ and $N(0,10)$ to reflect a variety of priors with a range of reasonable means and variances; we choose $L(x,\theta) = (x-\theta)^2$, or $L(x,\theta) = |x-\theta|$; and we choose λ to range from .5 to 5. We recall that in the context of $p^*(x|\theta)$, λ has two meanings. First, it behaves like a scale or dispersion parameter. For larger λ's, $p^*(x|\theta)$ is more flat, for smaller λ, $p^*(x|\theta)$ is more concentrated or more sharply peaked. The value of λ also affects the set of likelihoods over which we have optimized. Larger λ values correspond to a smaller set, and vice versa. We require λ to be in $[0,\infty)$, otherwise $p^*(x|\theta)$ is independent of θ and so is meaningless. In our work we generally found that values of λ in [.1, 10] gave reasonable results.

We used the iteration method described in section 1.3 to find the MIL $p^*(x|\theta)$. That is, we chose an initial distribution $m_{(0)}(\cdot)$, and got $p^*_{(1)}(x|\theta)$ by (1.3.1). Then, we plugged $p^*_{(1)}(x|\theta)$ into (1.3.2) to get $m^*_{(1)}(\cdot)$. We continued this cycling to get the n step likelihood $p^*_{(n)}(x|\theta)$ until the absolute difference of the two consecutive approximations to the MIL were no greater than a prespecified $\epsilon > 0$. Here, we found that a reasonable choice for ϵ ranged from 10^{-6} to 10^{-4}, and that the number of iterations required in our calculations for a fixed x and θ was of the order 10 to 10^2.

Our results for the cases listed above were generally consistent. For all the above choices of the prior distribution, λ and the loss function, the posterior density was unimodal and concentrated on the positive half line, with mode between 4 and 8.5, and relatively small posterior variance, which can be controlled by the parameter λ. 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. Figure 4.a shows some of the posteriors we obtained. Note that the posteriors assign essentially all their mass to the positive half-line.

Also, for Model I we tried some priors with larger variance (from the data plot, we find variance 10 is reasonable) and we deleted the first three observations from each monkey, since it seems that there are some outliers in these observations. The results are plotted in Figure 4.b. We see similar skewness toward the right axe on θ.

Note that Theorems 3.3.1 and 3.3.2 do not apply directly to (4.2.2.1) because it uses a product of univariate MIL's rather than a single MIL for a 4-variate outcome. However, the conclusions of those two theorems are qualitatively consistent with the results here. The posterior in (4.2.2.1) is seen to concentrate at a point as λ increases which is consistent with (iii) of Theorem 3.3.1. In addition, as λ decreases, the posterior is seen to converge to a dispersed distribution that is similar to the prior used, as suggested by (iii) of Theorem 3.3.1.

Model II

We investigate an alternative model based on MIL's to show how one might examine robustness against modeling strategy for paired data. As an alternative model, instead of using four differences in the data to estimate one parameter, we considered using the sample mean for the four first entry pairs and the sample mean from the four second entry pairs to estimate two parameters reflecting the means of the two groups of monkeys. Again, we assume independence between the two groups to simplify the modeling. This may seems not reasonable for the practical data set. To model the exact dependence structure in the data seems difficult, here we only intend to do another initial analysis using the MIL in a different way and compare the conclusion. Then, we can marginalize the posterior to get credible regions for the difference in the two parameters.

Since there are two parameters, we use a two-dimensional prior which for the present we assume factors, that is, we assume $w(\cdot,\cdot) = w_1(\cdot) \cdot w_2(\cdot)$. Also, we assume the components of X are independent and the components of Y are independent.

From $w_1(\cdot)$, we get the MIL $p_1^*(x^4|\theta_1) = \prod_{i=1}^4 p_1^*(x_i|\theta_1)$, and from $w_2(\cdot)$ we get the MIL $p_2^*(y^4|\theta_2) = \prod_{i=1}^4 p_2^*(y_i|\theta_2)$. Now, we can form the two-dimensional posterior

$$w(\theta_1,\theta_2|x^4,y^4) = \frac{p_1^*(x^4|\theta_1)p_2^*(y^4|\theta_2)}{m^*(x^4,y^4)} \qquad (4.2.2.2)$$

$$= w_1(\theta_1|x^4)w_2(\theta_2|y^4),$$

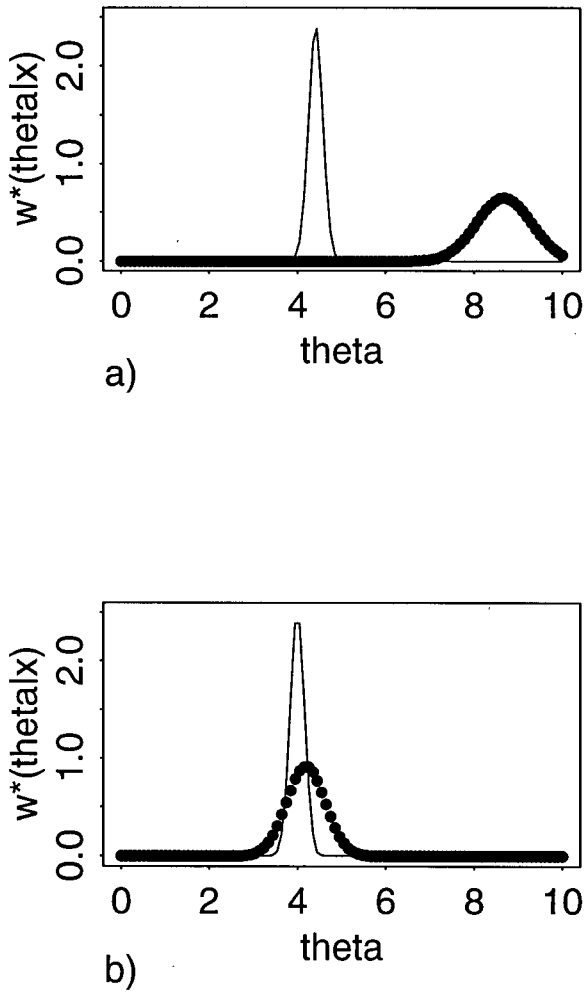

Figure 4: Posteriors from Model I. In 4 (a), the posteriors plotted here were formed from MIL's based on choosing w to be $N(0,1)$, L to be squared error loss, and $\lambda = .7$ (points) or $\lambda = 1.5$ (solid). In 4 (b) The posteriors plotted here were formed from MIL's based on choosing w to be $N(0,10)$, L to be squared error loss, and $\lambda = .7$ (dots) or $\lambda = 1.5$ (solid).

where $m^*(x^4, y^4)$ is the marginal from $p(x^4, y^4|\theta_1, \theta_2) = p_1^*(x^4|\theta_1)p_2^*(y^4|\theta_2)$ and the prior $w(\theta_1, \theta_2) = w_1(\theta_1)w_2(\theta_2)$, and $w_1(\theta_1|x^4)$ and $w_2(\theta_2|y^4)$ are the corresponding one-dimensional marginals.

Now, we can apply the transformation

$$\psi = \theta_1 + \theta_2, \quad \phi = \theta_1 - \theta_2$$

in the bivariate posterior. After integrating out ψ, we get a posterior for ϕ. (In our C program, we used discretization summation to approximate the integration.)

For this model, we also tried several priors (with variances ranging from about 1 to 10), losses (squared error and absolute difference) and λs (around 0.0001 to around 0.05). We found that when λ is small, around .0001, and the prior is $N(0, 1)$, the posterior is nearly $N(0, 1)$. In view of Theorem 3.3.1 this is not a surprise. As λ increases, the posterior shifts so as to concentrate on positive values of θ. However, when λ is much above .09 or much below .0001 our implementation of the Blahut-Arimoto algorithm is numerically unstable because the integrand function is close to a product of delta functions. This problem did not occur with Model I because the posterior there is based on a product of four densities whereas in Model II the product has 8 densities. The problem seems to be that as λ increases, the w_1 and w_2 concentrate at different points so that the product is too small for the computer to store. One consequence of this is that we cannot observe the convergence of the posterior to unit mass at a point that is suggested by Theorem 3.3.1. Moreover, in addition to having used a product of MIL's, we have marginalized (4.2.2.2) making the conclusion of Theorem 3.3.1 more distant.

Despite being unable to observe the concentration of the model at a point with increasing λ, Figure 5 shows the posteriors we obtained for two values of λ, .0001 and .09. Intermediate values of λ give posteriors roughly between these two posteriors. Note that for $\lambda = .0001$ the posterior reverts to a dispersed distribution resembling the prior. As the common value of the λ's increases, w^* shifts away from being centered at zero and again assigns essentially all its mass to the positive half-line (see Figure 5.a). The point is to note that if λ is chosen in Model II to be as close as possible to the values we used for Model I (without exceeding the limit of .09 so we can still compute) the inferences we make from the two models are qualitatively the same, namely we have evidence that the difference in rates for the FR50 and IRT30 groups is positive. Thus the two modeling strategies confirm each other. We

note that the conclusions from Model II do not seem as strong as from Model I: we attribute this to the choice of λ here being much smaller than the choice of λ in Model I, that is, the Bayes risk bound l used here is larger than that used in Model I. So, the set $\mathcal{P}_l$ here is much larger. Currently we do not have a good formal technique for choosing λ (or l). We will discuss this later in Chapter 6.

We also used $N(0, 10)$ as the prior for Model I and the priors for θ_1 and θ_2 in Model II, and did the same analysis. In this case, the posteriors form the two models are more spread out, see Figure 5 (b), especially for small values of λ_1 and λ_2, since the corresponding Bayes risk bound is large which makes the allowable distortion large and hence less accurate inferences. However, for the moderate value of 0.05 for λ_1 and λ_2, much of the posterior mass lies on the right of zero, this leads to similar conclusion as the $N(0, 1)$ prior was used.

More Alternatives

Having recourse to MIL's permits the elaboration of other models that do not require the extreme data summarization used in models I and II. This summarization is used here only to make it easier to get computational results, and is justifiable chiefly on the basis that it is not far wrong. One of the ways in which models I and II can be criticized is that they, unlike Nader and Reboussin's original model, are insensitive to the sample size used to form the summary statistics.

Other models can be considered. For example, assuming independence between the 57 sessions for each monkey, we can model each of the 57 observations form the two groups by the same MIL, and take the product of the 57 MIL's for the whole data set; or assuming no independence, we generate a 57 dimensional MIL to model the whole data. In this later case, we get a dependence model. Intermediate dependence structures can also be used. Unfortunately, there remains the problem of how to get the right dependence structure from the data. We will discuss this question partially in Section 6.2.1.

It is this plethora of modeling strategies that are equally plausible which motivated the work in the next Chapter.

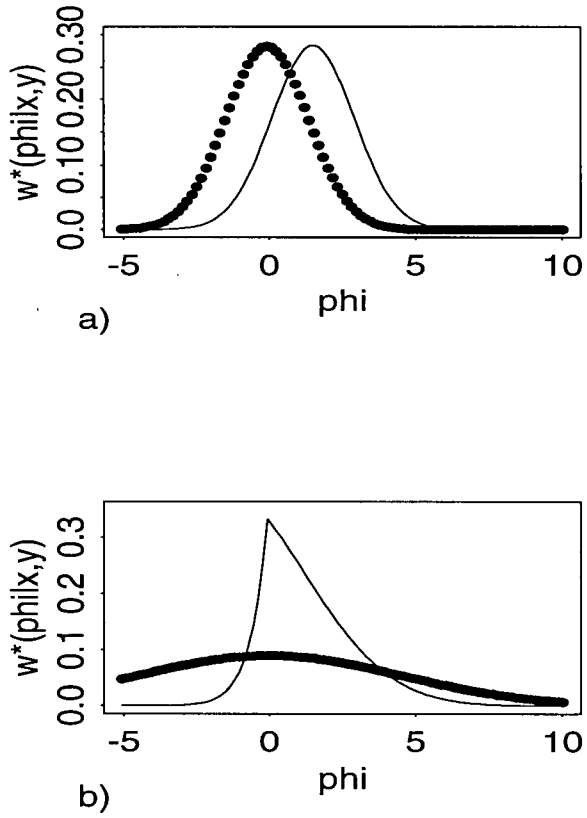

Figure 5: Posteriors from Model II. In 5 (a) the posteriors plotted here were formed from MIL's based on choosing $w_1 = w_2$ to be $N(0,1)$, L to be squared error loss, and $\lambda_1 = \lambda_2 =$ .0001 (dots), or $\lambda_1 = \lambda_2 =$ .05 (solid). In 5 (b) the posteriors plotted here were formed from MIL's based on choosing $w_1 = w_2$ to be $N(0,10)$, L to be squared error loss, and $\lambda_1 = \lambda_2 =$ .0001 (bold), or $\lambda_1 = \lambda_2 =$ .05 (solid).

Chapter 5

Robustness of Modeling Strategies for Paired Data

Motivated by the inferential similarity of two different modeling strategies we have tried to investigate formally the degree to which three modeling strategies applicable in the problems of the previous chapters would agree in general. We begin by defining general cases of the two models we have used and defining a general case for a third model that is equally plausible but that we did not use. Then, we seek conditions under which these three models will be equivalent, and we present results which partially characterize how discrepant the inferences from these models will be. In this Chapter, some of our results are for n-fold likelihoods, some are for products of univariate likelihoods and some are for other special cases. We indicate applicability of each result.

5.1 Introduction and Definition of Models

Recall that in the example in the previous chapter, we assume that we have two independent data sets X^n and Y^n, and we are interested in modeling the data with various likelihoods so as to make inference about the parameter θ. The parameter θ is a quantification of some population trait of interest.

We have used several different models. In Model I from Section 4.2.2, we generate the MIL for $Z^n = X^n - Y^n$ directly to get $p_{(1)}(z^n|\theta) = p^*(z^n|\theta)$ and get the corresponding posterior $w_{(1)}(\theta|z^n)$. In Model II, by contrast, we model the two sets of data X^n and Y^n by $p_1^*(x^n|\theta_1)$ and $p_2^*(y^n|\theta_2)$ which use the two marginal priors $w_1(\theta_1)$ and $w_2(\theta_2)$ from a joint prior $w(\theta_1, \theta_2)$. We got the posterior $w(\theta_1, \theta_2|X^n, Y^n)$, and then applied the transformation $\theta = \theta_1 - \theta_2$, $\phi = \theta_1 + \theta_2$. Marginalizing out ϕ gives the posterior $w_{(2)}(\theta|X^n, Y^n)$ of θ. In Chapter 4, we used the product of uni-dimensional MILs to form the model. In this Chapter, we present a robustness analysis for paired data from general likelihoods; they

may be dependent or independent among their variables, often including the various MIL's as special cases. Also, some results are only for n-dimensional MILs. The point here is to compare different modeling strategies, assuming the same likelihoods been used in each of these models.

We can consider other models. In particular, we define a third model, Model III: just as in Model II, we use $p_1(x^n|\theta_1)$ and $p_2(y^n|\theta_2)$ for X^n and Y^n respectively, then use the transformation $Z^n = X^n - Y^n, S^n = X^n + Y^n$ to get the density for (Z^n, S^n). Then integrating out S^n gives the density $p_{(3)}(z^n|\theta_1, \theta_2)$ for Z^n. Using the transformation $\theta = \theta_1 - \theta_2$, $\phi = \theta_1 + \theta_2$ and marginalizing out ϕ gives the posterior $w_{(3)}(\theta|z^n)$ of θ. In some cases $p_{(3)}(z^n|\theta_1, \theta_2)$ will reduce to the form $p_{(3)}(z^n|\theta)$, where $\theta = \theta_1 - \theta_2$, without further transformation. Later in this chapter we will deal with these conditions. This model is different from Model II in general. It uses the transformation of parameters to get the likelihood in θ first and then gets the posterior. Whereas in Model II, we get the two dimensional posterior first, then use the transformation on parameters and marginalize to get the posterior for θ.

There are numerous reasonable models for consideration, here as an attempt to do some robustness analysis for the paired data, we only consider the above three commonly used models. In practice, we can consider more general transformations in the models:

$$z^n = f_1(x^n, y^n), \qquad s^n = f_2(x^n, y^n). \tag{5.1.1}$$

Also, we have used

$$\theta = g_1(\theta_1, \theta_2), \qquad \phi = g_2(\theta_1, \theta_2). \tag{5.1.2}$$

Note that in Model I, we model the data transformation, while in Model II, we model the parameter transformation. These two modeling strategies are widely used in practice, it is natural to investigate the robustness of these methods for paired data.

If there were only 2 ways, Model I and Model II, to analyze a data set, we could do both. If they agree, as we have shown in Chapter 4, then we could be content and stop. However, there are many alternative techniques. Consider the general form of Model III in which used both a transformation in the data and a transformation on the parameters. Model $X_1, ..., X_n$ as iid $p(x|\theta_1)$ and $Y_1, ...Y_n$ as iid $p(y|\theta_2)$. Here we assumed the X_i's and

the Y_j's are from the same parametric family only with different parameters. Assume X^n and Y^n are independent. Now, we can use $Z^n = f(X^n, Y^n)$ so as to derive, by convolution, a density for Z^n. The density for Z_i is, in general,

$$p(z_i|\theta_1, \theta_2) = \int p_1(g_1(s_i, z_i)|\theta_1) p_2(g_2(s_i, z_i)|\theta_2) J(s^n, z^n) ds_i, \tag{5.1.3}$$

where $x_i = g_1(s_i, z_i)$, $y_i = g_2(s_i, z_i)$ is the inverse transformation, and $J(s^n, z^n) = \prod_{i=1}^n J(s_i, z_i)$ is the transformation Jacobian. For compatibility of the data transformation, let $\theta = f_1(\theta_1, \theta_2)$ be the parameter of interest. We can use the MIL for $p_1(\cdot|\theta_1)$ and for $p_2(\cdot|\theta_2)$ so as to obtain a minimally informative convolution in (5.1.3).

In some cases, which we identify presently, the left hand side of (5.1.3) reduces to a density of the form $p(z^n|\theta)$, where $\theta = f(\theta_1, \theta_2)$. More generally, however, this reduction does not occur. Whether or not the reduction occurs, the parameter of interest is $\theta = f_1(\theta_1, \theta_2)$: we would therefore use a prior for θ (or θ and ϕ) and make inferences as before.

If we want to relax the assumption of independence between X_i and Y_i, (5.1.3) changes. We would use

$$\int p(g_1(s, z), g_2(s, z)|\theta_1, \theta_2) J(s, z) ds$$

in place of (5.1.3) and have to find a MIL for a bivariate random variable, with two parameters.

For the present, we note some further alternatives. One could use a location family, one could use extra data; one could use a likelihood that was not minimally informative — perhaps based on physical modeling. One could avoid the extreme data summarization used here by modeling the day-to-day dependence through a time series approach.

