1984

Applications of the distance measures between the prior and posterior distributions

Cindy Lynn Martin
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/rtd

Part of the Statistics and Probability Commons

Recommended Citation
Martin, Cindy Lynn, "Applications of the distance measures between the prior and posterior distributions " (1984). Retrospective Theses and Dissertations. 8189.
https://lib.dr.iastate.edu/rtd/8189

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Retrospective Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
INFORMATION TO USERS

This reproduction was made from a copy of a document sent to us for microfilming. While the most advanced technology has been used to photograph and reproduce this document, the quality of the reproduction is heavily dependent upon the quality of the material submitted.

The following explanation of techniques is provided to help clarify markings or notations which may appear on this reproduction.

1. The sign or “target” for pages apparently lacking from the document photographed is “Missing Page(s)”. If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting through an image and duplicating adjacent pages to assure complete continuity.

2. When an image on the film is obliterated with a round black mark, it is an indication of either blurred copy because of movement during exposure, duplicate copy, or copyrighted materials that should not have been filmed. For blurred pages, a good image of the page can be found in the adjacent frame. If copyrighted materials were deleted, a target note will appear listing the pages in the adjacent frame.

3. When a map, drawing or chart, etc., is part of the material being photographed, a definite method of “sectioning” the material has been followed. It is customary to begin filming at the upper left hand corner of a large sheet and to continue from left to right in equal sections with small overlaps. If necessary, sectioning is continued again—beginning below the first row and continuing on until complete.

4. For illustrations that cannot be satisfactorily reproduced by xerographic means, photographic prints can be purchased at additional cost and inserted into your xerographic copy. These prints are available upon request from the Dissertations Customer Services Department.

5. Some pages in any document may have indistinct print. In all cases the best available copy has been filmed.
Martin, Cindy Lynn

APPLICATIONS OF THE DISTANCE MEASURES BETWEEN THE PRIOR AND POSTERIOR DISTRIBUTIONS

Iowa State University

University Microfilms International 300 N. Zeeb Road, Ann Arbor, MI 48106
Applications of the distance measures between
the prior and posterior distributions

by

Cindy Lynn Martin

A Dissertation Submitted to the
Graduate Faculty in Partial Fulfillment of the
Requirements for the Degree of
DOCTOR OF PHILOSOPHY

Major: Statistics

Approved:

Signature was redacted for privacy.

In Charge of Major Work

Signature was redacted for privacy.

For the Major Department

Signature was redacted for privacy.

For the Graduate College

Iowa State University
Ames, Iowa

1984
TABLE OF CONTENTS

1. INTRODUCTION 1

2. A MEASURE OF CONSONANCE BETWEEN THE PRIOR AND POSTERIOR DISTRIBUTIONS 13
   2.1. Introduction 13
   2.2. Distance Calculations 16
   2.3. Robust Considerations 24

3. THE USE OF THE DISTANCE MEASURE IN THE DETECTION OF OUTLIERS 35
   3.1. Introduction 35
   3.2. Applications to the Linear Model 39
   3.3. Modeling Outliers: An Open Topic 60

4. A MEASURE OF INFORMATION 67
   4.1. Introduction 67
   4.2. A New Measure of Information 74

5. A PREDICTION PROBLEM 84

6. REFERENCES 100

7. ACKNOWLEDGMENTS 102
1

1. INTRODUCTION

In any statistical investigation, an analyst will eventually find herself faced with the task of comparing and contrasting all available sources of information in order to make a decision or inference concerning some unknown state of nature. For example, suppose one wanted to investigate a parameter of interest, say \( \theta \), and an experiment was conducted which generated a sample \( x \). For a Bayesian statistician, the task begins by identifying the conditional distribution of \( x \) given \( \theta \), \( f(x|\theta) \), and the prior distribution of \( \theta \), \( g(\theta) \). Then, once the sample \( x \) has been observed we calculate the posterior distribution of \( \theta \) given \( x \),

\[
p(\theta|x) = \frac{f(x|\theta)g(\theta)}{\int f(x|\theta)g(\theta)d\theta}.
\]

Thus, updating prior beliefs by incorporating the information from the sample.

Given this posterior distribution, depending on the particular loss function assumed, and the specific estimation problem being addressed, we have all the tools needed to make a decision or inference. But, the question to be considered here is, what if these varying sources of information seem to be incongruous with respect to one another. Specifically, what if after observing \( x \), we calculate a posterior distribution for \( \theta \) which is very different from the prior distribution for \( \theta \)? Does this mean our prior beliefs were
incorrect, or incomplete in some sense, or did we just happen to observe a surprising \( x \)? And, what does it mean for our posterior distribution to be "different" from our prior distribution?

To begin then, notice that the statistician has at hand three sources of information from which to draw an inference, the prior, the likelihood function, and the observed \( x \). Throughout this thesis, we assume that the likelihood function is correct. This reduces the problem to one of comparing and contrasting the information in the prior with the information in the \( x \). Since the posterior distribution for \( \theta \) reflects the influence of the sample on our prior beliefs, it would be informative to contrast this distribution with the prior density.

This will be done by calculating the distance between the prior and posterior distributions for \( \theta \). The basic idea being that if the posterior distribution for \( \theta \) seems very different, in terms of say shape or location, from the prior distribution, then some further investigation would seem to be called for. There are two avenues of inquiry we could pursue. First, we consider the possibility of a misspecified prior, an idea suggested by what is known as the "Robust Bayesian Viewpoint," see Berger (1982). Or, if we are satisfied with the prior density, perhaps we have observed a "surprising" data set, or maybe our data contains outliers. Antithetically, if the two distributions seem quite consonant, we would feel justified in making our inferences according to standard Bayesian practices.
There are a number of functions available which measure the distance between two density functions. The two that will be used here are the squared Hellinger's distance, see Eaton (1981), i.e.

\[ d^2(P(\theta|x), g(\theta)) = \int \left[ (P(\theta|x))^{\frac{1}{2}} - (g(\theta))^{\frac{1}{2}} \right]^2 d\theta, \]

and the Kullback-Liebler directed divergence,

\[ d(P(\theta|x), g(\theta)) = \int P(\theta|x) \log \frac{P(\theta|x)}{g(\theta)} d\theta. \]

Note, the Kullback-Liebler measure is technically considered a directed divergence since it lacks the property of symmetry.