The class of all models is enormous. Even after restriction to the subclass of all statistically plausible models, there remain too many to enumerate and evaluate in every particular instance. Moreover, there is no guarantee that all models in this subclass will give the same inferences for the parameter of interest. It is worthwhile therefore to have some theoretical guidelines for when to expect two modeling strategies to agree and for when to expect them to disagree.

In short, the task of this chapter is to begin an investigation into the robustness of inferences to change in modeling strategy for paired data. Many have investigated sensitivity to

prior selection. Sensitivity to small changes in the likelihood has also been studied, although not much. Sensitivity to outliers or, more generally, data has been extensive . However, in all cases, the modeling strategy (by which we mean transformation of data, transformation of parameters and the nature of the link between the likelihood and the data, including the loss function if there is one) has never been the focus of a robustness study. Here we undertake to begin this in several cases.

5.2 Equivalence of Models

For simplicity, we consider the models for the specific form as in the beginning of this chapter. Thus, the results in this subsection are general; they are true for any n-dimensional likelihoods, independent or not, any form of MIL or not, including all the models with the product of 1-dimensional MILs we considered in Chapter 4.

Let the prior for Model I to be

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

and choose the likelihood for Model I to be

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

for some density $\pi(\cdot)$. We see that if the joint likelihood $p(s,z|\phi,\theta) = \frac{1}{2}p_1(\frac{s+z}{2}|\frac{\phi+\theta}{2})p_2(\frac{s-z}{2}|\frac{\phi-\theta}{2})$ for (S,Z) satisfies a sufficiency-like condition between Z and θ, then the three models are equivalent. The sufficiency-like condition is that the joint density of Z and S, obtained from the joint density of X and Y by transformation, can be factored into two parts. One part is a function of Z and θ only, the other part is independent of θ. Specifically, we have the following

Proposition 5.2.1 Suppose the joint likelihood for (S,Z) satisfies

$$p(s,z|\phi,\theta) = g(z,\theta)h(z,s,\phi)$$

for some functions $g(\cdot,\cdot)$ and $h(\cdot,\cdot,\cdot)$, and the prior satisfies

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

for some $w_1(\cdot)$ and $w_2(\cdot)$, then

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

Proof: Since the likelihood for Model I is

$$p_{(1)}(z|\theta) = \frac{1}{2}g(z,\theta)\int\int h(z,s,\phi)\pi(\phi)dsd\phi,$$

the posterior density for Model I is

$$\begin{aligned} w_{(1)}(\theta|Z) &= \frac{g(Z,\theta)\int\int h(Z,s,\phi)\pi(\phi)dsd\phi w_1(\theta)}{\int g(Z,\xi)\int\int h(Z,s,\phi)\pi(\phi)dsd\phi w_1(\xi)d\xi} \\ &\propto g(Z,\theta)w_1(\theta). \end{aligned} \tag{5.2.1}$$

Similarly, the likelihood for Model II factors as

$$p_{(2)}(x,y|\theta_1,\theta_2) = g(z,\theta)h(z,s,\phi),$$

giving that the posterior density for Model II is

$$\begin{aligned} w_{(2)}(\theta|X,Y) &= \frac{\frac{1}{2}g(Z,\theta)\int h(Z,S,\phi)w(\frac{\phi+\theta}{2},\frac{\phi-\theta}{2})d\phi}{\frac{1}{2}\int\int g(Z,\xi)h(Z,S,\phi)w(\frac{\phi+\xi}{2},\frac{\phi-\xi}{2})d\phi d\xi} \\ &\propto g(Z,\theta)w_1(\theta). \end{aligned} \tag{5.2.2}$$

The likelihood for Model III is

$$p_{(3)}(z|\theta_1,\theta_2) = g(z,\theta)\int h(z,s,\phi)ds,$$

thus

$$\begin{aligned} w_{(3)}(\theta|Z) &= \frac{g(Z,\theta)\int h(Z,s,\phi)ds\int w(\frac{\phi+\theta}{2},\frac{\phi-\theta}{2})d\phi}{\int g(Z,\xi)\int\int h(Z,s,\phi)w(\frac{\phi+\xi}{2},\frac{\phi-\xi}{2})dsd\phi d\xi} \\ &\propto g(Z,\theta)w_1(\theta). \end{aligned} \tag{5.2.3}$$

Now (5.2.1), (5.2.2) and (5.2.3) together complete the proof. □

Remark. We comment that if $w_1(\cdot) = w_2(\cdot)$ is the standard normal then $p_1(\cdot|\theta_1)$ is $N(\theta_1,\sigma^2)$ and $p_2(\cdot|\theta_2)$ is $N(\theta_2,\sigma^2)$, and the condition in Proposition 5.2.1 is satisfied.

5.3 Robustness against Modelling Strategies for Paired Data

The previous bounds on the differences between Models I and II were bounds in an average sense useful for comparing whole models. For practical purposes bounds that are pointwise in the data are more useful: they permit comparison of inferences given a particular data set. We first consider the simple case of the transformation of data and parameters:

$$S^n = X^n + Y^n, \qquad Z^n = X^n - Y^n;$$

$$\phi = \theta_1 + \theta_2, \qquad \theta = \theta_1 - \theta_2.$$

If we use priors $w_2(\theta_1, \theta_2), w_3(\theta_1, \theta_2)$ in Models II and III respectively, we can get an upper bound on the L_1 distance between the two posteriors without averaging over the data. In Proposition 5.3.1 and Proposition 5.3.3, the likelihoods involved are general in the sense described in the beginning of this Chapter they include all the models we considered in Chapter 4. Corollary 5.3.1, Proposition 5.3.2 and Theorem 5.3.1 are only for the n-dimensional MILs.

Proposition 5.3.1. If the priors of models 2 and 3 are respectively $w_2(\theta_1, \theta_2)$, $w_3(\theta_1, \theta_2)$, then

(i) For any data x^n and y^n,

$$\int \left| w_{(2)}(\theta | x^n, y^n) - w_{(3)}(\theta | z^n) \right| d\theta \leq 2M_{(2,3)}(x^n, y^n),$$

where

$$M_{(2,3)}(x^n, y^n) = M_{(2,3)}(s^n, z^n) = \sup_{\phi,\theta} M_{(2,3)}(s^n, z^n, \phi, \theta),$$

$$M_{(2,3)}(s^n, z^n, \phi, \theta) = \min\{\left|1 - R_{(2,3)}(s^n, z^n, \phi, \theta)\right|, \left|1 - [R_{(2,3)}(s^n, z^n, \phi, \theta)]^{-1}\right|\}$$

and

$$R_{(2,3)}(s^n, z^n, \phi, \theta) = \frac{\int p_1(\frac{v^n + z^n}{2} | \frac{\phi+\theta}{2}) p_2(\frac{v^n - z^n}{2} | \frac{\phi-\theta}{2}) dv^n w_3(\frac{\phi+\theta}{2}, \frac{\phi-\theta}{2})}{p_1(\frac{s^n + z^n}{2} | \frac{\phi+\theta}{2}) p_2(\frac{s^n - z^n}{2} | \frac{\phi-\theta}{2}) w_2(\frac{\phi+\theta}{2}, \frac{\phi-\theta}{2})}.$$

If $w_2(\cdot, \cdot) = w_3(\cdot, \cdot) = w(\cdot, \cdot)$, then the factor $M_{(2,3)}(s^n, z^n)$ in the upper bound is independent of $w(\cdot, \cdot)$, so we have

$$\sup_{w\in W}\left|w_{(2)}(\theta|x^n,y^n)-w_{(3)}(\theta|z^n)\right|d\theta\le 2M_{(2,3)}(x^n,y^n),$$

where W is the collection of all the two-dimensional priors.

(ii) The posterior means under Model II and III satisfy

$$\left|E_{(2)}(\theta|x^n,y^n)-E_{(3)}(\theta|z^n)\right|\le(E_{(2)}(|\theta|\,|x^n,y^n)+E_{(3)}(|\theta|\,|z^n))M_{(2,3)}(x^n,y^n),$$

for any data set x^n,y^n.

(iii) The posterior variances under Models II and III satisfy

$$\left|\mathrm{Var}_{(2)}(\theta|x^n,y^n)-\mathrm{Var}_{(3)}(\theta|z^n)\right|\le h_{(2,3)}(x^n,y^n)M_{(2,3)}(x^n,y^n),$$

where

$$h_{(2,3)}(x^n,y^n)=h_{(2,3)}(s^n,z^n)=E_{(2)}(\theta^2\,|x^n,y^n)+E_{(3)}(\theta^2\,|z^n)$$
$$+2(E_{(2)}(|\theta|\,|x^n,y^n)+E_{(3)}(|\theta|\,|z^n))$$

Proof: (i) Let

$$g_2(s^n,z^n|\theta)=\int p_1(\frac{s^n+z^n}{2}|\frac{\phi+\theta}{2})p_2(\frac{s^n-z^n}{2}|\frac{\phi-\theta}{2})w_2(\frac{\phi+\theta}{2},\frac{\phi-\theta}{2})d\phi,$$
$$g_3(z^n|\theta)=\int\int p_1(\frac{s^n+z^n}{2}|\frac{\phi+\theta}{2})p_2(\frac{s^n-z^n}{2}|\frac{\phi-\theta}{2})w_3(\frac{\phi+\theta}{2},\frac{\phi-\theta}{2})d\phi ds^n.$$

Now,

$$\begin{aligned}w_{(2)}(\theta|x^n,y^n) &= \frac{\frac{1}{2}\int p_1(x^n|\frac{\phi+\theta}{2})p_2(y^n|\frac{\phi-\theta}{2})w_2(\frac{\phi+\theta}{2},\frac{\phi-\theta}{2})d\phi}{\int\int p_1(x^n|\theta_1)p_2(y^n|\theta_2)w_2(\theta_1,\theta_2)d\theta_1d\theta_2}\\ &= \frac{\int p_1(\frac{s^n+z^n}{2}|\frac{\phi+\theta}{2})p_2(\frac{s^n-z^n}{2}|\frac{\phi-\theta}{2})w_2(\frac{\phi+\theta}{2},\frac{\phi-\theta}{2})d\phi}{\int\int p_1(\frac{s^n+z^n}{2}|\frac{\xi_1+\xi_2}{2})p_2(\frac{s^n-z^n}{2}|\frac{\xi_1-\xi_2}{2})w_2(\frac{\xi_1+\xi_2}{2},\frac{\xi_1-\xi_2}{2})d\xi_1d\xi_2}\\ &= \frac{g_2(s^n,z^n|\theta)}{\int g_2(s^n,z^n|\xi)d\xi},\end{aligned}$$

and likewise, for Model II we have

$$\begin{aligned}w_{(3)}(\theta|z^n) &= \frac{\int\int p_1(\frac{s^n+z^n}{2}|\frac{\phi+\theta}{2})p_2(\frac{s^n-z^n}{2}|\frac{\phi-\theta}{2})w_3(\frac{\phi+\theta}{2},\frac{\phi-\theta}{2})d\phi ds^n}{\int\int\int p_1(\frac{s^n+z^n}{2}|\frac{\xi_1+\xi_2}{2})p_2(\frac{s^n-z^n}{2}|\frac{\xi_1-\xi_2}{2})w_3(\frac{\xi_1+\xi_2}{2},\frac{\xi_1-\xi_2}{2})d\xi_1d\xi_2ds^n}\\ &= \frac{g_3(z^n|\theta)}{\int g_3(z^n|\xi)d\xi}.\end{aligned}$$

Thus, the difference in the posteriors is

$$\int \left| w_{(2)}(\theta|x^n, y^n) - w_{(3)}(\theta|z^n) \right| d\theta = \int \left| \frac{g_2(s^n, z^n|\theta)}{\int g_2(s^n, z^n|\xi)d\xi} - \frac{g_3(z^n|\theta)}{\int g_3(z^n|\xi)d\xi} \right| d\theta$$

$$\leq \int \left| \frac{g_2(s^n, z^n|\theta)}{\int g_2(s^n, z^n|\xi)d\xi} - \frac{g_3(z^n|\theta)}{\int g_2(s^n, z^n|\xi)d\xi} \right| d\theta + \int \left| \frac{g_3(z^n|\theta)}{\int g_2(s^n, z^n|\xi)d\xi} - \frac{g_3(z^n|\theta)}{\int g_3(z^n|\xi)d\xi} \right| d\theta$$

$$= \frac{1}{\int g_2(s^n, z^n|\xi)d\xi} \int |g_2(s^n, z^n|\theta) - g_3(z^n|\theta)| \, d\theta$$

$$+ \frac{\int g_3(z^n|\theta)d\theta \, |\int g_2(s^n, z^n|\xi)d\xi - \int g_3(z^n|\xi)d\xi|}{\int g_2(s^n, z^n|\xi)d\xi \int g_3(z^n|\xi)d\xi}$$

$$= \frac{1}{\int g_2(s^n, z^n|\xi)d\xi} \left[\int |g_2(s^n, z^n|\theta) - g_3(z^n|\theta)| \, d\theta + \left| \int g_2(s^n, z^n|\xi)d\xi - \int g_3(z^n|\xi)d\xi \right| \right]$$

$$\leq \frac{2}{\int g_2(s^n, z^n|\xi)d\xi} \int |g_2(s^n, z^n|\theta) - g_3(z^n|\theta)| \, d\theta$$

$$= 2 \frac{\int \left| \int p_1(\frac{s^n+z^n}{2}|\frac{\phi+\theta}{2}) p_2(\frac{s^n-z^n}{2}|\frac{\phi-\theta}{2}) w_2(\frac{\phi+\theta}{2}, \frac{\phi-\theta}{2}) d\phi \right.}{\int\int p_1(\frac{s^n+z^n}{2}|\frac{\xi_1+\xi_2}{2}) p_2(\frac{s^n-z^n}{2}|\frac{\xi_1-\xi_2}{2}) w_2(\frac{\xi_1+\xi_2}{2}, \frac{\xi_1-\xi_2}{2}) d\xi_1 d\xi_2}$$

$$\left. - \int\int p_1(\frac{v^n+z^n}{2}|\frac{\phi+\theta}{2}) p_2(\frac{v^n-z^n}{2}|\frac{\phi-\theta}{2}) dv^n w_3(\frac{\phi+\theta}{2}, \frac{\phi-\theta}{2}) d\phi \right| d\theta$$

$$\leq 2 \frac{\int\int \left| p_1(\frac{s^n+z^n}{2}|\frac{\phi+\theta}{2}) p_2(\frac{s^n-z^n}{2}|\frac{\phi-\theta}{2}) w_2(\frac{\phi+\theta}{2}, \frac{\phi-\theta}{2}) \right.}{\int\int p_1(\frac{s^n+z^n}{2}|\frac{\xi_1+\xi_2}{2}) p_2(\frac{s^n-z^n}{2}|\frac{\xi_1-\xi_2}{2}) w_2(\frac{\xi_1+\xi_2}{2}, \frac{\xi_1-\xi_2}{2}) d\xi_1 d\xi_2}$$

$$\left. - \int p_1(\frac{v^n+z^n}{2}|\frac{\phi+\theta}{2}) p_2(\frac{v^n-z^n}{2}|\frac{\phi-\theta}{2}) dv^n w_3(\frac{\phi+\theta}{2}, \frac{\phi-\theta}{2}) \right| d\phi d\theta$$

$$= 2 \frac{\int\int \left| 1 - R_{(2,3)}(s^n, z^n, \phi, \theta) \right| p_1(\frac{s^n+z^n}{2}|\frac{\phi+\theta}{2}) p_2(\frac{s^n-z^n}{2}|\frac{\phi-\theta}{2}) w_2(\frac{\phi+\theta}{2}, \frac{\phi-\theta}{2}) d\phi d\theta}{\int\int p_1(\frac{s^n+z^n}{2}|\frac{\xi_1+\xi_2}{2}) p_2(\frac{s^n-z^n}{2}|\frac{\xi_1-\xi_2}{2}) w_2(\frac{\xi_1+\xi_2}{2}, \frac{\xi_1-\xi_2}{2}) d\xi_1 d\xi_2}$$

$$= 2E_1 \left| 1 - R_{(2,3)}(s^n, z^n, \phi, \theta) \right|, \tag{5.3.1}$$

where the expectation E_1 is taken over (ϕ, θ) with respect to the density

$$q_1(\phi, \theta) = \frac{p_1(\frac{s^n+z^n}{2}|\frac{\phi+\theta}{2}) p_2(\frac{s^n-z^n}{2}|\frac{\phi-\theta}{2}) w_2(\frac{\phi+\theta}{2}, \frac{\phi-\theta}{2})}{\int\int p_1(\frac{s^n+z^n}{2}|\frac{\xi_1+\xi_2}{2}) p_2(\frac{s^n-z^n}{2}|\frac{\xi_1-\xi_2}{2}) w_2(\frac{\xi_1+\xi_2}{2}, \frac{\xi_1-\xi_2}{2}) d\xi_1 d\xi_2}.$$