In Chapter 1 then, we begin by considering when the prior is consonant with the posterior, and what a lack of such consonance would imply. We define the measure of consonance to be the distance between the prior and posterior distributions of \( \theta \). Distance calculations will be presented for the normal distribution, the general exponential family, and for those distributions where the necessary integration was tractable. The problem of how to interpret the resulting measure will then be discussed.

This idea of questioning the consonance of the prior with the observed sample, before making any inferences, is very much in the spirit of what Berger (1982) calls the "robust Bayesian" viewpoint. The outlook here is that a Bayesian statistician should try to behave in such a way that her analysis is robust with respect to all those
priors which seem reasonable after the prior elicitation process has been completed. Because in reality, it is often quite difficult to specify the exact prior distribution of the parameters. The analyst is usually able to identify qualities of the prior such as its percentiles, or maybe certain shape features, but values such as moments are often too difficult to accurately specify. The basic reason for this is that one does not usually have very detailed information concerning the "tail" of the prior distribution.

Because of this inability to identify the true prior, call it \( \pi_t \), the statistician may have to settle for her best approximation to the prior, say \( \pi_A \). The robust Bayesian could then proceed to define a class of priors, \( \Gamma \), which are "close" in some sense to \( \pi_A \). An example of one commonly used class is \( \Gamma = \{ \pi : \pi(\cdot) = (1-\epsilon)\pi_A + \epsilon P(\cdot) \} \), where \( P \) is an arbitrary probability distribution, and \( \epsilon \) relates to how strongly we believe our prior approximation.

Then, what would it mean in this case for a Bayesian to behave in a robust manner?

The most general definition states that any inference or decision is posterior robust with respect to \( \Gamma \) if it is satisfactory with respect to \( \pi(\cdot|x) \) for all \( \pi \in \Gamma \). This definition is intentionally vague, leaving the door open as to how to interpret "satisfactory with respect to." As an example, in a decision-theoretic situation, an action \( a_0 \) is called \( \epsilon \)-posterior robust with respect to \( \Gamma \), for a sample \( x \), if
\[ \sup_{\pi \in \Gamma} |\rho(\pi, x, a_o) - \inf_{a \in A} \rho(\pi, x, a)| \leq \epsilon. \]

Here \( \rho(\pi, x, a) \) is the Bayesian posterior risk for the loss function \( L(\theta, a) \), i.e. \( \rho(\pi, x, a) = \int L(\theta, a) \pi(d\theta | x) \). Now, note that this measure depends on the particular \( x \) that we observe. So while one \( x \) may suggest that your action is posterior robust, for another \( x \) it may not be.

An alternative definition for robustness involves looking at the Bayes risk, \( r(\pi, \delta) \), which is the expected value of the posterior risk with respect to the marginal distribution of \( X \), i.e. \( r(\pi, \delta) = \int \rho(\pi, x, \delta(x)) m(dx) \). In this case, the procedure \( \delta_o \) is called \( \Gamma \)-procedure robust with respect to \( \Gamma \) if

\[ \sup_{\pi \in \Gamma} [r(\pi, \delta_o) - \inf_{\delta} r(\pi, \delta)] < \epsilon. \]

It is useful to note that you only need to consider procedure robustness when posterior robustness is lacking. This has a tendency to occur when the likelihood function is concentrated in some region of the parameter space for which you have incomplete information, for example, the tails.

Recall then, that the essential element in any Bayesian analysis is the posterior distribution. For as Berger (1982) states when referring to the information available about an uncertainty after we have seen the data, "the only trustworthy and sensible measures of this information are Bayesian posterior measures." Thus, given a class of
priors $\Gamma$, the logical class of posterior distributions would be

$$\Gamma^* = \{ \pi(\cdot | x) : \pi \in \Gamma \}.$$ 

Therefore, one idea we might consider, is to calculate the distances between the priors and posteriors in $\Gamma$ and $\Gamma^*$ respectively, for a fixed $x$. Unfortunately though, we can not make any absolute comparisons between these distances, since the distribution of the measure depends upon the prior used. Therefore, a solution to this could be to calculate the relative probability of observing our distance, or one more extreme, for each prior and posterior. This then, would give us some basis by which to compare the resulting measures, i.e. a larger relative probability for our measure of consonance would indicate a more robust prior.

Another approach to this problem of disconsonance, suggested by Box (1980), involves consideration of the marginal distribution of $x$. The idea is that a particularly small value of $h(x) = \int f(x|\theta)g(\theta)d\theta$ indicates that "surprising" data has occurred. The emphasis here is more on the notion that we have observed an unexpected $x$, as opposed to having misspecified the prior. He, therefore, suggests comparing the value of the marginal density at the observed $x$, call this $h(x^*)$, to the density $h(x)$, i.e. evaluate $\alpha = P(h(x) < h(x^*))$. Therefore, if for our sample $x$, we calculate a small $\alpha$ level, then we would want to reconsider our model before making any further inferences. Note, this can be viewed as a sort of Bayesian analogue to the classical
significance test.

An alternative to the above idea, which does not involve averaging over sample values which did not occur, would be to measure the distance between the prior and posterior distributions for \( \theta \), at each \( x \) in the sample space. Assuming of course that the sample space is finite. This would give us a relative ordering of the sample points in terms of the degree to which they are consonant with our prior beliefs. Therefore, in Chapter 2, an example of this method will be given and comparisons made with Box's significance level technique.

In Chapter 3, we will investigate how the distance measure can be applied to the study of outliers. Throughout this chapter, we will be concerned with making judgements about observations from our sample \( \mathbf{x} \). The likelihood function and the prior distribution will be assumed to be specified to our satisfaction. Here we will consider what we call a deleted posterior, i.e. the posterior distribution of \( \theta \) when we remove a subset \( (i) \) of the observations. This will be denoted by \( P_{(i)}(\theta|x) \). Note, this is the prior distribution for \( \theta \) given that we have observed all values except the set \( x_{(i)} \). Then, distances will be calculated between the full posterior, i.e. the one including all the sample observations, and each deleted posterior distribution. The idea here is that a large value for this distance would imply that this subset is very influential. This technique for identification of outliers will then be applied to the regression problem, where the data is a sample from a general linear model.
Johnson and Geisser (1983) also considered using a distance measure for the detection of outliers in a general linear model. Their approach was to calculate the Kullback-Liebler divergences between the marginal, or "predictive" densities derived from the full and deleted data sets. They then defined these divergences to be "predictive influence functions," and used them to obtain an ordering of the subsets according to the magnitude of these divergences. Although this idea seems well suited to the prediction problem, the use of the marginal density, which averages over the sample space of $X$, may not be the best approach.