Similarly, by adding and subtracting $g_2(s^n, z^n|\theta)/\int g_3(z^n|\xi)d\xi$, we have

$$\int \left| w_{(2)}(\theta|x^n, y^n) - w_{(3)}(\theta|z^n) \right| d\theta$$

$$\leq \int \left| \frac{g_2(s^n, z^n|\theta)}{\int g_2(s^n, z^n|\xi)d\xi} - \frac{g_2(s^n, z^n|\theta)}{\int g_3(z^n|\xi)d\xi} \right| d\theta + \int \left| \frac{g_2(s^n, z^n|\theta)}{\int g_3(z^n|\xi)d\xi} - \frac{g_3(z^n|\theta)}{\int g_3(z^n|\xi)d\xi} \right| d\theta$$

$$= \frac{\int g_2(s^n, z^n|\theta)d\theta}{\int g_2(s^n, z^n|\xi)d\xi} \frac{|\int [g_2(s^n, z^n|\xi) - g_3(z^n|\xi)]d\xi|}{\int g_3(z^n|\xi)]d\xi}$$

$$+ \frac{1}{\int g_3(z^n|\xi)d\xi} \int |g_2(s^n, z^n|\theta) - g_3(z^n|\theta)|\, d\theta$$

$$\leq \frac{2}{\int g_3(z^n|\xi)d\xi} \int |g_2(s^n, z^n|\theta) - g_3(z^n|\theta)|\, d\theta$$

$$= 2\frac{\int \left| \int p_1(\frac{s^n+z^n}{2}|\frac{\phi+\theta}{2}) p_2(\frac{s^n-z^n}{2}|\frac{\phi-\theta}{2}) w_2(\frac{\phi+\theta}{2}, \frac{\phi-\theta}{2}) d\phi \right.}{\int\int\int p_1(\frac{s^n+z^n}{2}|\frac{\phi+\theta}{2}) p_2(\frac{s^n-z^n}{2}|\frac{\phi-\theta}{2}) w_3(\frac{\phi+\theta}{2}, \frac{\phi-\theta}{2}) ds^n d\phi d\theta}$$

$$\left. \frac{- \int\int p_1(\frac{s^n+z^n}{2}|\frac{\phi+\theta}{2}) p_2(\frac{s^n-z^n}{2}|\frac{\phi-\theta}{2}) ds^n w_3(\frac{\phi+\theta}{2}, \frac{\phi-\theta}{2}) d\phi \right| d\theta}{\int\int\int p_1(\frac{s^n+z^n}{2}|\frac{\phi+\theta}{2}) p_2(\frac{s^n-z^n}{2}|\frac{\phi-\theta}{2}) w_3(\frac{\phi+\theta}{2}, \frac{\phi-\theta}{2}) ds^n d\phi d\theta}$$

$$\leq 2\frac{\int\int \left| p_1(\frac{s^n+z^n}{2}|\frac{\phi+\theta}{2}) p_2(\frac{s^n-z^n}{2}|\frac{\phi-\theta}{2}) w_2(\frac{\phi+\theta}{2}, \frac{\phi-\theta}{2}) d\phi \right.}{\int\int\int p_1(\frac{s^n+z^n}{2}|\frac{\phi+\theta}{2}) p_2(\frac{s^n-z^n}{2}|\frac{\phi-\theta}{2}) w_3(\frac{\phi+\theta}{2}, \frac{\phi-\theta}{2}) ds^n d\phi d\theta}$$

$$\left. \frac{- \int p_1(\frac{s^n+z^n}{2}|\frac{\phi+\theta}{2}) p_2(\frac{s^n-z^n}{2}|\frac{\phi-\theta}{2}) ds^n w_3(\frac{\phi+\theta}{2}, \frac{\phi-\theta}{2}) \right| d\phi d\theta}{\int\int\int p_1(\frac{s^n+z^n}{2}|\frac{\phi+\theta}{2}) p_2(\frac{s^n-z^n}{2}|\frac{\phi-\theta}{2}) w_3(\frac{\phi+\theta}{2}, \frac{\phi-\theta}{2}) ds^n d\phi d\theta}$$

$$= 2 \int\int \left| [R_{(2,3)}(s^n, z^n, \phi, \theta)]^{-1} - 1 \right|$$

$$\frac{\int p_1(\frac{s^n+z^n}{2}|\frac{\phi+\theta}{2}) p_2(\frac{s^n-z^n}{2}|\frac{\phi-\theta}{2}) w_3(\frac{\phi+\theta}{2}, \frac{\phi-\theta}{2}) ds^n d\phi d\theta}{\int\int\int p_1(\frac{s^n+z^n}{2}|\frac{\phi+\theta}{2}) p_2(\frac{s^n-z^n}{2}|\frac{\phi-\theta}{2}) w_2(\frac{\phi+\theta}{2}, \frac{\phi-\theta}{2}) ds^n d\phi d\theta}$$

$$= 2E_2 \left| [R_{(2,3)}(s^n, z^n, \phi, \theta)]^{-1} - 1 \right|, \tag{5.3.2}$$

where the expectation E_2 is taken over (ϕ, θ) with respect to the density

$$q_2(\phi, \theta) = \frac{\int p_1(\frac{s^n+z^n}{2}|\frac{\phi+\theta}{2}) p_2(\frac{s^n-z^n}{2}|\frac{\phi-\theta}{2}) w_3(\frac{\phi+\theta}{2}, \frac{\phi-\theta}{2}) ds^n}{\int\int\int p_1(\frac{s^n+z^n}{2}|\frac{\phi+\theta}{2}) p_2(\frac{s^n-z^n}{2}|\frac{\phi-\theta}{2}) w_3(\frac{\phi+\theta}{2}, \frac{\phi-\theta}{2}) ds^n d\phi d\theta}.$$

Now by (5.3.1) and (5.3.2) we get the desired conclusion.

(ii) Similarly,

$$\left| \int \theta w_2(\theta|x^n, y^n)d\theta - \int \theta w_3(\theta|z^n)d\theta \right| = \left| \int \Big(\frac{\theta g_2(s^n, z^n|\theta)}{\int g_2(s^n, z^n|\xi)d\xi} - \frac{\theta g_3(z^n|\theta)}{\int g_3(z^n|\xi)d\xi} \Big) d\theta \right|$$

$$\leq \left| \int \Big(\frac{\theta g_2(s^n, z^n|\theta)}{\int g_2(s^n, z^n|\xi)d\xi} - \frac{\theta g_3(z^n|\theta)}{\int g_2(s^n, z^n|\xi)d\xi} \Big) d\theta \right|$$

$$+ \left| \int \Big(\frac{\theta g_3(z^n|\theta)}{\int g_2(s^n, z^n|\xi)d\xi} - \frac{\theta g_3(z^n|\theta)}{\int g_3(z^n|\xi)d\xi} \Big) d\theta \right|$$

$$= \frac{\int |\xi| g_2(s^n, z^n|\xi) d\xi}{\int g_2(s^n, z^n|\xi) d\xi} \frac{|\int (\theta g_2(s^n, z^n|\theta) - \theta g_3(z^n|\theta)) d\theta|}{\int |\xi| g_2(s^n, z^n|\xi) d\xi}$$
$$+ \frac{\int |\xi| g_3(z^n|\xi) d\xi}{\int g_3(z^n|\xi) d\xi} \frac{|\int g_2(s^n, z^n|\xi) d\xi - \int g_3(z^n|\xi) d\xi|}{\int g_2(s^n, z^n|\xi) d\xi}$$
$$\leq \Big(E_{(2)}(|\theta| \, |x^n, y^n) + E_{(3)}(|\theta| \, |z^n) \Big) M_{(2,3)}(x^n, y^n),$$

since, as in the proof of (i) we have

$$\frac{|\int (\theta g_2(s^n, z^n|\theta) - \theta g_3(z^n|\theta)) d\theta|}{\int |\xi| g_2(s^n, z^n|\xi) d\xi} \leq M_{(2,3)}(x^n, y^n),$$
$$\frac{|\int g_2(s^n, z^n|\xi) d\xi - \int g_3(z^n|\xi) d\xi|}{\int g_2(s^n, z^n|\xi) d\xi} \leq M_{(2,3)}(x^n, y^n),$$

and

$$\frac{\int |\xi| g_2(s^n, z^n|\xi) d\xi}{\int g_2(s^n, z^n|\xi) d\xi} = E_{(2)}(|\theta| \, |x^n, y^n), \qquad \frac{\int |\xi| g_3(z^n|\xi) d\xi}{\int g_3(z^n|\xi) d\xi} = E_{(3)}(|\theta| \, |z^n).$$

(iii) We have

$$\Big| \mathrm{Var}_{(2)}(\theta|x^n, y^n) - \mathrm{Var}_{(3)}(\theta|z^n) \Big| \leq \Big| \int (\theta^2 w_{(2)}(\theta|x^n, y^n) - \theta^2 w_{(3)}(\theta|z^n)) d\theta \Big|$$
$$+ \left| \left(\int \theta w_{(2)}(\theta|x^n, y^n) d\theta \right)^2 - \left(\int \theta w_{(3)}(\theta|z^n) d\theta \right)^2 \right|,$$

Note the second term in the right hand side of the above is

$$\left| \int \Big(\theta w_{(2)}(\theta|x^n, y^n) - \theta w_{(3)}(\theta|z^n) \Big) d\theta \right| \int \Big(\theta w_{(2)}(\theta|x^n, y^n) + \theta w_{(3)}(\theta|z^n) \Big) d\theta,$$

so as in the proof of (ii), the above is bounded by

$$\Big(\frac{m_{\tilde{w}}(s^n, z^n)}{m_w(s^n, z^n)} + \frac{m_{\tilde{w}}(z^n)}{m_w(z^n)} \Big) M_{(2,3)}(s^n, z^n) + 2 \Big(\frac{m_{\bar{w}}(s^n, z^n)}{m_w(s^n, z^n)} + \frac{m_{\bar{w}}(z^n)}{m_w(z^n)} \Big) M_{(2,3)}(s^n, z^n)$$
$$= \Big(E_{(2)}(\theta^2 \, |x^n, y^n) + E_{(3)}(\theta^2 \, |z^n) + 2(E_{(2)}(|\theta| \, |x^n, y^n) + E_{(3)}(|\theta| \, |z^n)) \Big)$$
$$\times M_{(2,3)}(s^n, z^n),$$

where we have defined

$$\frac{m_{\tilde{w}}(s^n, z^n)}{m_w(s^n, z^n)} = E_{(2)}(\theta^2 \, |x^n, y^n), \qquad \frac{m_{\tilde{w}}(z^n)}{m_w(z^n)} = E_{(3)}(|\theta| \, |z^n),$$

in which

$$m_{\tilde{w}}(s^n, z^n) = \int\int p_1(\frac{s^n + z^n}{2}|\frac{\phi+\theta}{2}) p_2(\frac{s^n - z^n}{2}|\frac{\phi-\theta}{2})$$
$$\times\theta^2 w_2(\frac{\phi+\theta}{2}, \frac{\phi-\theta}{2}) d\phi d\theta,$$
$$m_{\tilde{w}}(z^n) = \int\int\int p_1(\frac{s^n + z^n}{2}|\frac{\phi+\theta}{2}) p_2(\frac{s^n - z^n}{2}|\frac{\phi-\theta}{2})$$
$$\times\theta^2 w_3(\frac{\phi+\theta}{2}, \frac{\phi-\theta}{2}) d\phi d\theta ds^n. \quad \square$$

Remarks: 1. Since $\left|1 - R_{(2,3)}(s^n, z^n, \phi, \theta)\right| \geq 1$ implies $\left|1 - [R_{(2,3)}(s^n, z^n, \phi, \theta)]^{-1}\right| < 1$, so $\forall x^n, y^n$, $M_{(2,3)}(x^n, y^n) \leq 1$.

2. Note that in the above Proposition, if $w_2(\cdot,\cdot) = w_3(\cdot,\cdot)$, then $R_{(2,3)}(s^n, z^n, \phi, \theta) = p(z^n|s^n)^{-1}$, the conditional distribution of Z^n given S^n. So, if Z^n is tightly distributed given S^n, then $R_{(2,3)}(s^n, z^n, \phi, \theta) \approx 1$, and so $M_{(2,3)}(x^n, y^n) \approx 0$.

Also, for given priors $w_2(\cdot,\cdot), w_3(\cdot,\cdot)$, we can choose the data set (x^n, y^n) such that $R_{(2,3)}(s^n, z^n, \phi, \theta) \approx 1$. For given data set (x^n, y^n), we can choose priors $w_2(\cdot,\cdot), w_3(\cdot,\cdot)$ such that $R_{(2,3)}(s^n, z^n, \phi, \theta) \approx 1$. Or in other words, we can identify data for which the models are indistinguishable for given priors and we can find priors which make the models indistinguishable for a data set.

The upper bound in Proposition 5.3.1 is not sharp. If we take $w_1(\cdot,\cdot) = w_2(\cdot,\cdot)$ and $p_1(\cdot|\theta) = p_2(\cdot|\theta) = N(\theta, 1)$, then by Proposition 5.2.1 we have $w_{(2)}(\theta|x^n, y^n) = w_{(3)}(\theta|z^n)$. However, $R(s^n, z^n, \theta, \phi) = \prod_{i=1}^n \sqrt{4\pi}\exp\{(s_i-\phi)^2/4\}$, so $M_{(2,3)}(s^n, z^n, \theta, \phi) = 1-\prod_{i=1}^n(\sqrt{4\pi}$ $\exp\{(s_i - \phi)^2/4\})^{-1} > 0$, and as n tends to infinity, $M_{(2,3)}(s^n, z^n, \theta, \phi)$ tends to 1 giving a trivial result. On the other hand, this reduction to a trivial limiting case makes sense because posteriors concentrate at the true value if the prior assigns mass on a neighborhood around it.

If we use the MIL for the likelihoods, ie. $p_1(x^n|\theta_1) = p_1^*(x^n|\theta_1)$, $p_2(y^n|\theta_2) = p_2^*(y^n|\theta_2)$, and take $w_1(\cdot,\cdot) = w_2(\cdot,\cdot)$. Then for the independent case (i.e. $p_k^*(x^n|\theta) = \prod_{i=1}^n p_k^*(x_i|\theta)$, for $k = 1, 2$, we have

$$R_{(2,3)}(s^n, z^n, \theta, \phi) =$$

$$\frac{\prod_{i=1}^n \int [m_1^*(\frac{v_i+z_i}{2}) m_2^*(\frac{v_i-z_i}{2})] e^{-\lambda_1 \sum_{i=1}^n L_1(\frac{v_i+z_i}{2}, \frac{\phi+\theta}{2}) - \lambda_2 \sum_{i=1}^n L_2(\frac{v_i-z_i}{2}, \frac{\phi-\theta}{2})} dv_i}{\prod_{i=1}^n [m_1^*(\frac{s_i+z_i}{2}) m_2^*(\frac{s_i-z_i}{2})] e^{-\lambda_1 \sum_{i=1}^n L_1(\frac{s_i+z_i}{2}, \frac{\phi+\theta}{2}) - \lambda_2 \sum_{i=1}^n L_2(\frac{s_i-z_i}{2}, \frac{\phi-\theta}{2})}}.$$

In some cases, we can calculate $R_{(2,3)}(s^n, z^n, \theta, \phi)$ in closed form. For example, if we further specify $L_k(x, y) = (x - y)^2$ for $k = 1, 2, \lambda_1 = \lambda_2 = 1$, and $w(\cdot, \cdot) = N(\mathbf{0}, I_2)$, then from Example 1.4.3, we know that $m_1^*(x) = m_2^*(x) \propto e^{-x^2}$. By the facts that

$$(\frac{v_i + z_i}{2} - \frac{\phi + \theta}{2})^2 + (\frac{v_i - z_i}{2} - \frac{\phi - \theta}{2})^2 = \frac{1}{2}[(v_i - \phi)^2 + (z_i - \theta)^2],$$

and

$$(\frac{v_i + z_i}{2})^2 + (\frac{v_i - z_i}{2})^2 = \frac{1}{2}[v_i^2 + z_i^2],$$

we have

$$\begin{aligned} R_{(2,3)}(s^n, z^n, \theta, \phi) &= \frac{\prod_{i=1}^n \int \exp\{-v_i^2 - \lambda(v_i - \phi)^2/2\} dv_i}{\prod_{i=1}^n \exp\{-s_i^2 - \lambda(s_i - \phi)^2/2\}} \\ &= \prod_{i=1}^n \sqrt{\frac{2\pi}{2\lambda + 1}} e^{\frac{\lambda+2}{2}(s_i - \frac{\lambda}{\lambda+2}\phi)^2}. \end{aligned}$$

We can state the corresponding results for Model I vs Model II and Model I vs Model III as in the following corollary.

Corollary 5.3.1 Let

$$p_1(x^n|\theta_1) = \frac{m_1^*(x^n) \exp\{-\lambda_1 L_1(x^n, \theta_1)\}}{\int m_1^*(t^n) \exp\{-\lambda_1 L_1(t^n, \theta_1)\} dt^n}$$

and

$$p_2(y^n|\theta_2) = \frac{m_2^*(y^n) \exp\{-\lambda_2 L_2(y^n, \theta_2)\}}{\int m_2^*(t^n) \exp\{-\lambda_2 L_2(t^n, \theta_2)\} dt^n}$$

be MIL's, and write the likelihood for Model I as

$$p_{(1)}(z^n|\theta) = \frac{m^*(z^n) \exp\{-\lambda L(z^n, \theta)\}}{\int m^*(t^n) \exp\{-\lambda L(t^n, \theta)\} dt^n}.$$

For Models I, II and III, choose the priors to be $w_1(\theta_1, \theta_2)$, $w_2(\theta_1, \theta_2)$ and $w_3(\theta_1, \theta_2)$, where $w_1(\theta) = \frac{1}{2} \int w_1(\frac{\phi+\theta}{2}, \frac{\phi-\theta}{2}) d\phi$. We have the following

(i) For comparing Model I to Model II we have

$$\int \left| w_{(1)}(\theta|z^n) - w_{(2)}(\theta|x^n, y^n) \right| d\theta \leq 2M_{(1,2)}(x^n, y^n),$$

$$\left|E_{(1)}(\theta|z^n) - E_{(2)}(\theta|x^n, y^n)\right| \leq (E_{(1)}(|\theta|\,|z^n) + E_{(2)}(|\theta|\,|x^n, y^n))M_{(1,2)}(x^n, y^n),$$

$$\left|\mathrm{Var}_{(1)}(\theta|z^n) - \mathrm{Var}_{(2)}(\theta|x^n, y^n)\right| \leq h_{(1,2)}(x^n, y^n)M_{(1,2)}(x^n, y^n),$$

where

$$M_{(1,2)}(x^n, y^n) = M_{(1,2)}(s^n, z^n) = \sup_{\phi,\theta} M_{(1,2)}(s^n, z^n, \phi, \theta),$$

$$M_{(1,2)}(s^n, z^n, \phi, \theta) = \min\{\left|1 - R_{(1,2)}(s^n, z^n, \phi, \theta)\right|, \left|1 - [R_{(1,2)}(s^n, z^n, \phi, \theta)]^{-1}\right|\}$$

and

$$R_{(1,2)}(s^n, z^n, \phi, \theta) = \frac{m_1^*(\frac{s^n+z^n}{2})\exp\{-\lambda_1 L_1(\frac{s^n+z^n}{2}, \frac{\phi+\theta}{2})\}}{\int m_1^*(t^n)\exp\{-\lambda_1 L_1(t^n, \frac{\phi+\theta}{2})\}dt^n} \times$$

$$\frac{m_2^*(\frac{s^n-z^n}{2})\exp\{-\lambda_2 L_2(\frac{s^n-z^n}{2}, \frac{\phi-\theta}{2})\}}{\int m_2^*(t^n)\exp\{-\lambda_2 L_2(t^n, \frac{\phi-\theta}{2})\}dt^n} \frac{\int m^*(t^n)\exp\{-\lambda L(t^n, \theta)\}dt^n}{m^*(z^n)\exp\{-\lambda L(z^n, \theta)\}} \frac{w_2(\frac{\phi+\theta}{2}, \frac{\phi-\theta}{2})}{w_1(\frac{\phi+\theta}{2}, \frac{\phi-\theta}{2})},$$

$$h_{(1,2)}(x^n, y^n) = h_{(1,2)}(s^n, z^n)$$

$$= E_{(1)}(\theta^2\,|z^n) + E_{(2)}(\theta^2\,|x^n, y^n) + 2(E_{(1)}(|\theta|\,|z^n) + E_{(2)}(|\theta|\,|x^n, y^n)).$$