Instead then, we consider the posterior conditional distributions of $\beta$, i.e. given the entire data set and the subset deleted data set. The distances, both Hellinger's and Kullback-Liebler, will be calculated for the conjugate multivariate normal-gamma prior and the noninformative uniform prior. These distances will be shown to be functions of a generalized measure of the influence a specific subset has on the estimation of $\beta$ first introduced by Cook (1977). Note here, that if by using this technique we decide that a specific subset is suspected of being outliers, we would then apply this information to our analysis when considering the prediction, or any other general estimation problem.

In the last section of Chapter 3, we consider the case where the possibility of outliers is modeled into our analysis. This idea was formally investigated for the general linear model by Box and Tiao
(1968). There, they were able to show that the posterior distribution for $\beta$ could be expressed as a weighted average of $2^n$ multivariate $t$ distributions, where each density represented the posterior distribution of $\beta$ given that a particular set of the observations were outliers. The weights here represent the probabilities of those particular observations being outliers. We will discuss how this idea of modeling outliers could be related to our distance measure, although the exact relationship remains an open question.

Up until this point, we have discussed ways of comparing and contrasting the different sources of information available to the statistician. In Chapter 4, we formally address the question of how to measure the amount of information given by an experiment $E$, which yields an observed sample value $x$.

The idea of attempting to quantify information is important, for it serves to establish a general rule of experimentation, i.e. the statistician should perform that experiment for which the expected gain in information is maximum. This idea is also crucial because it ties in with what Good (1971) calls the "explanatory power" of the data. That is, he showed that maximizing the expected explanatory power is equivalent to maximizing the expected amount of information given by an experiment.

Basu (1975) presented various guidelines, in the form of four key principles, which should be followed when endeavoring to define a
measure of information. These will be reviewed and discussed in the introduction to Chapter 4. There we will also consider a standard measure satisfying these principles that was introduced by Lindley (1956).

Specifically, Lindley's measure entails comparing the amount of information we have after the experiment, with the amount we had before. This is accomplished by contrasting the degree of concentration in the posterior density with the degree of concentration in the prior.

One problem we found with this measure though, was that since it focused in on the concentration of the densities, it often reduced to a function of the various variance components. Thus, if the variances were not a function of the observed \( x \), as in the case of a normal density with a conjugate normal prior, the information measured by Lindley for a particular \( x \), will not be a function of this \( x \).

Therefore, in Chapter 4, we introduce a new measure of statistical information. This measure seeks to combine the information in the posterior density with the information in the observed \( x \), as represented by the likelihood function, by means of a convex combination. The weight given to each source will depend upon an appropriate form of our previously introduced measure of consonance.

The idea here is that if the \( x \) we observe is consonant with our prior beliefs, our information measure should be primarily a function of the information in the posterior density, as represented by
its concentration. Whereas, if the \( x \) seems to be in conflict with the prior density, then we would want to downplay the posterior information and concentrate on the information contained in the likelihood function alone.

Thus, we have derived a measure satisfying Basu's principles which takes into account the concentration of the densities involved while remaining sensitive to the overall consonance of the observed value.

In the final chapter, we will look at a rather interesting application of this distance notion. Here we consider the standard regression problem of deciding how many parameters to include when we want to predict a future \( y \) value. The thought here was to calculate the predictive density of a future value where we include all the available parameters, and then to calculate the density where we utilize only a subset of the parameters. Then, we measure the distance between the two predictive densities. This gives us a method by which to compare, for a fixed subset of variables, which set would give us a predictive density closest to that of the full model.

Then, if we associate a suitably scaled cost with each variable we consider including, we can establish a mechanism by which to decide which model would be "best" to use in predicting a future \( y \).

In the case where we use the uniform, noninformative prior, the distance measures can be shown to be increasing functions of the
reduced sums of squares, i.e. the classical measure used in deciding the significance of subsets of independent variables.
2. A MEASURE OF CONSONANCE BETWEEN THE PRIOR AND POSTERIOR DISTRIBUTIONS

2.1. Introduction

In this chapter, we will introduce the idea of distance as a measure of consonance between the prior and posterior beliefs about $\theta$, because in practice most Bayesian techniques involve combining the information from our prior and our sample in order to make some sort of decision or inference. For example, in the case of estimating an unknown parameter $\theta$, under squared error loss, most Bayes' estimates turn out to be a convex combination of our prior belief about $\theta$ and the estimate that we get from our sample alone. But if these two quantities seem quite different from one another, are we really justified in taking this convex combination? Or should we perhaps further investigate our problem in order to find out the cause of the discordance?

To begin then, we will define the distance between the prior distribution for $\theta$, $g(\theta)$, and the posterior distribution for $\theta$ given $x$, $P(\theta|x)$, to be our measure of consonance for the model we are considering. Recall that throughout this thesis we will be working under the assumption that our conditional distribution for $x$ given $\theta$, $f(x|\theta)$, is not being questioned. Now, as previously mentioned, there are two measures of distance that will be used. The first is the Squared Hellinger's distance, which for a particular $x$ will be denoted by $d_1(x)$, where
\[ d_1(x) = d^2(P(\theta|x), g(\theta)) = \int \left[ \left( P(\theta|x) \right)^{\frac{1}{2}} - \left( g(\theta) \right)^{\frac{1}{2}} \right]^2 d\theta. \]

We will also use the Kullback-Liebler directed divergence, \( d_2(x) \), where
\[ d_2(x) = d(P(\theta|x), g(\theta)) = \int P(\theta|x) \log \frac{P(\theta|x)}{g(\theta)} d\theta. \]

Note, we could likewise consider \( d(g(\theta), P(\theta|x)) \), which would be analogously defined. Or, we could look at the sum of these two directed divergences, which is known as the divergence between the two densities. But for our purposes, it is usually sufficient to consider just one of the directed divergences. We will thus use the aforementioned \( d_2(x) \). As an aside, it is useful to note that it was shown by J. Hannan (1960) that \( d_2(x) \geq d_1(x) \) when \( P(\theta|x) \) and \( g(\theta) \) are both densities with respect to a dominating measure \( \mu \).

Now, before we calculate any of these distances, it will be useful to adjust both measures so their range lies between 0 and 1. This will aid us in our assessment of what should be considered a large distance, and what should not. The implication here being that a distance close to one means there is quite a difference between the prior and posterior distributions for \( \theta \), in terms of say location or spread. While a distance close to 0 would mean that there was quite a high degree of consonance between the two densities.

Note then, that our squared Hellinger's distance can be written as
\[
\begin{align*}
  d_1(x) &= \int [(P(\theta|x))^\frac{1}{2} - (g(\theta))^\frac{1}{2}]^2 d\theta \\
  &= \int [P(\theta|x) + g(\theta) - 2(P(\theta|x)g(\theta))^\frac{1}{2}] d\theta \\
  &= 2\left[1 - \int (P(\theta|x)g(\theta))^\frac{1}{2} d\theta\right].
\end{align*}
\]

Therefore, we will define the scaled Hellinger's distance to be

\[
d^*(x) = d_1(x) / 2 = 1 - \int (P(\theta|x)g(\theta))^\frac{1}{2} d\theta.
\]

Next, recall that our Kullback-Liebler directed divergence

\[
d_2(x) = \int P(\theta|x) \log \frac{P(\theta|x)}{g(\theta)}. \quad \text{Therefore, for any two densities,}
\]

\[
d_2(x) \text{ has a range of 0 to infinity. So let}
\]

\[
\tilde{d}(x) = \frac{d_2(x)}{d_2(x) + 1}.
\]

This will again insure that the measure falls between 0 and 1.

It should be noted here that the only time that either measure
will equal zero is when the posterior and prior distributions for \( \theta \)
are identical. Now, since \( P(\theta|x) = [f(x|\theta)g(\theta)]/h(x) \), where
\( h(x) = \int f(x|\theta)g(\theta)d\theta \), the only time this can happen is when
\( f(x|\theta) = h(x) \), for all \( \theta \). This would mean that \( x \) and \( \theta \) were
independent, in which case observing our \( x \) would give us no
information about \( \theta \). If this extreme case did arise, then we should
certainly go back and reexamine the problem we were considering to
see what could be done to correct for this situation.
2.2. Distance Calculations

First, we will derive the general form of the distance between two univariate normal distributions. Suppose then that $P_1(\theta)$ is a normal distribution with mean $\mu_1$ and variance $\sigma_1^2$, i.e.,

$P_1(\theta) \sim N(\mu_1, \sigma_1^2)$, and let $P_2(\theta) \sim N(\mu_2, \sigma_2^2)$. Then,

$P_1(\theta) = \frac{1}{\sqrt{2\pi \sigma_1^2}} \exp - \frac{1}{2\sigma_1^2} (\theta - \mu_1)^2$ and $P_2(\theta) = \frac{1}{\sqrt{2\pi \sigma_2^2}} \exp - \frac{1}{2\sigma_2^2} (\theta - \mu_2)^2$.

Therefore, referring to (2.1.1) we need to first evaluate

$\frac{1}{2\pi} \int [P_1(\theta)P_2(\theta)]^\frac{1}{2} d\theta$. Now,

$\int [P_1(\theta)P_2(\theta)]^\frac{1}{2} d\theta = \int \left( \frac{1}{2\pi} \frac{1}{\sqrt{\sigma_1^2 \sigma_2^2}} \exp - \frac{1}{2 \left( \frac{\sigma_1^2}{\sigma_2^2} \right)} (\theta - \mu_1)^2 - \frac{1}{2 \left( \frac{\sigma_2^2}{\sigma_1^2} \right)} (\theta - \mu_2)^2 \right)^\frac{1}{2} d\theta$

$= \frac{1}{\sqrt{2\pi}} \left( \frac{1}{\sigma_1^2 \sigma_2^2} \right)^\frac{1}{4} \int \exp - \frac{1}{4 \left( \frac{\sigma_1^2}{\sigma_2^2} \right)} (\theta - \mu_1)^2 - \frac{1}{4 \left( \frac{\sigma_2^2}{\sigma_1^2} \right)} (\theta - \mu_2)^2 d\theta$

$= \frac{1}{\sqrt{2\pi}} \left( \frac{1}{\sigma_1^2 \sigma_2^2} \right)^\frac{1}{4} \int \exp - \frac{1}{4 \left( \frac{\sigma_1^2}{\sigma_2^2} \right)} (\theta - \mu_1)^2 - \frac{1}{4 \left( \frac{\sigma_2^2}{\sigma_1^2} \right)} (\theta - \mu_2)^2 d\theta$

$= \frac{1}{\sqrt{2\pi}} \left( \frac{1}{\sigma_1^2 \sigma_2^2} \right)^\frac{1}{4} \exp - \frac{1}{4 \left( \frac{\sigma_1^2}{\sigma_2^2} \right)} (\theta - \mu_1)^2 - \frac{1}{4 \left( \frac{\sigma_2^2}{\sigma_1^2} \right)} (\theta - \mu_2)^2 d\theta$
\[
\begin{align*}
&= \left( \frac{1}{\sigma_1^2} \right)^{\frac{1}{2}} \exp - \frac{1}{4} \left( \frac{\mu_1^2 + \mu_2^2}{\sigma_1^2 + \sigma_2^2} \right) \int \frac{1}{\sqrt{2\pi}} \exp - \frac{1}{4} \left( \frac{\sigma_1^2 + \sigma_2^2}{\sigma_1^2 + \sigma_2^2} \right) \\
&\quad \left( \theta^2 - 2\theta \left( \frac{\mu_1^2 + \mu_2^2}{\sigma_1^2 + \sigma_2^2} \right) + \left( \frac{\mu_1^2 + \mu_2^2}{\sigma_1^2 + \sigma_2^2} \right)^2 \right) d\theta \\
&= \left( \frac{1}{\sigma_1^2} \right)^{\frac{1}{2}} \left( \frac{2\sigma_1^2}{\sigma_1^2 + \sigma_2^2} \right)^{\frac{1}{2}} \exp - \frac{1}{4} \left[ \frac{\mu_1^2 + \mu_2^2}{\sigma_1^2 + \sigma_2^2} \right] - \frac{\left( \frac{\mu_1^2 + \mu_2^2}{\sigma_1^2 + \sigma_2^2} \right)^2}{\left( \sigma_1^2 + \sigma_2^2 \right)^2} \\
&\quad \times \int \frac{1}{\sqrt{2\pi}} \exp - \frac{1}{2} \left( \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} \right) \left( \theta - \frac{\mu_1^2 + \mu_2^2}{\sigma_1^2 + \sigma_2^2} \right)^2 d\theta \\
&= \sqrt{2} \left( \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} \right)^{\frac{1}{2}} \exp - \frac{1}{4} \left[ \frac{\mu_1^2 + \mu_2^2}{\sigma_1^2 + \sigma_2^2} \right].
\end{align*}
\]