(ii) For comparing Model I and Model II we have

$$\int \left|w_{(1)}(\theta|z^n) - w_{(3)}(\theta|z^n)\right| d\theta \leq 2M_{(1,3)}(x^n, y^n),$$

$$\left|E_{(1)}(\theta|z^n) - E_{(3)}(\theta|z^n)\right| \leq (E_{(1)}(|\theta|\,|z^n) + E_{(3)}(|\theta|\,|z^n))M_{(1,3)}(x^n, y^n),$$

$$\left|\mathrm{Var}_{(1)}(\theta|z^n) - \mathrm{Var}_{(3)}(\theta|z^n))\right| \leq h_{(1,3)}(x^n, y^n)M_{(1,3)}(x^n, y^n),$$

where

$$M_{(1,3)}(x^n, y^n) = M_{(1,3)}(s^n, z^n) = \sup_{\phi,\theta} M_{(1,3)}(s^n, z^n, \phi, \theta),$$

$$M_{(1,3)}(s^n, z^n, \phi, \theta) = \min\{\left|1 - R_{(1,3)}(s^n, z^n, \phi, \theta)\right|, \left|1 - [R_{(1,3)}(s^n, z^n, \phi, \theta)]^{-1}\right|\}$$

and

$$R_{(1,3)}(s^n, z^n, \phi, \theta) = \frac{w_3(\frac{\phi+\theta}{2}, \frac{\phi-\theta}{2})}{w_1(\frac{\phi+\theta}{2}, \frac{\phi-\theta}{2})}\frac{A}{B}$$

where

$$A = \int m_1^*(\frac{s^n + z^n}{2})\exp\{-\lambda_1 L_1(\frac{s^n + z^n}{2}, \frac{\phi + \theta}{2})\}m_2^*(\frac{s^n - z^n}{2}) \times$$

$$\exp\{-\lambda_2 L_2(\frac{s^n - z^n}{2}, \frac{\phi - \theta}{2})\}ds^n \int m^*(t^n)\exp\{-\lambda L(t^n, \theta)\}dt^n$$

$$B = \int m_1^*(t^n)\exp\{-\lambda_1 L_1(t^n, \frac{\phi + \theta}{2})\}dt^n \times$$

$$\int m_2^*(t^n)\exp\{-\lambda_2 L_2(t^n, \frac{\phi-\theta}{2})\}dt^n m^*(z^n)\exp\{-\lambda L(z^n,\theta)\}$$

and

$$h_{(1,3)}(x^n, y^n) = h_{(1,3)}(s^n, z^n) =$$

$$E_{(1)}(\theta^2\,|z^n) + E_{(3)}(\theta^2\,|z^n) + 2(E_{(1)}(|\theta|\,|z^n) + E_{(3)}(|\theta|\,|z^n))).\square$$

Let us now examine the robustness of the three models with respect to sets of possible data points that are likely when MIL's are used. For this we define a notion of typicality based on Theorem 3.1.1. That result suggests sets on which posteriors should be close to their respective priors under the mixture density. First let

$$I_1^*(n) = E_{m_1^*} D(w_1^*(\cdot|X^n)||w_1(\cdot)),$$

$$I_2^*(n) = E_{m_2^*} D(w_2^*(\cdot|Y^n)||w_2(\cdot)),$$

and

$$I_0^*(n) = E_{m^*} D(w^*(\cdot|Z^n)||w(\cdot)).$$

Now, Theorem 3.1.1 gives conditions under which $I_i^*(n) \to 0$, as $n \to \infty$. To use this fact, let $C_i(n)$, for $i = 0, 1, 2$ be sequences of constants such that as $n \to \infty$ we have

$$C_i(n) \to \infty, \qquad C_i(n) I_i^*(n) \to 0.$$

Next, for $i = 0, 1, 2$, let S_i be subsets of the sample space defined by

$$S_0 = \{Z^n : D(w^*(\cdot|Z^n)||w(\cdot)) \le C_0(n) I_0^*(n)\},$$

$$S_1 = \{X^n : D(w_1^*(\cdot|X^n)||w_1(\cdot)) \le C_1(n) I_1^*(n)\},$$

and

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

We call such data sets canonical. Let P_i^* be the probability measure corresponding to $p_i^*, (i = 0, 1, 2)$. The following proposition gives a sense in which the MIL probability of these canonical sets is large.

Proposition 5.3.2 For any pre-assigned $\epsilon > 0$, there exist subsets $A_{i,n}(\epsilon)$ for $i = 0, 1, 2$ in the domain of Θ, so that as $n \to \infty$,

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

That is, for large sample sizes, the canonical sets have large $P_i^*(\cdot|\theta)$ probabilities, for all values of θ in a set of arbitrarily large probability under the prior distribution $W_i(\cdot)$.

Proof: We only prove the statement for $i = 1$, the other cases are similar. We will omit the θ in P_1^* when it is notationally convenient and no confusion. Note that

$$\begin{aligned} P_1^*(S_1^c|\theta) &= P_1^*\Big(D(w_1^*(\cdot|X^n)||w_1(\cdot)) > C_1(n)I_1^*(n)\Big) \\ &= E_{P_1^*}\chi\Big(D(w_1^*(\cdot|X^n)||w_1(\cdot)) > C_1(n)I_1^*(n)\Big) \\ &\leq \frac{1}{C_1(n)I_1^*(n)} E_{P_1^*} D(w_1^*(\cdot|X^n)||w_1(\cdot)), \end{aligned} \tag{5.3.7}$$

where $\chi(A)$ is the indicator function for set A. Recall that

$$E_{P_1^*} D(w_1^*(\cdot|X^n)||w_1(\cdot)) = \int \frac{p_1^*(x^n|\theta)}{m_1^*(x^n)} \int p_1^*(x^n|\xi) w(\xi) \log \frac{p_1^*(x^n|\xi)}{m_1^*(x^n)} d\xi dx^n. \tag{5.3.8}$$

Since

$$\frac{p_1^*(x^n|\theta)}{m_1^*(x^n)} = \frac{e^{-L_n(x^n,\theta)}}{\int m_1^*(t^n) \exp\{-L_n(t^n,\theta_1)\} dt^n},$$

we denote the inverse of the denominator of the right hand side of the last expression by $h(n,\theta)$. By definition we require

$$\int \frac{e^{-L_n(x^n,\theta)} w(\theta)}{\int m_1^*(t^n) \exp\{-L_n(t^n,\theta)\} dt^n} d\theta \leq 1$$

for all x^n. That is, for all x^n we have

$$\int e^{-L(x^n,\theta)} h(n,\theta) w(\theta) d\theta \leq 1. \tag{5.3.9}$$

We show that for any pre-assigned $\epsilon > 0$ there exists $A_{1,n}$, such that $e^{-L(x^n,\theta)} h(n,\theta)$ is bounded by a number N on $A_{1,n}$ uniformly in n and x^n with $W_1(A_{1,n}^c) \leq \epsilon$. Now, by way of contradiction, suppose there is an ϵ_0 and 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

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

Then,

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

contradicting (5.3.9). Now we have

$$\frac{p_1^*(x^n|\theta)}{m_1^*(x^n)} = e^{-L(x^n,\theta)}h(n,\theta) \leq N, \qquad \forall\theta \in A_{1,n},\ \forall n,\ \ \forall x^n. \tag{5.3.10}$$

By (5.3.8), (5.3.9) and (5.3.10), $\forall\theta \in A_{1,n}$ and $\forall n$ we have

$$P_1^*(S_1^c|\theta) \leq \frac{1}{C_1(n)I_1^*(n)} N \int\int p_1^*(x^n|\xi)w(\xi)\log\frac{p_1^*(x^n|\xi)}{m_1^*(x^n)}d\xi dx^n$$
$$= \frac{1}{C_1(n)I_1^*(n)} N I_1^*(n) = \frac{N}{C_1(n)} \to 0,$$

as $n \to \infty$, since $C_1(n) \to \infty$ by assumption. □

Let $\mathcal{B}$ be the Borel field on the unidimensional parameter space Θ, and let $W_{(i)}(\cdot|Data)$ be the posterior probability measure corresponding to $w_{(i)}(\cdot|Data)$. Here, $Data$ means the full data set or whatever summary statistics are being used to form the posterior. If we use MIL's (here we mean the multi-dimensional MIL's, not the product of one-dimensional MIL's) and assume that the data came from a canonical set then the posterior probabilities converge to a common limit in variation distance. We have the following.

Theorem 5.3.1 Let Model I be obtained from the prior $w(\cdot)$ as in (ii) of Proposition 5.2.1, and suppose Models II and III are obtained by using the priors $w_{(1)}(\cdot)$ and $w_2(\cdot)$ as described in Section 5.2. Under the conditions of Theorem 3.1.1, if Models I, II, and III are formed from MIL's then 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 $C_{i,j}(n)$ tends to zero as n tends to infinity. 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 $\theta_1 \in A_{1,n}, \theta_2 \in A_{2,n}$ where $A_{1,n}$ and $A_{2,n}$ are as in Proposition 5.3.2. Convergences in probability are assessed with the appropriate mixture density.

Remark 1: Theorem 5.3.1 is a sense in which Bayesian inferences using the MIL's and canonical data sets are insensitive to which of the three models is used, at least for large sample sizes. In particular, this means that the conclusions of Bayesian hypothesis tests based on posteriors using MIL's are also robust against modeling strategy for canonical data. This follows from recalling that a Bayes hypothesis test is based on the posterior odds ratio or, equivalently, the posterior probability of the null. This is a limited form of robustness against modeling strategy.

Remark 2: Our strategy of proof is to expand $W^*_{(i)}(B|Data)$'s as $W(B)$ plus the corresponding error terms. For example, we write

$$W^*_{(1)}(B|Z^n) = \int_B w(\theta)d\theta + \int_B \Big(w^*_{(1)}(\theta|Z^n) - w(\theta)\Big)d\theta,$$

and prove the second term is negligible. For $W^*_{(2)}(B|X^n, Y^n)$ and $W^*_{(3)}(B|Z^n)$, the treatments are similar. Note that this holds because convergences are assessed in a mixture density not in an *iid* density. We use the mixture density so that canonical data can be defined, however, in practice, one would assume that there is a true value of the parameter.

Proof: First we prove the conclusion for models I and II. We first 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

$$\Big|\int_B \Big(w^*_{(1)}(\theta|Z^n) - w(\theta)\Big)d\theta\Big| \leq \int |w^*_{(1)}(\theta|Z^n) - w(\theta)|d\theta$$

$$\leq \frac{1}{\sqrt{2\ln 2}}\sqrt{D(w^*(\cdot|Z^n)||w(\cdot))} \leq \frac{1}{\sqrt{2\ln 2}}\sqrt{C_0(n)I^*_0(n)},$$

Now,

$$\Big|W^*_{(1)}(B|Z^n) - W(B)\Big| \leq \frac{1}{\sqrt{2\ln 2}}\sqrt{C_0(n)I^*_0(n)}, \tag{5.3.11}$$

where the error term $\frac{1}{\sqrt{2\ln 2}}\sqrt{C_0(n)I^*_0(n)}$ is o(1) as $n \to \infty$, for any data X^n, Y^n giving $Z^n \in S_0$.

Similarly,

$$W^*_{(2)}(B|X^n, Y^n) = \int_B \Big(\frac{1}{2}\int w^*_1(\frac{\phi+\theta}{2}|X^n)w^*_2(\frac{\phi-\theta}{2}|Y^n)d\phi\Big)d\theta$$

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

where, the first term above is $W(B)$ and the error terms

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

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

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 gives bounds on the error terms. We have

$$\begin{aligned}
|J_{2,1}(X^n)| &\leq \frac{1}{2}\int\int |w_1^*(\frac{\phi+\theta}{2}|X^n) - w_1(\frac{\phi+\theta}{2})| w_2(\frac{\phi-\theta}{2}) d\phi d\theta \\
&= \int\int |w_1^*(\theta_1|X^n) - w_1(\theta_1)| w_2(\theta_2) d\theta_1 d\theta_2 \\
&\leq \frac{1}{\sqrt{2\ln 2}}\sqrt{C_1(n) I_1^*(n)} \int w_2(\theta_2) d\theta_2 = \frac{1}{\sqrt{2\ln 2}}\sqrt{C_1(n) I_1^*(n)}, \qquad (5.3.12)
\end{aligned}$$

thus, $J_{2,1}(X^n)$ tends to zero as n tends to infinity.

Similarly, for canonical data X^n, Y^n,

$$\begin{aligned}
|J_{2,2}(X^n, Y^n)| &\leq \int \frac{1}{2}\int |w_2^*(\frac{\phi-\theta}{2}|Y^n) - w_2(\frac{\phi-\theta}{2})| w_1^*(\frac{\phi+\theta}{2}|X^n) d\phi d\theta \\
&= \int\int |w_2^*(\theta_2|Y^n) - w_2(\theta_2)| w_1^*(\theta_1|X^n) d\theta_2 d\theta_1 \\
&\leq \frac{1}{\sqrt{2\ln 2}}\sqrt{C_2(n) I_2^*(n)}, \qquad (5.3.13)
\end{aligned}$$

so $J_{2,2}(X^n, Y^n)$ tends to zero as n tends to infinity. Thus, by (5.3.12) and (5.3.13),

$$\left|W_{(2)}^*(B|X^n, Y^n) - W(B)\right| \leq b_2(n), \qquad (5.3.14)$$

where $b_2(n) = \frac{1}{\sqrt{2\ln 2}}\left(\sqrt{C_1(n) I_1^*(n)} + \sqrt{C_2(n) I_2^*(n)}\right)$. Now by (5.3.11) and (5.3.14), the conclusion is true with $C_{1,2} = \frac{1}{\sqrt{2\ln 2}}\sqrt{C_0(n) I_0^*(n)} + b_2(n)$, which tends to zero as n tends to infinity.

Now we prove the conclusion for models I and III, II and III. We first express $W_{(3)}(B|Z^n)$ as $W(B)$ plus an error term. since

$$w_{(3)}(\theta|Z^n) = \frac{\frac{1}{4}\int\int p_1^*(\frac{s^n+Z^n}{2}|\frac{\phi+\theta}{2}) p_2^*(\frac{s^n-Z^n}{2}|\frac{\phi-\theta}{2}) ds^n w_1(\frac{\phi+\theta}{2}) w_2(\frac{\phi-\theta}{2}) d\phi}{\int\int\int p_1^*(\frac{s^n+Z^n}{2}|\xi_1) p_2^*(\frac{s^n-Z^n}{2}|\xi_2) w_1(\xi_1) w_2(\xi_2) d\xi_1 d\xi_2 ds^n},$$

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

where

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

and

$$m_{(3)}^*(Z^n) = \frac{1}{2}\int\int p_1^*\Big(\frac{s^n+Z^n}{2}|\xi_1\Big)p_2^*\Big(\frac{s^n-Z^n}{2}|\xi_2\Big)w_1(\xi_1)w_2(\xi_2)d\xi_1 d\xi_2 ds^n$$
$$= \frac{1}{2}\int m_1^*(\frac{s^n+Z^n}{2})m_2^*(\frac{s^n+Z^n}{2})ds^n. \tag{5.3.15}$$

So, $\forall B \in \mathcal{B}$, we have that $W_{(3)}^*(B|Z^n)$ can be written as

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

Let

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

then $\int h(s^n,Z^n)ds^n = 1$. So, for fixed $Z^n = z^n$, $h(\cdot,Z^n)$ is a probability density. Now

$$W_{(3)}^*(B|Z^n) = \int[\int_B \Big(\frac{1}{2}\int w_1^*(\frac{\phi+\theta}{2}|\frac{s^n+Z^n}{2})w_2^*(\frac{\phi-\theta}{2}|\frac{s^n-Z^n}{2})d\phi\Big)d\theta]h(s^n,Z^n)ds^n$$
$$= W(B) + \int J_{2,1}(\frac{s^n+Z^n}{2})h(s^n,Z^n)ds^n + \int J_{2,2}(\frac{s^n+Z^n}{2},\frac{s^n-Z^n}{2})h(s^n,Z^n)ds^n,$$

by adding and subtracting $W(B)$ and

$$\int\int_B \frac{1}{2}\int w_1^*(\frac{\phi+\theta}{2}|\frac{s^n+Z^n}{2})w_2(\frac{\phi-\theta}{2})h(s^n,Z^n)d\phi d\theta ds^n.$$

Recall that by the definition of S_1 in Proposition 5.3.2, we have

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

where $2S_1 - Z^n$ is the set of all s^n's, for fixed $Z^n = z^n$, such that $\frac{s^n+Z^n}{2} \in S_1$. By (5.3.12), the first term in the right hand side above is bounded by $\frac{1}{\sqrt{2\ln 2}}\sqrt{C_1(n)I_1^*(n)}$. For the second term on the right hand side, note that $J_{2,1}$ is bounded and $h(\cdot,Z^n)$ is a density, we may

apply Proposition 5.3.2 to assert that $P_1^*(2S_1^c - Z^n)$ tends to zero as n tends to infinity. Thus, the second term on the right hand side converges to zero in P_1^* probability, i.e. it is $o_{P_1^*}(1)$.

Similarly,

$$\begin{aligned}
&|\int J_{2,2}(\frac{s^n+Z^n}{2}, \frac{s^n-Z^n}{2})h(s^n, Z^n)ds^n| \\
\leq &|\int_{2S_2+Z^n} J_{2,2}(\frac{s^n+Z^n}{2}, \frac{s^n-Z^n}{2})h(s^n, Z^n)ds^n| \\
+&|\int_{2S_2^c+Z^n} J_{2,2}(\frac{s^n+Z^n}{2}, \frac{s^n-Z^n}{2})h(s^n, Z^n)ds^n|,
\end{aligned}$$

by (5.3.13), the first term above is bounded by $\frac{1}{\sqrt{2\ln 2}}\sqrt{C_1(n)I_1^*(n)}$. For the second term, the argument is similar as before: $J_{2,2}(\cdot,\cdot)$ is bounded and $h(\cdot, Z^n)$ is a density, Proposition 5.3.1 asserts that the set $2S_2^c + Z^n$ is $o_{p_2^*}(1)$. So is the second term in the right hand side above. Thus we have

$$|W_{(3)}^*(B|Z^n) - W(B)| \leq b_2(n) + o_{p_1^*}(1) + o_{p_2^*}(1). \tag{5.3.16}$$

Now since $b_2(n) \to 0$ as $n \to \infty$, by (5.3.11) and (5.3.16) we get the conclusion for Model I and III. By (5.3.14) and (5.3.16), we get the conclusion for models II and III. □

Since the three models rarely coincide but are similar in a general sense, we are interested in which pairs of them are closer together or further apart. The following proposition tells us that, roughly speaking, Models II and III are the closest, and Models I and III differ most. This is consistent with intuition. Since Models II and III start from the same likelihoods, the difference is that Model II is transformed once in the parameters and Model III is transformed twice, in both data and parameters. This additional transformation make Model III differs most from Model I. For Model I, the likelihood is different from that of Models II and III, and it differs from Models II and III than they do from each other.