Therefore,
\[
d^*(x) = 1 - \frac{\sqrt{2} \left( \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} \right)^{\frac{1}{2}}}{\sqrt{\sigma_1^2 + \sigma_2^2}} \exp - \frac{1}{4} \left( \frac{\mu_1^2 - \mu_2^2}{\sigma_1^2 + \sigma_2^2} \right). \tag{2.2.1}
\]

In order to calculate \( d(x) \) we need to first simplify the ratio \( P_1(\theta)/P_2(\theta) \). So in this case,
\[ \frac{P_1(\theta)}{P_2(\theta)} = \frac{\sigma_1^2}{\sigma_2^2} \exp \left\{ -\frac{1}{2} \left( \frac{(\theta - \mu_1)^2}{\sigma_1^2} - \frac{(\theta - \mu_2)^2}{\sigma_2^2} \right) \right\} \]

\[ = \frac{\sigma_1^2}{\sigma_2^2} \exp \left\{ -\frac{1}{2} \left( \theta^2 \left( \frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2} \right) - 2 \theta \left( \frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2} \right) + \left( \frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2} \right) \right) \right\}. \]

Therefore,

\[ \int P_1(\theta) \log \frac{P_1(\theta)}{P_2(\theta)} d\theta = \]

\[ \int \left( \frac{1}{\sqrt{2\pi \sigma_1^2}} \right) \exp \left\{ -\frac{1}{2} \left( \theta - \mu_1 \right)^2 \right\} \frac{1}{2} \ln \left( \frac{\sigma_2^2}{\sigma_1^2} \right) - \frac{\theta^2}{2} \left( \frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2} \right) \]

\[ + \theta \left( \frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2} \right) - \frac{1}{2} \left( \frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2} \right) \] \, d\theta

\[ = \frac{1}{2} \ln \left( \frac{\sigma_2^2}{\sigma_1^2} \right) - \frac{1}{2} \left( \frac{\sigma_1^2}{\sigma_2^2} \right) \left( \frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2} \right) + \mu_1 \left( \frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2} \right) - \frac{1}{2} \frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2} \]

\[ = \frac{1}{2} \left[ \ln \left( \frac{\sigma_2^2}{\sigma_1^2} \right) + \frac{(\mu_1 - \mu_2)^2}{\sigma_2^2} + \frac{\sigma_1^2}{\sigma_2^2} - 1 \right]. \quad (2.2.2) \]

Thus,
\[
\tilde{d}(x) = \frac{\frac{\sigma_1^2}{\ln(-f)} + \frac{(\frac{\mu_1 - \mu_2}{\sigma_2})^2}{\sigma_2^2} + \frac{\sigma_1^2}{\sigma_2^2} \cdot 1}{\ln(\frac{\sigma_1^2}{\ln(-f)} + \frac{(\frac{\mu_1 - \mu_2}{\sigma_2})^2}{\sigma_2^2} + \frac{\sigma_1^2}{\sigma_2^2} + 1)}.
\] (2.2.3)

Now, suppose we are interested in estimating the mean of a normal random variable \(X\) which has known variance \(\sigma^2\). Suppose also in this case that we decide to use the conjugate prior for \(\theta\), i.e. \(g(\theta)\) is the normal density with mean \(\mu\) and variance \(\tau^2\), then the posterior distribution of \(\theta\) given \(x\) will also be normally distributed. Specifically, \(P(\theta|x) \sim N\left(\frac{\tau^2 x + \mu \sigma^2}{\tau^2 + \sigma^2}, \frac{\sigma^2}{\tau^2 + \sigma^2}\right)\). In this case, using (2.2.1), with \(\mu_1 = \frac{\tau^2 x + \mu \sigma^2}{\tau^2 + \sigma^2}, \sigma_1^2 = \frac{\sigma^2}{\tau^2 + \sigma^2}, \mu_2 = \mu,\) and \(\sigma_2^2 = \tau^2\), we find that the scaled Hellinger's distance between our prior and posterior distributions for \(\theta\) is

\[
d^*(x) = 1 - \frac{\sqrt{2 \tau^2 (\sigma^4 + \sigma^2 \tau^2)^4}}{\sqrt{2 \tau^2 (\sigma^2 + \tau^2)^4}} \exp -\frac{1}{4} \left\{ \frac{\tau^4 (\mu - x)^2}{(\tau^2 + \sigma^2)(\tau^2 + \sigma^2)^4} \right\}. \] (2.2.4)

And utilizing (2.2.3), it can be shown that

\[
\tilde{d}(x) = \frac{\ln(1 + \frac{\tau^2}{\sigma^2}) + \frac{\tau^2}{\sigma^2} (\mu - x)^2 + \frac{\sigma^2}{\tau^2 + \sigma^2} - 1}{\ln(1 + \frac{\tau^2}{\sigma^2}) + \frac{\tau^2}{\sigma^2} (\mu - x)^2 + \frac{\sigma^2}{\tau^2 + \sigma^2} + 1}.
\] (2.2.5)
Notice here that in both cases we see that the distance will be minimum when the observed \( x \) is equal to the prior mean for \( \theta \). And, that the distances are monotonically increasing functions of \(|x-\mu|\).

This is important since in the Bayesian framework, with squared error loss, the estimate for \( \theta \) would be the posterior mean of \( \theta \), i.e.

\[
(\frac{-\tau}{\tau + \sigma^2})x + (\frac{\sigma^2}{\tau + \sigma^2})\mu.
\]

This then is an example where the estimate is a convex combination of the prior mean for \( \theta \) and the observed value. Therefore, a large difference between these two numbers would perhaps shed some doubt upon the estimate.