Proposition 5.3.3 Assume the general likelihoods as described in the beginning of Chapter 5 (so the results including all the models we considered in Chapter 4), we have
(i)

$$E_{(X^n,Y^n)}\Big(D(w_{(2)}(\cdot|X^n,Y^n)||w_{(1)}(\cdot|Z^n)) - D(w_{(2)}(\cdot|X^n,Y^n)||w_{(3)}(\cdot|Z^n))\Big) > 0.$$

(ii)

$$E_{(X^n,Y^n)}\Big(D(w_{(3)}(\cdot|Z^n)||w_{(1)}(\cdot|Z^n)) - D(w_{(3)}(\cdot|Z^n)||w_{(2)}(\cdot|X^n,Y^n))\Big) > 0.$$

(iii)

$$E_{(X^n,Y^n)}\Big(D(w_{(1)}(\cdot|Z^n)||w_{(3)})(\cdot|Z^n)) - D(w_{(1)}(\cdot|Z^n)||w_{(2)}(\cdot|X^n,Y^n)) > 0,$$

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

Proof: (i) Let the marginal density for Z^n be $m(z^n)$ obtained from the joint marginal $m_1(x^n)m_2(y^n)$ by the transformation

$$Z^n = X^n - Y^n, \quad S^n = X^n + Y^n$$

and integrating out s^n. Now

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

which is the same as $m_{(3)}(z^n)$, the marginal density for model 3,

$$m_{(3)}(z^n) = \frac{1}{2}\int p_1(\frac{s^n + Z^n}{2}|\eta_1)p_2(\frac{s^n - Z^n}{2}|\eta_2)w_1(\eta_1)w2(\eta_2)d\eta_1 d\eta_2$$

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

So,

$$E_{(X^n,Y^n)}\Big(D(w_{(2)}(\cdot|X^n,Y^n)||w_{(1)}(\cdot|Z^n)) - D(w_{(2)}(\cdot|X^n,Y^n)||w_{(3)}(\cdot|Z^n))\Big)$$

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

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

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

$$\log\frac{\int\int w_1(\frac{\phi+\theta}{2}|\frac{s^n+Z^n}{2})w_2(\frac{\phi-\theta}{2}|\frac{s^n-Z^n}{2})m_1(\frac{s^n+Z^n}{2})m_2(\frac{s^n-Z^n}{2})d\phi ds^n}{4m_{(3)}(Z^n)w_{(1)}(\theta|Z^n)}d\theta\Big). \qquad (5.3.17)$$

The log-sum inequality states that for any integer n and any non-negative numbers $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$, we have

$$(\sum_{i=1}^n a_i)\log\frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i} \leq \sum_{i=1}^n a_i \log\frac{a_i}{b_i},$$

with equality if and only if a_i/b_i is a constant over all i. Using this inequality, the right hand side of (5.3.17) is greater than

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

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

$$= \int \Big(\int\int\int \frac{1}{2} w_1(\frac{\phi+\theta}{2}|x^n) w_2(\frac{\phi-\theta}{2}|y^n) m_1(x^n) m_2(y^n) dx^n dy^n d\phi \Big)$$

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

$$= \int \Big(\frac{1}{2} \int\int\int w_1(\frac{\phi+\theta}{2}|x^n) w_2(\frac{\phi-\theta}{2}|y^n) m_1(x^n) m_2(y^n) dx^n dy^n d\phi \Big)$$

$$\log \frac{\frac{1}{2} \int\int\int w_1(\frac{\phi+\theta}{2}|x^n) w_2(\frac{\phi-\theta}{2}|y^n) m_1(x^n) m_2(y^n) dx^n dy^n d\phi}{E_{Z^n} w_{(1)}(\theta|Z^n)} d\theta \geq 0,$$

since it is the relative entropy between two densities of θ.

(ii) The proof is similar to that of (i).

(iii) It is enough to note that

$$E_{(X^n,Y^n)} \Big(D(w_{(1)}(\cdot|Z^n)||w_{(3)}(\cdot|Z^n)) - D(w_{(1)}(\cdot|Z^n)||w_{(2)}(\cdot|X^n,Y^n) \Big)$$

$$= E_{(X^n,Y^n)} \Big(\int w_{(1)}(\theta|Z^n) \Big(\log \frac{1}{2} \int w_1(\frac{\phi+\theta}{2}|X^n) w_2(\frac{\phi-\theta}{2}|Y^n) d\phi$$

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

$$m_1(\frac{s^n+Z^n}{2}) m_2(\frac{s^n-Z^n}{2}) d\phi ds^n \Big) d\theta$$

$$\geq \int\int E_{Z^n} w_{(1)}(\theta|Z^n)$$

$$\log \frac{\frac{1}{2} \int\int\int w_1(\frac{\phi+\theta}{2}|x^n) w_2(\frac{\phi-\theta}{2}|y^n) m_1(x^n) m_2(y^n) dx^n dy^n d\phi}{\frac{1}{4} \int\int\int w_1(\frac{\phi+\theta}{2}|\frac{s^n+z^n}{2}) w_2(\frac{\phi-\theta}{2}|\frac{s^n-z^n}{2}) m_1(\frac{s^n+z^n}{2}) m_2(\frac{s^n-z^n}{2}) ds^n dz^n d\phi} d\theta$$

$$= \int\int E_{Z^n} w_{(1)}(\theta|Z^n)$$

$$\log \frac{\int\int\int w_1(\frac{\phi+\theta}{2}|x^n) w_2(\frac{\phi-\theta}{2}|y^n) m_1(x^n) m_2(y^n) dx^n dy^n d\phi}{\int\int\int w_1(\frac{\phi+\theta}{2}|x^n) w_2(\frac{\phi-\theta}{2}|y^n) m_1(x^n) m_2(y^n) dx^n dy^n d\phi} = 0. \qquad \square$$

Comment: In Chapter 4, the data analysis used Models I and II and we found that they gave similar conclusions. According to Proposition 5.3.3, we can expect that the conclusions from Model III to be close to those of Model II.

Models II and III are closer because both use pairs of data points and integrate out a function of the data. That they give weaker results may be attributed to the marginalization procedure. In marginalizing, we take an average over all possible data points. This is different from fixing a particular data point, and so implicitly includes variation from two random variables whereas in Model I we have included variation from only one random variable. Since two random variables generally have more variability than one does, it seems Models II and III require more data to achieve inferences of the same strength as we got from Model I. Physically, one should ask if there are two random variables that are reasonable to model, or if we have only one random variable (like for Model I). This is a scientific question brought to light by our statistical analysis. However, statistical analysis alone cannot provide an answer.

The results from Model II and III are similar to that of Nader and Reboussin, Section 4.2, in which the conclusion is not clear-cut. Model I gives stronger inferences and clear answers. There is the question for the experimenter: are the model assumptions in Model I reasonable? i.e., do we have one random variable or must assume there are more of them?

By the above Proposition, Models 1 and III differs most on average. So the average discrepancy between Models I and III can be taken as a measure of robustness for the likelihood-prior quadruple $\{p_1(\cdot|\theta_1), p_2(\cdot|\theta_2), w_1(\theta_1), w_2(\theta_2)\}$ for the three modeling strategies. To simplify the problem, we may fix the priors, and for Model I, we take the prior $w_{(1)}(\theta) = \int \frac{1}{2} w_1(\frac{\phi+\theta}{2}) w_2(\frac{\phi-\theta}{2}) d\phi$, and the likelihood to be $p_{(1)}(z|\theta) = \frac{1}{4} p_1(\frac{s+z}{2}|\frac{\phi+\theta}{2}) p_2(\frac{s-z}{2}|\frac{\phi-\theta}{2}) ds d\phi$. Then the question becomes the robustness of the likelihood pairs $\{p_1(\cdot|\theta_1), p_2(\cdot|\theta_2)\}$ against the three modeling strategies. Specifically, we can use

$$R = \exp\{-E_{X,Y} D(w_{(3)}(\cdot|Z)||w_{(1)}(\cdot|Z))\} = \exp\{-E_{m_{(3)}(z)} D(w_{(3)}(\cdot|Z)||w_{(1)}(\cdot|Z))\}$$

to measure the robustness, where $Z = X - Y$. Or to simplify the calculation, we may use

$$R' = \exp\{-D(E_{X,Y}[w_{(3)}(\cdot|Z)]||D(E_{X,Y}[w_{(1)}(\cdot|Z)])\}$$

$$= \exp\{-D(w_{(3)}(\cdot)||E_{m_{(3)}(z)}[w_{(1)}(\cdot|Z)])\}$$

as the measure. Note that $0 \leq R \leq R' \leq 1$. Roughly speaking, the larger R' is, the more robust the likelihood pair is against the three models and R' can be arbitrarily close to 1 by choosing appropriate likelihoods and priors. To see this, note

$$E_{m_{(3)}(z)}[w_{(1)}(\cdot|Z)] = \int \frac{m_{(3)}(z)}{m_{(1)}(z)} p_{(1)}(z|\theta) w_{(1)}(\theta) dz,$$

where

$$m_{(1)}(z) = \frac{1}{4} \int \Big(\int \int p_1(\frac{s+z}{2}|\frac{\phi+\theta}{2}) p_2(\frac{s-z}{2}|\frac{\phi-\theta}{2}) ds d\phi \Big)$$
$$\Big(\frac{1}{2} \int w_1(\frac{\phi+\theta}{2}) w_2(\frac{\phi-\theta}{2}) d\phi \Big) d\theta,$$

and

$$m_{(3)}(z) = \frac{1}{4} \int \Big(\int \int p_1(\frac{s+z}{2}|\frac{\phi+\theta}{2}) p_2(\frac{s-z}{2}|\frac{\phi-\theta}{2}) ds$$
$$w_1(\frac{\phi+\theta}{2}) w_2(\frac{\phi-\theta}{2}) d\phi \Big) d\theta.$$

If we take the priors for Model I and III to be $w_{(1)}(\theta) = \int \frac{1}{2} w_1(\frac{\phi+\theta}{2}) w_2(\frac{\phi-\theta}{2}) d\phi = w_{(3)}(\theta)$ and the likelihoods for Models I and III, as $\{p_1(\cdot|\theta_1), p_2(\cdot|\theta_2)\}$, to be uniform densities on finite intervals, then the interval bounds may depend on θ_1, θ_2 respectively. If the intervals are large, then $m_{(1)}(\cdot) \approx m_{(3)}(\cdot)$, so $E_{m_{(3)}(z)}[w_{(1)}(\cdot|Z)] \approx \int p_{(1)}(z|\theta) w_{(1)}(\theta) dz = w_{(1)}(\cdot) = w_{(3)}(\cdot)$, thus $R' \approx 1$.

5.4 More Considerations for the Robustness Issue

Here we discuss some more considerations for the robustness issue addressed in the previous sections. These considerations are in the initial stage, the conditions imposed are too strong in practice, and the results are comparatively weak. However, we list these results here for further possible improvements.

We have compared Models I and II computationally in an example. Here we compare Models II and III theoretically. We begin by stating and proving sufficient conditions for the left hand side of (5.1.3) to reduce to a function $p(z^n|\theta)$ with $\theta = (\theta_1, \theta_2)$. Assume that $X_1, ..., X_n$ are *iid* $p_1(x|\theta_1)$, and that $Y_1, ..., Y_n$ are *iid* $p_2(y|\theta_2)$. Assume $Z^n = f(X^n, Y^n)$

has density of the form

$$p(z^n|\theta) = \prod_{i=1}^{n} p(z_i|\theta)$$

where

$$p(z|\theta) = \int p_1(g_1(s,z)|\theta_1)p_2(g_2(s,z)|\theta_2)J(s,z)ds,$$

and $\theta = f_1(\theta_1, \theta_2)$.

Because we have assumed independence, it is enough to examine the $n = 1$ case. Assume the distributions of X and Y are in a one parameter exponential family, ie.

$$p(x|\theta_1) = \exp\{\eta_1(\theta_1)T_1(x) - B_1(\theta_1)\}h_1(x), \tag{5.4.1}$$

$$p(y|\theta_2) = \exp\{\eta_2(\theta_2)T_2(x) - B_2(\theta_2)\}h_2(x). \tag{5.4.2}$$

Let $z = f_1(x,y), s = f_2(x,y)$, $\theta = f_1(\theta_1,\theta_2)$, $\phi = f_2(\theta_1,\theta_2)$. or $x = g_1(s,z), y = g_2(s,z)$ and $\theta_1 = g_1(\theta,\phi)$, $\theta_2 = g_2(\theta,\phi)$, and $J(s,z)$ is the transformation Jacobian. Now in the following Proposition, we give a sufficient condition for $p(z|\theta_1,\theta_2)$ to reduce to the form $p_{(3)}(z|\theta)$ for the exponential family. It basically says that if the exponents of p_1 and p_2 can be reformed into a sum of exponents from two exponential families for Z^n and S^n, then we get the reduction we want.

In this section, Proposition 5.4.1 and Theorem 5.4.2 are only for exponential families; Proposition 5.4.2, Proposition 5.4.4 and Theorem 5.4.1 are for the MILs in both the product form of n-dimensional dependent form; Proposition 5.4.3 is for any likelihoods as long as they satisfy the stated hypotheses.

Proposition 5.4.1 (Reduction in parameters for exponential families:) Suppose the distributions of p_1 and p_2 are given by (5.4.1) and (5.4.2) respectively. If

$$\begin{aligned}
&\eta_1(\theta_1)T_1(f_1(s,z)) - B_1(\theta_1) + \log h_1(f_1(s,z)) \\
&+\eta_2(\theta_2)T_2(f_2(s,z)) - B_2(\theta_2) + \log h_2(f_2(s,z)) + \log J(s,z) \\
&= \tilde{\eta}_1(\phi)\tilde{T}_1(s) - \tilde{B}_1(\phi) + \log \tilde{h}_1(s) + \tilde{\eta}_2(\theta)\tilde{T}_2(z) - \tilde{B}_2(\theta) + \log \tilde{h}_2(z)
\end{aligned} \tag{5.4.3}$$

for some functions $\tilde{\eta}_k(\cdot), \tilde{T}_k(\cdot), \tilde{B}_k(\cdot), \tilde{h}_k(\cdot)$ $(k = 1, 2)$, then

$$p(z^n|\theta_1, \theta_2) = p(z^n|\theta) = \exp\{\tilde{\eta}_2 \tilde{T}_2(z) - \tilde{B}_2(\theta)\}\tilde{h}_2(z).$$

ie. $p(z^n|\theta_1, \theta_2)$ reduces to a parameter only depend on θ.

Proof:

$$\begin{aligned} p(z|\theta_1, \theta_2) &= \int p_1(g_1(s,z)|\theta_1) p_2(g_2(s,z)|\theta_2) J(s,z) ds \\ &= \int \exp\{\eta_1(\theta_1) T_1(f_1(s,z)) - B_1(\theta_1) + \log h_1(f_1(s,z))\} \\ &\quad \exp\{\eta_2(\theta_2) T_2(f_2(s,z)) - B_2(\theta_2) + \log h_2(f_2(s,z)) + \log J(s,z)\} ds \\ &= \int \exp\{\tilde{\eta}_1(\phi)\tilde{T}_1(s) - \tilde{B}_1(\phi) + \log \tilde{h}_1(s) + \tilde{\eta}_2(\theta)\tilde{T}_2(z) - \tilde{B}_2(\theta) + \log \tilde{h}_2(z)\} ds \\ &= \exp\{\tilde{\eta}_2(\theta)\tilde{T}_2(z) - \tilde{B}_2(\phi)\}\tilde{h}_2(z). \qquad \square \end{aligned}$$

Remark: If $p_1(x|\theta_1)$ and $p_2(y|\theta_2)$ are normal densities with θ_1, θ_2 be their respective location parameters, (without loss of generality assume their variance is 1) then $\eta_i(\theta_i) = \theta_i$, $T_i(x) = x$, $B_i(\theta) = \theta^2$, $h_i(x) = 1$ (i = 1,2). Assume $f_1(x,y) = x + y, f_2(x,y) = x - y$, then (6.2.2.2) is satisfied with $\tilde{\eta}_i(\theta) = \frac{1}{2}\theta$, $\tilde{T}_i(s) = s$, $\tilde{B}_i(\theta) = \frac{1}{2}\theta^2$, $\tilde{h}_i(x) = 1$, $(i = 1, 2)$ and $J(s,z) = \frac{1}{2}$.

Next we show that the conclusion of Proposition 5.4.1 holds for certain MIL's. Note that MIL's have a form similar to that assumed in Proposition 5.4.1, but they are different.

Proposition 5.4.2 Let $p_1(x^n|\theta_1)$ and $p_2(y^n|\theta_2)$ be MIL's denoted by $p_1^*(x^n|\theta_1)$ and $p_2^*(y^n|\theta_2)$. (i) If the marginal priors $w_1(\cdot)$ and $w_2(\cdot)$ and the marginal densities $m_1^*(\cdot)$ and $m_2^*(\cdot)$ are the uniform densities on R^n, and the $L_k(\cdot,\cdot)$ in p_k^* $(k = 1, 2)$ satisfies

$$L_1(x, \theta_1) + L_2(y, \theta_2) = \underline{r}_1(f_1(x,y), f_1(\theta_1, \theta_2)) + \underline{r}_2(f_2(x,y), f_2(\theta_1, \theta_2)),$$
$$\forall x, y, \theta_1, \theta_2 \in R^1, \qquad (5.4.4)$$

with $\underline{r}_k(t,\theta) = \underline{r}_k(|t-\theta|) \geq 0, (k=1,2)$, then again $p(z^n|\theta_1,\theta_2)$ reduces to $p(z^n|\theta)$ and

$$p(z^n|\theta) = c(n)\exp(-\lambda_2\underline{r}_2(z^n,\theta))\int \exp(-\lambda_1\underline{r}_1(s^n,\theta))J(s^n,z^n)ds^n.$$

where $c(n)$ is the normalizing constant.

(ii) For $n=1$, if the marginal priors $w_1(\cdot)$ and $w_2(\cdot)$ are the uniform densities on R^1, then $m_1^*(\cdot)$ and $m_2^*(\cdot)$ are the uniform densities on R^1.

Remark 1: If $L_1(t,\theta) = L_2(t,\theta) = (t-\theta)^2$, $f_1(x,y) = x+y$, then (5.4.4) is true with $\underline{r}_1(t,\theta) = \underline{r}_2(t,\theta) = \frac{1}{2}(t-\theta)^2$.