Now, suppose we have available to us a sample \( x = (x_1, \ldots, x_n) \) where each \( X_i \) given \( \theta \) is normally distributed with mean \( \theta \) and variance \( \sigma^2 \). If we agree to again use the conjugate prior, then 

\[ P(\theta|x) \] will be normally distributed with mean \( \frac{n\tau \bar{x} + \sigma^2 \mu}{n\tau^2 + \sigma^2} \), and variance \( \frac{\sigma^2 \tau^2}{\sigma^2 + n\tau^2} \). This time we find that

\[ d^*(x) = 1 - \frac{\sqrt{2\pi (\sigma^4 + n\tau^2 \sigma^2)^4}}{\sqrt{\pi^2 \tau^2 + n\tau^4}} \exp \left[ -\frac{1}{4} \frac{n^2 \tau^4 (\mu - \bar{x})^2}{(n\tau^2 + \sigma^2)(2\tau^2 + n\tau^4)} \right]. \quad (2.2.6) \]

and

\[ \gamma(x) = \frac{\ln(1 + \frac{n\tau^2}{\sigma^2}) + \frac{n^2 \tau^2}{(n\tau^2 + \sigma^2)^2} (\mu - \bar{x})^2 + \frac{\sigma^2}{\sigma^2 + n\tau^2} - 1}{\ln(1 + \frac{n\tau^2}{\sigma^2}) + \frac{n^2 \tau^2}{(n\tau^2 + \sigma^2)^2} (\mu - \bar{x})^2 + \frac{\sigma^2}{\sigma^2 + n\tau^2} + 1}. \quad (2.2.7) \]
Notice here that analogous to the previous case, both distances will assume their minimum when the classical estimate for $\theta$, i.e. $\bar{x}$, is equal to the prior mean. And, the distances are again monotonically increasing functions of $|\mu - \bar{x}|$.

Now, for other than the case when we have two normal distributions, the calculations for the Kullback-Liebler divergences become quite unmanageable. There are some approximations available in certain cases, but these would not really be useful for our purposes. Therefore, we will give only the Hellinger's distances for the instances where our likelihood function is binomial, Poisson, and exponential. In all cases, the corresponding conjugate prior will be used.

First then, suppose that the conditional distribution of $x$ given $\theta$ is the binomial density with parameters $n$ and $\theta$, i.e. $f(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$. Then the conjugate prior would be the beta distribution with parameters $\alpha$ and $\beta$, that is, $g(\theta) = \theta^{\alpha-1} (1-\theta)^{\beta-1} / B(\alpha, \beta)$. In this case it is easy to show that the posterior distribution for $\theta$ given $x$ will again follow a beta distribution, with parameters $\alpha+x$ and $\beta+n-x$. Therefore, to calculate $d^*(x)$ we must first look at the expression $\int (P(\theta|x)g(\theta))^{1/2} d\theta$. Now,

$$\int (P(\theta|x)g(\theta))^{1/2} d\theta = \int \left[ \frac{\theta^{\alpha-1} (1-\theta)^{\beta-1}}{B(\alpha, \beta)} \cdot \frac{\theta^{\alpha+x-1} (1-\theta)^{\beta+n-x-1}}{B(\alpha+x, \beta+n-x)} \right]^{1/2} d\theta$$

$$= \frac{1}{\sqrt{B(\alpha, \beta) B(\alpha+x, \beta+n-x)}} \int \theta^{2\alpha+x-2} (1-\theta)^{2\beta+n-x-2} \frac{1}{2} d\theta$$
Therefore,
\[
\frac{B(\alpha + x/2, \beta + n-x/2)}{\sqrt{B(\alpha, \beta)B(\alpha+x, \beta+n-x)}}.
\]

If our conditional distribution of \( x \) given \( \theta \) is the Poisson distribution, i.e. \( f(x|\theta) = \frac{\theta^x e^{-\theta}}{x!} \), then the conjugate prior would be the gamma distribution with parameters \( \alpha \) and \( \beta \). This means that \( g(\theta) = \theta^{\alpha-1}e^{-\theta}\beta^\alpha/\Gamma(\alpha) \). Here the posterior distribution of \( \theta \) given \( x \) is also a gamma distribution, with parameters \( \alpha+x \) and \( \beta+1 \). In this situation,

\[
\int (P(\theta|x)g(\theta))^{1/2}d\theta = \int \left\{ \frac{\beta^\alpha e^{\theta} e^{-\theta \beta \theta} \Gamma(\alpha+x)}{\Gamma(\alpha)} \cdot \frac{\theta^{\alpha+x} e^{-\theta} \Gamma(\beta+1)}{\Gamma(\alpha+x)} \right\}^{1/2} d\theta
\]

\[
= \frac{\theta^{\alpha/2} e^{(\beta+1)\theta^{-1}}}{\Gamma(\alpha) \Gamma(\alpha+x)} \int_0^\infty \theta^{\alpha-1/2} e^{-\theta (\beta+1/2)} d\theta
\]

\[
= \frac{\theta^{\alpha/2} (\beta+1)^{-1/2}}{\Gamma(\alpha) \Gamma(\alpha+x)} \cdot \int_0^\infty \Gamma(\alpha+x/2) / (\beta+1/2)^{\alpha+x/2}.
\]
Therefore,

\[
d^*(x) = 1 - \frac{\alpha/2 (\beta+1)^2 \Gamma(\alpha+x/2)}{\sqrt{\Gamma(\alpha) \Gamma(\alpha+x) (\beta+1)/2}}.
\] (2.2.9)

In the general case then, suppose the likelihood function is a member of the one parameter exponential family, i.e. \( f(x|\theta) = c(\theta)h(x) \exp[\eta(\theta)T(x)] \). Also, suppose our conjugate prior is of the form \( g(\theta) = k[c(\theta)]^a \exp(\beta.g(\theta)) \). Then the posterior distribution for \( \theta \) given \( x \) will be,

\[
P(\theta|x) = \frac{[c(\theta)]^{\alpha+1} \exp[\eta(\theta)(T(x)+\beta)]}{\int [c(\theta)]^{\alpha+1} \exp[\eta(\theta)(T(x)+\beta)]d\theta}.
\]

For convenience, let \( \int [c(\theta)]^{\alpha+1} \exp[\eta(\theta)(T(x)+\beta)]d\theta = r(x) \). Then,

\[
\int (P(\theta|x)g(\theta))^\frac{1}{2}d\theta = \int \left( \frac{k[c(\theta)]^{2\alpha+1} \exp[\eta(\theta)(T(x)+2\beta)]}{r(x)} \right)^\frac{1}{2} d\theta
\]

\[
= \left( \frac{k}{r(x)} \right)^\frac{1}{2} \int [c(\theta)]^{\alpha+\frac{1}{2}} \exp[\eta(\theta)(\frac{T(x)}{2} + \beta)]d\theta.
\]