Remark 2: If $m_1^*(\cdot)$ or $m_2^*(\cdot)$ is not uniform, $p(z^n|\theta_1,\theta_2)$ will not necessarily reduce to the form $p(z^n|\theta)$. For example, if we take $w_1(\cdot) = w_2(\cdot) = N(0,1)$, $L_1(t,\theta) = L_2(t,\theta) = (t-\theta)^2$, $\lambda_1 = \lambda_2 = 1$, $n=1$, $z^n = x^n + y^n$, then form Example 1.4.3, we know that $m_k^*(x) = \frac{1}{\sqrt{\pi}}e^{-x^2}(k=1,2)$, and so $p(z^n|\theta_1,\theta_2) \propto \exp\{-\frac{1}{2}(\theta_1-z)^2 - \frac{1}{2}(\theta_2+z)^2\}$ which cannot be reduced to the form $p(z|\theta)$.

Proof:

(i) By definition, the density for Z^n is

$$\int p_1^*(g_1(s^n,z^n)|\theta_1)p_2^*(g_2(s^n,z^n)|\theta_2)J(s^n,z^n)ds^n$$

$$= \int \frac{m_1^*(g_1(s^n,z^n))m_2^*(g_2(s^n,z^n))e^{-\lambda_1 L_1(g_1(s^n,z^n),\theta_1)-\lambda_2 L_2(g_2(s^n,z^n),\theta_2)}}{\int m_1^*(t^n)e^{-\lambda_1 L_1(t^n,\theta_1)}dt^n \int m_2^*(t^n)e^{-\lambda_2 L_2(t^n,\theta_2)}dt^n} J(s^n,z^n)ds^n.$$

By the assumptions on the priors and marginals we get

$$\int \frac{\exp\{-\lambda_1\underline{r}_1(s^n,\phi) - \lambda_2\underline{r}_2(z^n,\theta)\}}{\int \exp\{-\lambda_1 L_1(t^n,\theta_1)\}dt^n \int(t^n)\exp\{-\lambda_2 L_2(t^n,\theta_2)\}dt^n} J(s^n,z^n)ds^n$$

which reduces, by (5.4.3), to

$$c(n)\exp(-\lambda_2\underline{r}_2(z^n,\theta))\int \exp(-\lambda_1\underline{r}_1(s^n,\theta))J(s^n,z^n)ds^n.$$

(ii) In the case $n=1$, since $m_1^*(\cdot)$ is uniquely determined by the inequality

$$\int \frac{\exp\{-\lambda_1 L_1(x,\theta_1)\}w_1(\theta_1)}{\int m_1^*(t)\exp\{-\lambda_1 L_1(t,\theta_1)\}dt}d\theta_1 \leq 1,$$

with equality for all x in the support of $m_1^*(\cdot)$. We see that if $w_1(\cdot)$ is uniform on R^1, then the uniform density for $m_1^*(\cdot)$ satisfies the above inequality. The same holds for $m_2^*(\cdot)$. □

As we will see later, the problem of the equivalence of models is much harder than that of the mean of the model, the conditions imposed are so stringent that only in very rare cases the equivalence can be guaranteed.

Now we show that for the exponential families we discussed in Proposition 5.4.1, if the density for $Z = f_1(X, Y)$ reduces to a likelihood in θ alone, then Model II and III are identical in a formal sense, namely the posteriors are the same.

Proposition 5.4.3

(i) Suppose the joint prior $w(\cdot, \cdot)$ satisfies

$$\frac{w(\phi, \theta)}{\int w(\xi, \theta) d\xi} = C,$$

and

$$p_1(\frac{s^n + z^n}{2} | \frac{\phi + \theta}{2}) p_2(\frac{s^n - z^n}{2} | \frac{\phi - \theta}{2}) = q_1(s^n, z^n) q_2(z^n | \phi, \theta)$$

for some non-negative q_1, q_2 where q_1 satisfies

$$\int q_1(s^n, z^n) ds^n = C,$$

then

$$w_{(2)}(\theta | x^n, y^n) = w_{(3)}(\theta | z^n).$$

(ii) Suppose $p_1(x|\theta_1)$ and $p_2(x|\theta_2)$ are given by (5.4.1) and (5.4.2) respectively, and (5.4.3) is satisfied. Assume further that $J(s, z)$ is constant, that $w(\theta_1, \theta_2) = w_1(\theta_1) w_2(\theta_2)$, and that

$$w(g_1(\phi, \theta), g_2(\phi, \theta)) J(\phi, \theta) = \bar{w}_1(\phi) \bar{w}_2(\theta), \quad \forall \phi, \theta$$

for some integrable $\bar{w}_1(\cdot)$ and $\bar{w}_2(\cdot)$. Then

$$w_{(2)}(\theta | x^n, y^n) = w_{(3)}(\theta | z^n).$$

Proof:(i) Let us abuse notation to write

$$f(s^n, z^n, |\phi, \theta) = p_1(\frac{s^n + z^n}{2}|\frac{\phi + \theta}{2})p_2(\frac{s^n - z^n}{2}|\frac{\phi - \theta}{2}),$$

and, for simplicity, denote

$$w(\frac{\phi + \theta}{2}, \frac{\phi - \theta}{2}) = w(\phi, \theta).$$

Since

$$w_{(2)}(\theta|x^n, y^n) \propto \int f(s^n, z^n, |\phi, \theta)w(\phi, \theta)d\phi,$$

and by the definition of Model III,

$$w_{(3)}(\theta|z^n) \propto \int \int f(s^n, z^n, |\phi, \theta)d\phi ds^n \int w(\phi, \theta)d\phi,$$

we have $w_{(2)}(\theta|x^n, y^n) = w_{(3)}(\theta|z^n)$ if and only if

$$\int f(s^n, z^n, |\phi, \theta)w(\phi, \theta)d\phi = q_1(s^n, z^n) \int \int f(s^n, z^n, |\phi, \theta)d\phi ds^n \int w(\phi, \theta)d\phi,$$

for some non-negative q_1, the last expression is equivalent to

$$\int [f(s^n, z^n, \phi, \theta)\frac{w(\phi, \theta)}{\int w(\xi, \theta)d\xi} - q_1(s^n, z^n) \int f(s^n, z^n, \phi, \theta)ds^n]d\phi = 0. \tag{5.4.5}$$

A sufficient condition for (5.4.5) is that the integrand itself be zero, i.e. $\forall z^n, \phi, \theta$,

$$f(s^n, z^n, |\phi, \theta)\frac{w(\phi, \theta)}{\int w(\xi, \theta)d\xi} = q_1(s^n, z^n) \int f(s^n, z^n, |\phi, \theta)ds^n.$$

If we omit the fixed variables z^n, ϕ, θ for simplicity, we get

$$q(s^n)f(s^n) = \lambda \int f(s^n)ds^n, \tag{5.4.6}$$

which is a Fredholm equation of the second type (See, for example, Kondo, 1991), where $q(s^n) = (q_1(s^n, z^n))^{-1}$, and $\lambda = \int w(\xi, \theta)d\xi / w(\phi, \theta)$. To solve this equation, divide by $\sqrt{q(s^n)}$ on both sides of (5.4.6) and let $Y(s^n) = \sqrt{q(s^n)}f(s^n)$. Now (5.4.6) becomes

$$Y(s^n) = \lambda \int \frac{1}{\sqrt{q(t^n)q(s^n)}}Y(t^n)dt^n. \tag{5.4.7}$$

Expression (5.4.7) has a solution if and only if

$$\lambda(\phi,\theta)\int \frac{ds^n}{q(s^n,z^n)} = 1.$$

Since this is guaranteed by the assumption on $q_1(\cdot,\cdot)$, the solution is given by

$$f(s^n,z^n,|\phi,\theta) = q_1(s^n,z^n)q_2(z^n,\phi,\theta) \tag{5.4.8}$$

where

$$q_2(z^n,\phi,\theta) = \int f(s^n,z^n,|\phi,\theta)ds^n$$

is non-negative, and for any non-negative $q_2(z^n,\phi,\theta)$, by substituting (5.4.8) into (5.4.6), it is seen that (5.4.8) is a solution for (5.4.6).

(ii) Since p_1 and p_2 are exponential families, we have

$$w_{(2)}(\theta|x^n,y^n) \propto \int \exp\{\eta_1(g_1(\phi,\theta))\sum_{i=1}^n T_1(x_i,y_i) + \eta_2(g_2(\phi,\theta))\sum_{i=1}^n T_2(x_i,y_i)$$
$$-nB_1(g_1(\phi,\theta)) - nB_2(g_2(\phi,\theta))\}w(g_1(\phi,\theta),g_2(\phi,\theta))J(\phi,\theta)d\phi,$$

Therefore

$$\eta_1(g_1(\phi,\theta))\sum_{i=1}^n T_1(x_i,y_i)) + \eta_2(g_2(\phi,\theta))\sum_{i=1}^n T_2(x_i,y_i)) - nB_1(g_1(\phi,\theta)) - nB_2(g_2(\phi,\theta))$$
$$= \tilde{\eta}_1(\phi)\sum_{i=1}^n \tilde{T}_1(g_1(x_i,y_i)) + \tilde{\eta}_2(\theta)\sum_{i=1}^n \tilde{T}_2(g_2(x_i,y_i)) - n\tilde{B}_1(\phi) - n\tilde{B}_2(\theta),$$

so we have

$$w_{(2)}(\theta|x^n,y^n) \propto \exp\{\tilde{\eta}_2(\theta)\sum_{i=1}^n \tilde{T}_2(g_2(x_i,y_i)) - n\tilde{B}_2(\theta)\tilde{w}_2(\theta)\}. \tag{5.4.9}$$

On the other hand, the density of Z^n has the reduced form

$$p_{(3)}(z^n|\theta) = \exp\{\tilde{\eta}_2(\theta)\tilde{T}_2(z^n) - n\tilde{B}_2(\theta)\}\tilde{h}_2(z^n),$$

where $\tilde{T}_2(z^n) = \sum_{i=1}^n \tilde{T}_2(z_i)$ and $\tilde{h}_2(z^n) = \prod_{i=1}^n \tilde{h}_2(z_i)$. So

$$w_{(3)}(\theta|z^n) \propto \exp\{\tilde{\eta}_2(\theta)\tilde{T}_2(z^n) - n\tilde{B}_2(\theta)\}\int w(g_1(\phi,\theta),g_2(\phi,\theta))J(\phi,\theta)d\phi$$

$$\propto \exp\{\tilde{\eta}_2(\theta)\tilde{T}_2(z^n) - n\tilde{B}_2(\theta)\}\tilde{w}_2(\theta). \tag{5.4.10}$$

For $w(\cdot,\cdot)$ satisfying the given conditions, the right hand side of (5.4.9) and (5.4.10) are proportional to

$$\exp\{\tilde{\eta}_2(\theta)\tilde{T}_2(z^n) - n\tilde{B}_2(\theta)\}\tilde{w}_2(\theta),$$

i.e. $w_{(2)}(\theta|x^n, y^n) = w_{(3)}(\theta|z^n)$, $\square$.

Remark: Let $w_1(\cdot) = w_2(\cdot)$ be the $N(0,1)$ density, $f_1(\theta_1,\theta_2) = \theta_1+\theta_2$, $f_2(\theta_1,\theta_2) = \theta_1-\theta_2$, then (ii) is satisfied with $J(\phi,\theta) = \frac{1}{2}$, $\bar{w}_1(\cdot)$ and $\bar{w}_2(\cdot)$ be the $N(0,1)$ density.

Note the sufficient conditions in (i) of Proposition 5.4.3 are not necessary. For example, let $n = 1$, $p_1(x|\theta) = p_2(x|\theta) = p^*(x|\theta) = m^*(x)e^{-L(x,\theta)}/$
$\int m^*(t)e^{-L(t,\theta)}dt$ with $L(x,y) = (x-y)^2$ and the prior $w(\theta_1,\theta_2)$ to be $N(\mathbf{0}, I_2)$. By Proposition 5.4.4 in the following we know that $w_{(2)}(\theta|x^n, y^n) = w_{(3)}(\theta|z^n)$, but the condition in Proposition 5.4.3 (i) is not satisfied. In fact from Example 1.4.3 we know that $p^*(\cdot|\theta) \sim N(\theta, 1/4)$, and

$$p_1(\frac{s+z}{2}|\frac{\phi+\theta}{2})p_2(\frac{s-z}{2}|\frac{\phi-\theta}{2}) = \exp\{-(s-\phi)^2 + (z-\theta)^2\},$$

which does not satisfy the conditions in (i) of Proposition 5.4.3.

We assume the uniform priors and marginals for MIL's as in Proposition 5.4.2 and give conditions to ensure the same results as in Proposition 5.4.3. Note that even if we use the uniform prior to generate the MIL's, we can still choose proper priors to form posteriors for inference.

Parallel to Proposition 5.4.3 for exponential families, we give conditions for Model II and III to give the same posterior when MIL's are used.

Proposition 5.4.4 Assume $p_1(x^n|\theta_1) = p_1^*(x^n|\theta_1), p_2(y^n|\theta_2) = p_2^*(y^n|\theta_2)$.
(i) Assume (i) of Proposition 5.4.2 and $w(\theta_1,\theta_2) = w_1(\theta_1)w_2(\theta_2)$. If we take the prior for inference on θ to be

$$w(\theta) = \int w(g_1(\phi,\theta), g_2(\phi,\theta))J(\phi,\theta)d\phi,$$

then

$$w_{(2)}(\theta|x^n, y^n) = w_{(3)}(\theta|z^n).$$

(ii) Let $n = 1, f_1(x,y) = x+y, f_2(x,y) = x-y, \lambda_1 = \lambda_2 = 1, L_k(x,\theta) = (x-\theta)^2, (k = 1,2)$. if $w(\cdot,\cdot)$ factors, i.e. $w(\cdot,\cdot) = w_1(\cdot)w_2(\cdot)$ and $w_1(\cdot) = w_2(\cdot)$ is the $N(0,1)$ density, then

$$w_{(2)}(\theta|x,y) = w_{(3)}(\theta|z)$$

Proof: (i) By Proposition 5.2.1, the density for Z^n is

$$p(z^n|\theta_1,\theta_2) = p(z^n|\theta) \propto \exp(-\lambda_2 \underline{r}_2(z^n,\theta)),$$

so

$$\begin{aligned} w_{(3)}(\theta|z^n) &= \frac{\exp\{-\lambda_2 \underline{r}_2(z^n,\theta)\}w(\theta)}{\int \exp\{-\lambda_2 \underline{r}_2(z^n,\xi)\}w(\xi)d\xi} \\ &= \frac{\exp\{-\lambda_2 \underline{r}_2(z^n,\theta)\}\int w(g_1(\phi,\theta),g_2(\phi,\theta))J(\phi,\theta)d\phi}{\int \exp\{-\lambda_2 \underline{r}_2(z^n,\xi)\}\int w(g_1(\xi_1,\xi_2)J(\xi_1,\xi_2)d\xi_1 d\xi_2} \\ &\propto \exp\{-\lambda_2 \underline{r}_2(z^n,\theta)\}\int w(g_1(\phi,\theta),g_2(\phi,\theta))J(\phi,\theta)d\phi. \end{aligned}$$

On the other hand,

$$w(\theta_1,\theta_2|x^n,y^n) = \frac{p_1^*(x^n|\theta_1)p_2^*(y^n|\theta_2)w(\theta_1,\theta_2)}{\int p_1^*(x^n|\xi_1)p_2^*(y^n|\xi_2)w(\xi_1,\xi_2)d\xi_1 d\xi_2},$$

so

$$w_{(2)}(\theta|x^n,y^n) = \int w(g_1(\phi,\theta),g_2(\phi,\theta)|x^n,y^n)J(\phi,\theta)d\phi$$

$$= \frac{m_1^*(x^n)m_2^*(y^n)}{m^*(x^n,y^n)}\int \frac{e^{-\lambda_1 L_1(x^n,g_1(\phi,\theta))-\lambda_2 L_2(y^n,g_2(\phi,\theta))}w(g_1(\phi,\theta),g_2(\phi,\theta))J(\phi,\theta)}{\int m_1^*(t^n)e^{-\lambda_1 L_1(t^n,g_1(\phi,\theta))}dt^n \int m_2^* e^{-\lambda_2 L_2(t^n,g_2(\phi,\theta))}dt^n}d\phi.$$

Since $w(\theta_1,\theta_2) = w_1(\theta_1)w_2(\theta_2)$, $m^*(x^n,y^n)$ factors into $m_1^*(x^n)m_2^*(y^n)$, so the above is

$$\int \frac{\exp\{-\lambda_1 \underline{r}_1(s^n,\phi)\}\exp\{-\lambda_2 \underline{r}_2(z^n,\theta)\}w(g_1(\phi,\theta),g_2(\phi,\theta))J(\phi,\theta)}{\int \exp\{-\lambda_1 L_1(t^n,g_1(\phi,\theta))\}dt^n \int \exp\{-\lambda_2 L_2(t^n,g_2(\phi,\theta))\}dt^n}d\phi$$

$$\propto \exp\{-\lambda_2 \underline{r}_2(z^n, \theta)\} \int w(g_1(\phi, \theta), g_2(\phi, \theta)) J(\phi, \theta) d\phi.$$

Thus,

$$w_{(3)}(\theta | z^n) = w_{(2)}(\theta | x^n, y^n).$$

(ii) We only assumed the MIL families, before marginalizing to get a posterior for θ given z, the posterior for (θ_1, θ_2) given z is given by

$$w_{(3)}(\theta_1, \theta_2 | z) = c(z) a(\theta_1, \theta_2 | z)$$

where

$$a(\theta_1, \theta_2 | z) = \frac{w(\theta_1, \theta_2) \int m_1^*(g_1(s,z)) m_2^*(g_2(s,z)) e^{-\lambda_1 L_1(g_1(s,z), \theta_1) - \lambda_2 L_2(g_2(s,z), \theta_2)} J(s,z) ds}{\int m_1^*(t) e^{-\lambda_1 L_1(t, \theta_1)} dt \int m_2^*(t) e^{-\lambda_2 L_2(t, \theta_2)} dt}$$

and

$$c(z) = \int \int \frac{\int m_1^*(g_1(s,z)) m_2^*(g_2(s,z)) e^{-\lambda_1 L_1(g_1(s,z), \xi_1) - \lambda_2 L_2(g_2(s,z), \xi_2)} w(\xi_1, \xi_2) J(s,z) ds}{\int m_1^*(t) e^{-\lambda_1 L_1(t, \xi_1)} dt \int m_2^*(t) e^{-\lambda_2 L_2(t, \xi_2)} dt} d\xi_1 d\xi_2.$$

Now, recall the posterior for θ under Model II is

$$w_{(2)}(\theta | x, y) = \frac{m_1^*(x) m_2^*(y)}{m^*(x,y)}$$
$$\times \int \frac{w(g_1(\phi, \theta), g_2(\phi, \theta)) e^{-\lambda_1 L_1(x, g_1(\phi, \theta)) - \lambda_2 L_2(y, g_2(\phi, \theta))} J(\phi, \theta)}{\int m_1^*(t) e^{-\lambda_1 L_1(t, g_1(\phi, \theta))} dt \int m_2^*(t) e^{-\lambda_2 L_2(t, g_2(\phi, \theta))} dt} d\phi.$$

Since $w(\cdot, \cdot) = w_1(\cdot) w_2(\cdot)$ with $w_1 = w_2 = N(0,1)$, and we have chosen $n = 1$, $L_k(x, \theta) = (x - \theta)^2$, $\lambda_k = 1$for $k = 1, 2$, $g_1(x,y) = \frac{x+y}{2}$ and $g_2(x,y) = \frac{x-y}{2}$, then we know from Example 1.4.3 that $m_k^*(x) = \frac{1}{\sqrt{\pi}} e^{-x^2}$ for $k = 1, 2$, and $J(\cdot, \cdot) = 1/2$. So

$$w_{(3)}(\theta_1, \theta_2 | z) \propto \exp\{-(z - \frac{\theta_1 - \theta_2}{2})^2 - \frac{1}{2}(\theta_1^2 + \theta_2^2)\}.$$

Thus,

$$w_{(3)}(\theta | z) = \frac{1}{2} \int w_{(3)}(\frac{\phi + \theta}{2}, \frac{\phi - \theta}{2}) d\phi \propto \exp\{-\frac{1}{2}(\theta - z)^2\}.$$

By the expression for $w_{(2)}(\theta | x^n, y^n)$ we get $w_{(2)}(\theta | x^n, y^n) \propto e^{-\frac{1}{2}(\theta - z)^2}$ and so, $w_{(3)}(\theta_1, \theta_2 | z) = w_{(2)}(\theta | x, y)$. □

If the conditions in Proposition 5.4.2 are not satisfied, we might still want to know how far away the posteriors from the different models are. In the following we use $||\cdot||$ to denote the L_1 norm, and $U(\cdot), U(\cdot,\cdot)$ to denote the uniform density on R^1, R^2 respectively. Let $\tilde{\theta}_{(i)}$ be the Bayes estimator of θ under model i using the common convex loss $h(\cdot,\cdot)$ (i.e. under a posterior $w_{(i)}(\theta|x^n, y^n)$ for $i = 1,2,3$. Note that h can be different from the L_k's used to get the MIL's p_k^*'s for $k = 1,2,3$. Let $E_{(i)}(\theta|X^n, Y^n), \mathrm{Var}_{(i)}(\theta|X^n, Y^n)$ denote the posterior expectation and variance of θ under Model i.

Theorem 5.4.1 Assume the MIL's for the likelihoods, then $\forall \epsilon > 0$, $\exists \delta_i > 0, (i = 1,2,3,4)$, such that if $||w(\cdot,\cdot) - U(\cdot,\cdot)|| \leq \delta_1$, $||m_1^*(\cdot) - U(\cdot)|| \leq \delta_2$, $||m_2^*(\cdot) - U(\cdot)|| \leq \delta_3$ and $L_k((\cdot,\cdot)$'s satisfy

$$\sup_{x,y,\theta_1,\theta_2} |L_1(x,\theta_1) + L_2(y,\theta_2) - \underline{r}_1(f_1(x,y), f_1(\theta_1,\theta_2)) - \underline{r}_2(f_2(x,y), f_2(\theta_1,\theta_2))| \leq \delta_4,$$

with $\underline{r}_k(t,\theta) = \underline{r}_k(|t-\theta|) \geq 0, (k = 1,2)$, then we have

(i)

$$||w_{(2)}(\theta|x^n, y^n) - w_{(3)}(\theta|z^n)|| \leq \epsilon,$$

and hence

$$|E_{(2)}(\theta|X^n, Y^n) - E_{(3)}(\theta|Z^n)| \leq \epsilon,$$

$$|\mathrm{Var}_{(2)}(\theta|X^n, Y^n) - \mathrm{Var}_{(3)}(\theta|Z^n)| \leq \epsilon.$$

(ii) For any $\epsilon > 0$, the Bayes estimators from the two models satisfy

$$|\hat{\theta}_{(2)} - \hat{\theta}_{(3)}| \leq \epsilon.$$

Proof: (i) Clearly, as a functional of $w(\cdot,\cdot)$, $m_1(\cdot)$ and $m_2(\cdot)$, the posteriors $w_{(j)}(\theta|x^n, y^n)$ $(j = 2,3)$ are continuous (in L_1 norm). Let $w_{(j),U,\tilde{L}}(\theta|x^n, y^n)$ $(j = 1,2,3)$ be the posterior density corresponding to $w(\cdot,\cdot) = U(\cdot,\cdot)$, $m_k(\cdot) = U(\cdot)$ and some loss function $\underline{r}_k$ satisfies the assumption of Theorem 5.4.1 $(k = 1,2)$. $w_{(j),\tilde{L}}(\theta|x^n, y^n)$ $(j = 1,2,3)$ be the quantity corresponding to $w(\cdot,\cdot) = U(\cdot,\cdot)$. By Propositions 5.4.2 and 5.4.3, $w_{(2),U,\underline{r}}(\theta|x^n, y^n) =$

$w_{(3),U,\underline{r}}(\theta|z^n)$. So

$$\begin{aligned}
&\int |w_{(2)}(\theta|x^n,y^n) - w_{(3)}(\theta|z^n)|d\theta \\
&\leq \int |w_{(2)}(\theta|x^n,y^n) - w_{(2),\underline{r}}(\theta|x^n,y^n)|d\theta \\
&\quad + \int |w_{(3)}(\theta|z^n) - w_{(3),U,\tilde{L}}(\theta|z^n)|d\theta. \qquad (5.4.11)
\end{aligned}$$

For simplicity, in the following we only discuss the first term in the right hand side of (5.4.11). It is

$$\begin{aligned}
&\int |w_{(2)}(\theta|x^n,y^n) - w_{(2),U,\tilde{L}}(\theta|x^n,y^n)|d\theta \\
&\leq \int |w_{(2)}(\theta|x^n,y^n) - w_{(2),\underline{r}}(\theta|x^n,y^n)|d\theta \\
&+ \int |w_{(2),\underline{r}}(\theta|x^n,y^n) - w_{(2),U,\underline{r}}(\theta|x^n,y^n)|d\theta.
\end{aligned}$$

The first term above can be made as small as we want by the continuity of $w_{(2)}(\theta|x^n,y^n)$ as a functional of $L_k(\cdot,\cdot)$'s. the second term in the above is

$$\begin{aligned}
&\int \left| \frac{m_1^*(x^n)m_2^*(y^n)e^{-\lambda_2\bar{c}_2\underline{r}_2(z^n,\theta)}}{m^*(x^n,y^n)} \right. \\
&\int \frac{e^{-\lambda_1\bar{c}_1\underline{r}_1(s^n,\phi)}w(g_1(\phi,\theta),g_2(\phi,\theta))J(\phi,\theta)}{\int m_1^*(t^n)e^{-\lambda_1\tilde{L}_1(t^n,g_1(\phi,\theta))}dt^n \int m_2^*(t^n)e^{-\lambda_2\tilde{L}_2(t^n,g_2(\phi,\theta))}dt^n}d\phi \\
&\quad - \frac{U(x^n)U(y^n)e^{-\lambda_2\bar{c}_2\underline{r}_2(z^n,\theta)}}{U(x^n,y^n)} \\
&\left. \int \frac{e^{-\lambda_1\bar{c}_1\underline{r}_1(s^n,\phi)}U(g_1(\phi,\theta),g_2(\phi,\theta))J(\phi,\theta)}{\int U(t^n)e^{-\lambda_1\tilde{L}_1(t^n,g_1(\phi,\theta))}dt^n \int U(t^n)e^{-\lambda_2\tilde{L}_2(t^n,g_2(\phi,\theta))}dt^n}d\phi \right| d\theta.
\end{aligned}$$

By adding and subtracting appropriate terms, the above can be bounded by $a_1||m_1^* - U|| + a_2||m_2^* - U|| + a_3||w - U||$ for some constants a_1, a_2 and a_3, except for θ, x^n, y^n in a set of small Lebesgue measure. The second term in the right-hand side of (5.4.11) can be bounded in a similar manner. Thus the right-hand side of (5.4.11) can be made smaller than ϵ for suitable choices of the δ's.

Also, since $E_{(i)}(\theta|x^n,y^n)$ and $\mathrm{Var}_{(i)}(\theta|x^n,y^n)$ are continuous functionals of $w_{(i)}$, the last two conclusions of (i) follow.

(ii) Since the loss is convex in a, the Bayes solution $\tilde{\theta}_{(i)}$ exists and is the unique minimizer of the posterior risk:

$$\tilde{\theta}_{(i)} = \arg\inf_{a\in A} R_{w_{(i)}}(a|x^n, y^n),$$

where

$$R_{w_{(i)}}(a|x^n, y^n) = \int h(a,\theta) w_{(i)}(\theta|x^n, y^n) d\theta,$$

and A is the action space. Since $R_{w_{(i)}}(a|x^n, y^n)$ is a continuous functional of $w_k(\cdot)$, $(k = 1, 2)$, so is $\tilde{\theta}_{(i)}$, (i = 2, 3). Also, $R_{w_{(i)}}(a|x^n, y^n) = R_{w_{(j)}}(a|x^n, y^n)$ for $(i, j = 2, 3)$ under the conditions of Proposition 5.4.4, so when these conditions are deviated a little (in the sense given in the L_1 conditions), the $R_{w_{(i)}}(a|x^n, y^n)$'s will also change a little, thus the conclusion true. □

Next, we establish a version of Theorem 5.4.1 for model II and III when exponential families are used.

Theorem 5.4.2 Assume

$$p_1(x|\theta_1) = \exp\{\eta_1(\theta_1)T_1(x) - B_1(\theta_1)\}h_1(x),$$

$$p(y|\theta_2) = \exp\{\eta_2(\theta_2)T_2(x) - B_2(\theta_2)\}h_2(x),$$

if we assume (5.4.3) is satisfied. and take $w(\theta) = \int w(g_1(\phi,\theta), g_2(\phi,\theta))J(\phi,\theta)d\phi$, then
(i) For any prespecified ϵ, we can choose δ such that

$$|E_{(2)}(\theta|X^n, Y^n) - E_{(3)}(\theta|Z^n)| \le \epsilon,$$

$$|Var_{(2)}(\theta|X^n, Y^n) - Var_{(3)}(\theta|Z^n)| \le \epsilon,$$

whenever $w(g_1(\phi,\theta))$ can be approximated, in the L_1 sense, by the product of two independent densities, i.e. $||w(g_1(\phi,\theta), g_2(\phi,\theta)) - \tilde{w}_1(\phi)\tilde{w}_2(\theta)|| \le \delta$ for some integrable $\tilde{w}_1(\cdot)$ and $\tilde{w}_2(\cdot)$.
(ii) The Bayes estimators $\tilde{\theta}_{(2)}$ and $\tilde{\theta}_{(3)}$ satisfy

$$|\tilde{\theta}_{(2)} - \tilde{\theta}_{(3)}| \le \epsilon.$$

Proof: (i) Since $E_{(i)}(\theta|X^n, Y^n)$ and $\text{Var}_{(i)}(\theta|X^n, Y^n)$ are continuous functional of $w(\cdot,\cdot)$. The $E_{(i)}$'s are equal, and the $\text{Var}_{(i)}$'s are equal for $w(g_1(\phi,\theta)$, $g_2(\phi,\theta)) = \tilde{w}_1(\phi)\tilde{w}_2(\theta)$, so for any prespecified ϵ, we can choose δ such that

$$|E_{(2)}(\theta|X^n, Y^n) - E_{(3)}(\theta|Z^n)| \leq \epsilon,$$

$$|Var_{(2)}(\theta|X^n, Y^n) - Var_{(3)}(\theta|Z^n)| \leq \epsilon,$$

whenever $||w(g_1(\phi,\theta), g_2(\phi,\theta)) - \tilde{w}_1(\phi)\tilde{w}_2(\theta)|| \leq \delta$.

For the second conclusion, the proof is similar to that in Theorem 5.4.1. □

We may also investigate the inference range for the three different Models. For a function $h(\cdot)$ integrable with respect to $w_{(i)}(\theta|x^n, y^n)$, $i = 1, 2, 3$, let $\mathcal{W}$ be the collection of the three posteriors $w_{(i)}(\theta|x^n, y^n)$, $i = 1, 2, 3$ based on the three likelihoods from the three models. Consider the interval

$$\left(E_{\min_{w\in\mathcal{W}}} h(\Theta), E_{\max_{w\in\mathcal{W}}} h(\Theta)\right)$$

and length of it. Since

$$w_{(2)}(\theta|x^n, y^n) = \frac{g(x^n + y^n, x^n - y^n|\theta)}{\int g(x^n + y^n, x^n - y^n|\xi)d\xi},$$

and

$$w_{(3)}(\theta|x^n, y^n) = \frac{\int g(s^n, x^n - y^n|\theta)ds^n}{\int\int g(s^n, x^n - y^n|\xi)ds^n d\xi},$$

we have

$$E_{w_{(2)}}[h(\Theta)] = \frac{\int h(\theta)g(x^n + y^n, x^n - y^n|\theta)d\theta}{\int g(x^n + y^n, x^n - y^n|\xi)d\xi},$$

and

$$E_{w_{(3)}}[h(\Theta)] = \frac{\int\int h(\theta)g(x^n + y^n, x^n - y^n|\theta)d\theta ds^n}{\int\int g(s^n, x^n - y^n|\xi)d\xi ds^n}.$$

So the difference

$$\begin{aligned} & E_{w_{(2)}}[h(\Theta)] - E_{w_{(3)}}[h(\Theta)] \\ &= \frac{\int g(x^n + y^n, x^n - y^n|\theta)[h(\theta) - 1]d\theta}{\int g(x^n + y^n, x^n - y^n|\theta)d\theta} \\ &+ \frac{\int\int h(\theta)[g(s^n, x^n - y^n|\theta) - 1]d\theta ds^n}{\int\int g(s^n, x^n - y^n|\xi)d\xi ds^n}. \end{aligned} \tag{5.4.12}$$

Similarly, assume $w(\theta) = \int w(\frac{\phi+\theta}{2}, \frac{\phi-\theta}{2})d\phi$. Since

$$w_{(1)}(\theta|x^n, y^n) = \frac{p(x^n - y^n|\theta) \int w(\frac{\phi+\theta}{2}, \frac{\phi-\theta}{2})d\phi}{\int\int p(x^n - y^n|\theta)w(\frac{\phi+\theta}{2}, \frac{\phi-\theta}{2})d\phi d\theta},$$

we get

$$\begin{aligned}
&E_{w_{(1)}}[h(\Theta)] - E_{w_{(2)}}[h(\Theta)] \\
&= \frac{\int h(\theta)p(x^n - y^n|\theta)w(\theta)d\theta}{\int p(x^n - y^n|\xi)w(\xi)d\xi \int g(x^n + y^n, x^n - y^n|\xi)d\xi} \\
&\times \int\int [p_1(x^n|\frac{\phi+\theta}{2})p_2(y^n|\frac{\phi-\theta}{2}) - p(x^n - y^n|\theta)]w(\frac{\phi+\theta}{2}, \frac{\phi-\theta}{2})d\phi d\theta \\
&+ \frac{\int\int h(\theta)[p(x^n - y^n|\theta) - p_1(x^n|\frac{\phi+\theta}{2})p_2(y^n|\frac{\phi-\theta}{2})]w(\frac{\phi+\theta}{2}, \frac{\phi-\theta}{2})d\phi d\theta}{\int g(x^n + y^n, x^n - y^n|\xi)d\xi}.
\end{aligned} \tag{5.4.13}$$

Likewise, we have,

$$\begin{aligned}
&E_{w_{(1)}}[h(\Theta)] - E_{w_{(3)}}[h(\Theta)] \\
&= \frac{\int h(\theta)p(x^n - y^n|\theta)w(\theta)d\theta}{\int p(x^n - y^n|\xi)d\xi \int\int g(s^n, z^n|\xi)d\xi ds^n} \\
&\int\int [p_1(\frac{\eta+z^n}{2}|\frac{\phi+\theta}{2})p_2(\frac{\eta-z^n}{2}|\frac{\phi-\theta}{2}) - p(x^n - y^n|\theta)]w(\frac{\phi+\theta}{2}, \frac{\phi-\theta}{2})d\phi d\theta \\
&+ \frac{\int\int h(\theta)[p(x^n - y^n|\theta) - p_1(x^n|\frac{\phi+\theta}{2})p_2(y^n|\frac{\phi-\theta}{2})]w(\frac{\phi+\theta}{2}, \frac{\phi-\theta}{2})d\phi d\theta}{\int g(x^n + y^n, x^n - y^n|\xi)d\xi}.
\end{aligned} \tag{5.4.14}$$

Likewise, we have,

$$\begin{aligned}
&E_{w_{(1)}}[h(\Theta)] - E_{w_{(3)}}[h(\Theta)] \\
&= \frac{\int h(\theta)p(x^n - y^n|\theta)w(\theta)d\theta}{\int p(x^n - y^n|\xi)d\xi \int\int g(s^n, z^n|\xi)d\xi ds^n} \\
&\int\int [p_1(\frac{\eta+z^n}{2}|\frac{\phi+\theta}{2})p_2(\frac{\eta-z^n}{2}|\frac{\phi-\theta}{2}) - p(x^n - y^n|\theta)]w(\frac{\phi+\theta}{2}, \frac{\phi-\theta}{2})d\phi d\theta \\
&+ \frac{\int\int h(\theta)[p(x^n - y^n|\theta) - p_1(\frac{\eta+z^n}{2}|\frac{\phi+\theta}{2})p_2(\frac{\eta-z^n}{2}|\frac{\phi-\theta}{2})]w(\frac{\phi+\theta}{2}, \frac{\phi-\theta}{2})d\phi d\theta}{\int\int g(s^n, z^n|\xi)d\xi ds^n}.
\end{aligned} \tag{5.4.15}$$

By using (5.4.13), (5.4.14) and (5.4.15) we may find the interval range as the posteriors vary among the three models.

Chapter 6

Discussion and Further Research

6.1 Discussion

Here we have proposed a technique for choosing a likelihood based on a given prior, a loss function and a distortion parameter. Given those three quantities, one can optimize the Shannon mutual information over a class of likelihoods to find the likelihood which makes the weakest possible assumptions in a precise information theoretic sense.