If we define \( \ell(x) \) to be \( \int [c(\theta)]^{\alpha+\frac{1}{2}} \exp[\eta(\theta)(\frac{T(x)}{2} + \beta)]d\theta \), then the general form of the distance for the exponential family will be

\[
d^*(x) = 1 - \left( \frac{k}{r(x)} \right)^\frac{1}{2} \ell(x).
\] (2.2.10)
2.3. Robust Considerations

Now that the distances we need are calculated, we can consider the question of how to interpret the resulting values. For example, suppose we are interested in estimating the mean, say \( \theta \), of the likelihood function. We specify some prior distribution for \( \theta \) and conduct an experiment which yields a sample \( x \). We then calculate the posterior distribution given this \( x \), and evaluate the scaled distance function. Suppose this distance turns out to be close to one, but we are reasonably certain that the \( x \) that we have sampled is accurate. One possible explanation for this problem is that the prior was misspecified in some way.

This notion of questioning the choice of the prior after a sample value has been observed is very much in keeping with the robust Bayesian viewpoint that was discussed in the introduction. There it was mentioned that a decision is considered posterior robust if it satisfies certain criteria with respect to the posterior distributions derived from a certain class of priors, say \( \Gamma \). One measure of this is known as \( \varepsilon \)-posterior robustness. In this case, an action \( a_0 \) is \( \varepsilon \)-posterior robust with respect to \( \Gamma \) if for the observed \( x \),

\[
\sup_{\pi \in \Gamma} \rho(\pi, x, a_0) - \inf_{a \in A} \rho(\pi, x, a) \leq \varepsilon.
\]

Here \( \rho(\pi, x, a_0) \) is the posterior Bayes risk.

Now, if the procedure we decide to use is posterior robust, there really is no need for further concern. But if it is not, then lacking
an opportunity for further elicitation of the prior, we must try and decide on an alternate way to proceed. The real problem is, that often in practice we opt to use the conjugate prior simply because it is most convenient, or the calculations needed when we use other priors are impossible to evaluate beyond an approximation. But, conjugate priors by definition have the same type of tails as the likelihood functions. And, as mentioned, using a prior whose tail is flatter than that of the likelihood tends to yield better results.

One approach then for handling the problem of a lack of posterior robustness that was discussed in the introduction involves marginalizing the posterior risk with respect to the predictive density for $x$, and then checking for Ĉ-procedure robustness.

An alternative idea that we considered was to calculate the distances, for a fixed $x$, between each of the prior and posterior distributions in $\Gamma$ and $\Gamma^*$ respectively. Our intuitive notion was that this distance would be maximum when the data and the prior information were most in conflict, while it would be minimum when they were most in accordance. Therefore, it was conjectured that use of the prior which minimized this distance would lead to the more robust procedure. For example, if your likelihood function was, say $f(x|\theta) \sim N(0,1)$, and you had reason to believe that $\theta$ was "close" to 0, then you might consider using the conjugate normal prior with mean 0, and some variance $\tau^2$. Whereas, someone else who may not agree that $\theta$ is any more likely to be centered about one value
than another may choose the uniform prior. Then suppose you take a sample and observe a very large $x$. Intuition might lead you to think that using the more robust uniform prior, for which really no one $x$ should be more surprising than any other, would give you a smaller distance measure than if you had used the conjugate prior. The problem is though, that you cannot make a comparison of this sort. For in practice, the distances measured using the flatter tailed priors will tend to be larger than those of, say, the exponential family just by virtue of the fact that they start out with a much larger variance.

Now, one way around this situation would be to compare not the absolute distances over the class $\Gamma$, but the relative probabilities of the distances. Specifically, for each prior, $g$, and the observed $x^*$ one could calculate $P(d_g(x) > d_g(x^*))$. Then, the prior for which this probability is maximum could be considered the most robust. Note here that we have been forced to consider a frequentist approach, but in this case it seems the only solution to the problem of comparing these measures.

What seems then to be the most reasonable suggestion is to calculate the distance for the prior you think best reflects your a priori information, then if this distance turns out to be very large, you should view this as a signal which alerts the statistician to some sort of dissonance within the model. Where, one source of the trouble could be an improperly specified prior. In this case, one
might try varying the prior over some reasonable class, while observing the behavior of the distance measure.

Now, as mentioned in the introduction, another value that has been suggested as a sort of trouble signal is $h(x)$, the value of the predictive density of the sample observation. In this case, a small value of $h(x)$ would be interpreted as a sign, say that a "surprising" data value had occurred. Note that the value is surprising with respect to the way you have modeled the problem, i.e. your choice of a likelihood function and a prior. And since we have decided not to question the likelihood function, it means that the value is surprising with respect to the prior.

Box (1980) suggested that one way to use $h(x)$ was to compare the value of the predictive density of your observation, say $h(x^*)$, to the totality of all possible $x$ values. Specifically, he suggested using $\alpha = P(h(x) < h(x^*))$ as a measure of surprise, or what he calls the "predictive check." His rationale for this choice was that "since the posterior distribution combined the information from the prior and the sample, it could not be used to distinguish a surprising observation. But, since the predictive density in a sense contrasts these two sources, it could be used to check their compatibility."

Suppose then, for example, that we have a random sample of size $n$ from a normal population with unknown mean $\theta$ and known variance $\sigma^2$. And, we decide to use the conjugate normal prior with mean $\mu$ and variance $\tau^2$. In this case, the predictive density of the sample
\( \tilde{x} = (x_1, \ldots, x_n) \) is

\[
h(\tilde{x}) = \sigma^{-\frac{(n-1)}{2}} \frac{\sigma^2}{2} \exp\left[-\frac{1}{2} \left( \frac{(n-1)s^2}{\sigma^2} + \frac{(\tilde{x} - \mu)^2}{n^{-1} \sigma^2 + \tau^2} \right) \right].
\]

Then the predictive check would be found by calculating

\[
\alpha = P(h(x) < h(\tilde{x})) = P(\chi^2_n > g(\tilde{x})),
\]

where

\[
g(\tilde{x}) = \frac{(\tilde{x} - \mu)^2}{n^{-1} \sigma^2 + \tau^2} + \frac{(n-1)s^2}{\sigma^2}.
\]

As evidenced then by this example, this method does have certain obvious advantages. The \( \alpha \) here reduces to a probability involving a known distribution for which there are tables available. Also, the expression \( g(\tilde{x}) \) seems reasonable in that the first part is a function of a standard statistic used to measure the difference between two estimates of a mean, while the second part compares two variance estimates. But again, while these are nice advantages, this method does have the disadvantage of using as its reference set other possible \( x \)'s. This frequentist idea is not really in keeping with the Bayesian attitude of making inferences only from the \( x \) you have observed.