The assumptions implicit in this likelihood are also weak in two other statistical senses, formalized by the first two theorems. Theorem 3.1.1 shows that in the limit of large sample sizes, the expected relative entropy distance between a prior and a posterior formed from the minimally informative likelihood tends to zero. That is, the Shannon mutual information goes to zero. Theorem 3.2 states a small sample sense in which the MIL is minimally informative. Theorems 3.3.1 and 3.3.2 show how posteriors formed from minimally informative likelihoods depend on the distortion parameter. When the allowed Bayes risk increases, the posterior tends to the prior. When the allowed Bayes risk decreases to zero, the posterior tends to concentrate at the data. We may consider generalizing these theorems to the case of the n-fold product of univariate MIL's.

The main drawback to the use of these likelihoods is that at this point we have not compared the inferences they give to the inferences one would get from a true parametric family. In particular, one can conjecture that highest posterior density sets from the present likelihoods would be wider than one would get from the true parametric family and that there would be further deformation due to the fact that the parameter introduced here might not be exactly the same as the actual location parameter of interest, this would appear to follow from the estimating equation context in Chapter 3. Another drawback is that the computing required to find these likelihoods for several parameters or more than

one outcome may be onerous.

The formulation of the minimally informative likelihood permits a robustness analysis against choice of loss, λ and prior. In Figures 1 and 2, Section 1.6, we observed how the shape of $p^*(\cdot|\theta)$ varies as the prior and the parameter λ vary. For relatively small values of λ, the location of the $p^*(\cdot|\theta)$'s are approximately at θ, this may be due to the loss function punishing data points for being far away from the parameter and the exponential component of the MIL. For larger values of λ, the location of $p^*(\cdot|\theta)$ moves, independently of the prior, around the point where the minimal average loss is achieved, which in the squared error case, is the mean of the prior. This is predicted by Theorem 3.3.1. However the shape seems does not change significantly. For the left skewed $\exp(-(x+10)), (x \geq -10)$ prior, the non-skewed $N(0,1)$ prior and the right skewed $\exp(-(x-15)), (x \leq 15)$, the corresponding $p^*(\cdot|\theta)$ all have a roughly symmetric form around their locations. We may also investigate how the MIL varies if we fix the mean of a prior and increase its variance; or if we fix the mean and vary the skewness of a prior.

In addition, one must choose how many parameters to include and how to condition on the data (in a posterior). Often, there are several ways one can do this. For instance, in a very simple case one might have an independent sequence of paired data. One can marginalize a bivariate likelihood to get a model for the difference in each pair, as in Model I. 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, as in Models II and III. It is not clear in general when taking differences in the data will be equivalent to taking differences in the parameters so that the two strategies will give compatible inferences. In the usual case, one does not have a plausible likelihood which can be used in either approach. Our method generates one which also permits a sensitivity analysis of the modeling strategy for paired data.

In general, the n-dimensional MIL $p^*(x^n|\theta)$ and the n-fold MIL $\prod_{i=1}^n p^*(x_i|\theta)$ are very different. The former may be used for dependent data. In information theory, θ represents the send message, x^n may be interpreted as the messages received by n receivers, there should be reasonably high dependence among those messages. Whereas $p^*(x^n|\theta)$ is the

channel which permits the slowest data transmission subject to a distortion constraint. The latter may be used for independent data. Intermediate between an n-fold product and an n-fold density we may define the following. Let $\{x_{n_i}^{n_{i+1}}\}$ be a partition of the data x^n, where $x_{n_i}^{n_{i+1}}$ stands for $(x_{n_i}, x_{n_i+1}, ..., x_{n_{i+1}-1})$. If there is dependence among the data in each substring but it is reasonable to model different substrings as independent, then we can model the data by $\prod_i p^*(x_{n_i}^{n_{i+1}}|\theta)$.

If there is reason to believe the data are independent, one should use the product of MILs. If there is obvious dependence structure, for example, pair wise dependence, one may use a bivariate MIL. It is unclear how generally applicable the assumptions of Theorem 3.3.1 are. In case this result is applicable, it is an argument for modeling of dependent data from a sample of size n by an n-dimensional MIL when n is too large. If the result is not applicable, then an n-dimensional MIL may be a candidate model.

In the data analysis, we used the MIL for a location parameter. We can also use it for different types of parameters. For example, for a scale parameter, we may take the loss to be $L(x, \theta) = C(x/\theta)^\alpha$ for some positive constant C and real number α. If one still wishes to use a summary statistic, one may take the sufficient statistic $\sum_{i=1}^n x_i^\alpha$ in this case. For the location-scale parameter (μ, σ), we may take the loss to be $L(x, \mu, \sigma) = (x - \mu)^2/\sigma^2$ or some other suitable alternatives.

In our examples, inferences seems do not appear to be particularly dependent on priors. This robustness may be due in part to the imprecision in specification of the parameter (pre optimization) on which the prior is used. That is, when the optimization procedure identifies the exact interpretation of the parameter, it may also, as a concomitant, reduce the influence of the prior. This is a natural conjecture if the requirement of non-informativity tends to make MIL's more similar than the priors which produced them. In this sense, the SMI may be a contraction mapping, and this contraction may be the main waay minimal informativity is being achieved.

The MIL method seems work well for the initial data analysis in Chapter 4, as it does not require detailed physical modeling. In addition, it makes relatively few assumptions. These assumptions are the inputs of $w(\cdot), L(\cdot, \cdot)$ and the dispersion parameter λ. Our method assigns likelihoods based on minimizing the strength of assumptions, so it is easy to get likelihoods which can be applied to summary statistics. (earlier, we discussed how to handle

the seeming independence of our method from the sample size on which a summary statistic was based, see Section 4.2.1. Partially as a consequence of optimization, we obtain some robustness against choice of prior since we are minimizing the strength of assumption going into the formulation of the likelihood. In addition, robustness against local perturbations of the inputs is relatively straight forward to evaluate.

Frequentist and Bayesian evaluations of robustness usually assume perturbations are local. We do not have to do this here. It is scientifically more important to compare models that close in physical motivation but not close mathematically (i.e., one is not just a perturbed form of the other). This can be examined by evaluating the compatibility of their inferences. For example, Models I, II and III in Chapter 4 are not close in mathematical formulation in any quantified sense, but being formed from pairs and differences they are "close" interpretationally. Classical frequentist robustness results do not appear to handle this case even though it may be of more importance to scientists. Conventional Bayesian robustness does not either. It is only certain model selection techniques that permit this sort of comparison indirectly when they choose the mode of a posterior distribution over a class of models. Here, we are not concerned with model selection so much as with corroboration of inferences by similar yet distinct modeling strategies.

Another potential use of the MIL for statistical problems is that, it provides a reference model for initial data analysis in a Bayesian frame work when little data are available and it is difficult to model them.

There are also potential use of the MIL in information theory context. In fact, we see from Section 2.1.5 that the MIL provides the optimal code for data compression. It is also the channel over which the slowest transmission of the source is achieved.

6.2 Further Topics Regarding The MIL

We see in Section 2.2 that, there is the problem of choosing the Bayes risk bound l; and there are other formulations of the minimization problem, for example use penalty term(s) instead of constraint. In case there is little knowledge for both the likelihood and prior, we may consider a joint optimization procedure for selecting both the likelihood and prior. Also, as we noted that, l and λ determine each other and they have a sort of reciprocal

relationship. In many cases, choose λ is more convenient, and λ behaves like a smoothing parameter. So we can consider some smoothing technique for choosing λ.

In cases without an exact interpretation of θ, it is hard to get a prior distribution of θ. But if we have some vague knowledge about the data distribution, for example its Fisher information $\gamma(\theta)$, then we can form Jeffreys' non-informative prior for the problem, and add the constraint so that the likelihood has the Fisher information $\gamma(\theta)$. Or we may choose an initial prior $w_0(\cdot)$ which can reasonably the parameter, get $p_{MIL,0}$ form this prior, then based on its Fisher information $\gamma_0(\theta)$ get Jeffreys' non-informative prior $w_1(\cdot)$ as the next stage prior, and continue this iteratively we get a minimally informative likelihood-prior pair.

A modification of this search for joint minimal information is sequential. Consider the following procedure. Start with a prior $w_0(\theta)$ which is non-informative and is not derived from a likelihood. (We note that Jeffreys' prior is non-informative but depends on a likelihood through its Fisher information.) For instance, suppose $w_0(\theta)$ is that. From $p_0^*(\cdot|\theta)$ for the first data point x_1. From this get the posterior density $w_1(\theta|x_1) = p_0^*(x_1|\theta)w_0(\theta)/m_0(x_1)$ and use it for fixed x_1 as a prior so that optimization of the SMI gives a likelihood $p_1^*(\cdot|\theta)$, for use on x_2. (Note $p_1^*(\cdot|\theta)$ also depends on x_1 but we have suppressed this.) In this sequential fashion we can develop an adaptive formulation for a joint likelihood by taking a product of these sequential likelihoods. At this point we cannot even conjecture how this procedure will perform. We mention it as a further possibility to explore.

It is also interesting to investigate how much will be lose if we use the MIL, instead of using the true likelihood for inference.

From Example 1.4.3, we see that $E_{p_\lambda^*}(X|\theta)$ is biased for θ because it is a weighted average of μ and θ. We may investigate whether this is the general case, and try method to reduce the bias. In principle there may be cases where it is possible to ensure $E_{p_\lambda^*}(X|\theta) = \theta$ is satisfied. We may also consider this as a constraint in the optimization procedure.

In Chapter 5, we studied the robustness of modeling strategies for paired data. There we investigated the use of general likelihoods, in some cases the special cases for both the dependent and independent MILs, also some cases only for the n-dimensional MILs. We may investigate some special cases only for the independent MILs, since in practice this form may be of wider use than the n-dimensional MILs, although our theoretical work is

mostly for the multivariate MILs.

Also, in Chapter 5, some bounds on the differences for the three modeling strategies are not sharp, so we are not sure how robust the modeling strategies are in such cases. To access it, we may do some numerical comparisons for some practical problems to get some ideas about how these models differ.

Theorem 3.1.1 asserts that under suitable conditions, $I(\Theta, X^n) \to 0$ as $n \to \infty$. That is, Theorem 3.1.1 syas that the MIL for a lot of data must be vacuous given all the dependence structures infinitely much data might have. So, it may be useful to associate to a data string of length n, say $x_1, ..., x_n$, an integer k between 1 and n. This integer k is to be regarded as the information content of $x_1, ..., x_n$ measured in terms of k independent data points. That is, because a sequence of dependent data points behaves like a smaller sequence of independent data points we convert the x^n to k smaller sequences each representing information equivalent to one data point. Now, we have grouped the data set $x_1, ..., x_n$ into k subsets of equal size. Naturally one would want to use several plausible values for k to see how they affect the conclusions. In our example in Chapter 4, we summarized $x_1, ..., x_{57}$ into one summary statistic, essentially thereby regarding the information content of $x_1, ..., x_{57}$ as being equal to that of one data point.

In general cases, the data may rarely be independent, but the generally high dependence in the multivariate MIL may not appropriate. To control the dependence structure in some extent in the multivariate MILs, we may consider using the *copula* to construct multivariate MILs.

Using a copula to construct multivariate model has been popular in recent years. The copula method is an attempt to construct multivariate models with arbitrary given marginals and to some extent, control the dependence structure (see Sklar 1959, Joe, 1993).

Definition: A mapping $C : (0,1)^m \to (0,1)$ is called a *copula* if

(1) it is a continuous distribution function;

(2) each of its univariate marginals is a uniform distribution.

Let $F(\mathbf{x})$ be a multivariate distribution function with marginal distributions $F_1(x_1), ..., F_m(x_m)$, and let $u_1, ..., u_m$ be random variables from $U(0,1)$, then the multivariate distribution $C_F(\cdot, ...)$ defined by

$$C_F(u_1, ..., u_m) = F\Big(F_1^{-1}(u_1), ..., F_m^{-1}(u_m)\Big)$$

is a copula by the above definition.

One of the most commonly used copulas is the multivariate normal copula. It is constructed from the multivariate normal distribution $N(\mathbf{0}, \Gamma)$, where $\Gamma = (\gamma_{ij})$ is its covariance matrix, and for convenience, all the diagonal elements are 1. We denote its distribution function by $\boldsymbol{\Phi}_\Gamma$, its marginal *cdf*s by $\Phi_{\Gamma,1}, ..., \Phi_{\Gamma,m}$. Then the m-dimensional copula is defined as

$$C_{\boldsymbol{\Phi}_\Gamma}(\mathbf{u}|\Gamma) = \boldsymbol{\Phi}_\Gamma\Big(\Phi_{\Gamma,1}^{-1}(u_1), ..., \Phi_{\Gamma,m}^{-1}(u_m)\Big), \qquad \mathbf{u} \in (0,1)^m.$$

Its density function is

$$c_{\boldsymbol{\Phi}_\Gamma}(\mathbf{u}|\Gamma) = |\Gamma|^{-1/2} \exp\Big\{-\frac{1}{2}\mathbf{x}^T(\Gamma^{-1} - I)\mathbf{x}\Big\},$$

where $\mathbf{x} = (x_1, ..., x_m)^T$ with $x_i = \Phi_{\Gamma,i}^{-1}(u_i)$, $i = 1, ..., m$, and I is the identity matrix of dimension m (see Xu, 1996).

In this way, we can construct a multivariate minimally informative distribution $F^*(\cdot, ..., \cdot)$ with given marginals $F_1^*(\cdot), ..., F_m^*(\cdot)$, and partially known dependence structure by means of the m-dimensional normal copula, i.e., if

$$F^*(x_1, ..., x_m|\Gamma) = C_{\boldsymbol{\Phi}}\Big(F_1^{*-1}(x_1), ..., F_m^{*-1}(x_m)\Big),$$

then its marginals are $F_1^*(\cdot), ..., F_m^*(\cdot)$ and its dependence structure can be controlled by the chosen Γ.

We may also consider a non-parametric MIL, by treating the unknown data distribution $F(\cdot)$ as the parameter of interest, and use the Dirichlet process as the prior for distributions. Such method was first proposed by Ferguson (1973) for the non-parametric Bayes method. Let $(R, \mathcal{B})$ be a measurable space, where R is the real line and $\mathcal{B}$ is the σ-algebra of Borel subsets of R. Let $\alpha(\cdot)$ be a finite non-null measure on $(R, \mathcal{B})$. A stochastic process $P(\cdot)$ is a *Dirichlet process* with parameter α and we write $P \in \mathcal{D}(\alpha)$, if for any finite partition $\{B_1, ..., B_m\}$ of R, the random vector $(P(B_1), ..., P(B_m))$ has a Dirichlet distribution with parameter $(\alpha(B_1), ..., \alpha(B_m))$. It has the property that if $F \in \mathcal{D}(\alpha)$, then the posterior of F given $X_1, ..., X_n$ is $\mathcal{D}(\alpha + \sum_{i=1}^n \delta_{X_i})$, where δ_x is the indicator function at x (see Ferguson, 1973). Thus, if one can find the Bayes rule for the no-sample problem ($n = 0$), then the Bayes rule for the n-sample problem is given by replacing α with $\alpha + \sum_{i=1}^n \delta_{X_i}$. For given loss function $L(\cdot, \cdot)$ (for example, we may take it as the L_r norm, the variational distance,

or the Kullback-Leibler divergence, etc.) and $l > 0$, the Bayes risk bound for the no-sample problem is

$$\mathcal{E}L(F, \hat{F}) \leq l,$$

where $F(t) = P((-\infty, t])$ and $\hat{F}$ is chosen to minimize the SMI in this setting subject to the above constraint. Here, we need to formulate the SMI in a meaningful way, so that the optimization is feasible.

References

[1] Akaike, H. (1977). On entropy maximization principle, *Applications of Statistics*, North-Holland Publishing Company.

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

[3] Bernardo, J. M. (1979). Reference posterior distribution for Bayesian inference, *J. Roy. Statist. Soc., Ser. B*, No.2, 113-147.

[4] Barron, A. R. and Cover, T. M. (1989). Minimum Complexity Density Estimation, Technical Report ♯ 28, Department of Statistics, University of Illinois.

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

[6] Blahut, R. E. (1972b) An hypothesis testing approach to information theory, *Ph.D. Thesis, Cornell Univ.*.

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

[8] Clarke, B. S. and Barron, A. R (1990). Information-Theoretic Asymptotics of Bayes Methods, *IEEE Trans. Inform. Theory*, vol. 36, no. 3, pp. 453-471.

[9] Clarke, B. S. and Barron, A. R (1994). Jeffreys' prior is asymptotically least favorable under entropy risk, *J. Statist. Planning and Inference*, vol. 41, pp. 37-60.

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

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

[12] Csiszar, I. (1975). I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 1975, Vol.3, No.1, 146-158.

[13] Farnum, N. R. and Stanton, L. W. (1989). *Quantitative forecasting methods*, PWS-KENT.

[14] Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* Vol. 1, No. 2, 209-230.

[15] Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Math. Statist.* **81**, 1208-1212.

[16] Haughton, M. A. (1988). On the choice of a model to fit data from an exponential family. *Ann. Statist*, Vol. 16, No. 1, 342- 355.

[17] Jaynes, E. T. (1957). Information Theory and Statistical Mechanics", *Physical Review*, 106, 620-630.

[18] Joe, Harry. (1989). Relative Entropy Measures of Multivariate Dependence. *J. Amer. Statist. Assoc.*

[19] Joe, Harry. (1993). Parametric family of multivariate distribution with given margins. *J. Mult. Anal.* **46**, 262-282.

[20] Jørgensen, B. and Labouriau, R. S. (1994). Exponential families and theoretical inference. *private communication.*

[21] Kolmogorov, A. N. (1965). Three Approaches to the Quantitative Definition of Information, *Problemy Peredachi Informatsii*, Vol. 1, pp.3-11.

[22] Kondo, J. (1991). *Integral Equations*, Oxford Univ. Press.

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

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

[25] Schwartz, G. (1978). Estimating the dimension of a model. *Ann. Statist*, Vol. 6, No. 2, 461-464.

[26] Sklar, A. (1959). Fonctions de répartition á n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, **8**, 229-231.

[27] Soofi, E. S. (1994). Capturing the intangible concept of information. *J. Amer. Statist. Assoc.* Vol. 89, No. 428, 1243-1254.

[28] Strasser, H. (1981). Consistency of maximum likelihood and Bayes estimates, *Ann. Statist.* Vol. 9, No. 5, 1107-1113.

[29] Wald, A. (1950). *Statistical Decision Functions*, Wiley.

[30] Walker, A. M. (1967). On the asymptotic behavior of posterior distributions, *Journal of the Royal Statistical Society*, Series B (31), 80-88.

[31] Xu, J. J. (1996). Ph.D. Thesis, Department of Statistics, University of British Columbia, Canada.