A measure then which contrasts the information in the posterior, at a given \( x \), to the information in the prior, would be the previously defined distance measure. Here though, the emphasis has changed in that a large distance value would indicate that the \( x \) we have observed seems to be giving us information which is conflicting in some sense.
with our a priori views about the behavior of the parameter. Unfortunately, \( d(x) \) does not have a nice distributional form. So we do not have the advantage of checking our measure against standard tabulated values. What we can do is use the fact that \( 0 \leq d(x) \leq 1 \) to give us some feel for the consonance of our information.

If our sample space, \( X \), is finite we could let \( x \) vary over the sample space, for a fixed prior, and get an ordering for the \( x \)'s, i.e. the minimum distance would be for the least surprising \( x \), and the maximum for the most surprising one.

For example, suppose we have a binomial random variable, i.e. \( f(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \), where \( n = 3 \). And suppose we use the conjugate beta prior with \( \alpha = 2 \) and \( \beta = 1 \), i.e. \( g(\theta) = 2\theta \), \( 0 \leq \theta \leq 1 \). Then using (2.2.8) we have

\[
\begin{align*}
d^*(x) &= 1 - \frac{B(2 + x/2, 5/2 - x/2)}{\sqrt{B(2,1) B(2+x,4-x)}} \\
&= 1 - \frac{\sqrt{2} B(2 + x/2, 5/2 - x/2)}{\sqrt{B(2+x,4-x)}},
\end{align*}
\]

since \( B(2,1) = \Gamma(2) \cdot \Gamma(1)/\Gamma(3) = 1/2 \). For this example, we find that \( d^*(0) = .304 \), \( d^*(1) = .148 \), \( d^*(2) = .05 \), and \( d^*(3) = .114 \). This would lead us to conclude that \( x = 0 \) is the most surprising value, while \( x = 2 \) is the value most consonant with our prior beliefs. Note that this result seems quite reasonable since our prior estimate for \( \theta \), \( \frac{\alpha}{\alpha+\beta} \), is 2/3, and observing \( x = 2 \) would give us the classical sample
estimate $x/n = 2/3$. On the other hand, the observation $x = 0$ would give us the sample estimate farthest from our prior guess. If we were to look at the marginal distribution of $x$ we would find that $h(x) = \frac{x+1}{10}$. In this case, using the ordering suggested by Box (1980), $x = 0$ would also be the most surprising value, but the least surprising would be $x = 3$, and the $\alpha$ level would monotonically increase with $x$. The reason for this ordering does not seem as intuitively appealing.

Using the uniform prior with this likelihood function points out a weakness of the use of the marginal density. For if we have $g(\theta) = 1$, $0 < \theta < 1$, which is a beta density with $\alpha = \beta = 1$, then the distance measure is

$$d^*(x) = 1 - \frac{B(1 + x/2, 5/2 - x/2)}{B(1 + x, 4 - x)}.$$

We find here that $d^*(0) = .2$, $d^*(1) = .075$, $d^*(2) = .075$, and $d^*(3) = .2$, leading us to conclude that observing $x = 1$ and $x = 2$ is equivalent with respect to their being more consonant with the prior. But in this case, $h(x) = 1/4$, which does not give us any indication of the relative consonance of the possible values.

Unfortunately though, when our sample space is infinite we cannot use this method. In this case, we could perhaps vary the $x$ over a certain range of values and observe how the distance measure behaves.

So far we have always discussed observing one $x$, but obviously
this technique applies when we have a sample $x = (x_1, \ldots, x_n)$. Here we would calculate the distance between $P(\theta | x)$ and $g(\theta)$, and interpret a large distance measure as an indication of a surprising sample. Referring back to the normal example investigated by Box (1980) where he found $a$ to be equal to the probability that a chi-squared random variable with $n$ degrees of freedom was greater than a function of the difference between $\bar{x}$ and the prior mean, and a ratio of variance estimates, recall that the distance measure also is a function of these quantities, i.e.

$$d^*(x) = 1 - \frac{\left(\frac{2}{\tau} \left(\frac{4}{\sigma^2 + \tau} \right)^{\frac{1}{2}} \right)}{\sqrt{2 \tau^2 + n \tau^4}} \exp \left\{ -\frac{1}{4} \frac{n^2 \tau^4 (\mu - \bar{x})^2}{(n \tau^2 + \tau^2)(2 \tau^2 + n \tau^4)} \right\}$$

and,

$$\tilde{d}(x) = \frac{\ln \left(1 + \frac{n \tau^2}{\sigma^2} \right) + \frac{n^2 \tau^2}{(n \tau^2 + \tau^2)^2} (\mu - \bar{x})^2 + \frac{\sigma^2}{\sigma^2 + n \tau^2} - 1}{\ln \left(1 + \frac{n \tau^2}{\sigma^2} \right) + \frac{n^2 \tau^2}{(n \tau^2 + \tau^2)^2} (\mu - \bar{x})^2 + \frac{\sigma^2}{\sigma^2 + n \tau^2} + 1}.$$ 

Therefore, a small $\alpha$ level for Box would correspond to a large value for $d(x)$.

For example, Box (1980) discussed a sample consisting of four analytical tests of yield performed on a single batch from an industrial process where it was thought that the testing variance was $\sigma^2 = 1$, the prior process mean was $\mu = 70$, and the batch to
batch variance was $\sigma^2 = 2$. The sample resulted in $x = (77, 74, 75, 78)$, with $\bar{x} = 76$. Here he calculated $\alpha = .001$, thus in his view discrediting use of the posterior distribution for estimative purposes. While, for the same example, we calculate $d^*(x) = .968$ and $\tilde{d}(x) = .938$, both indicating a large degree of disconsonance.

Suppose then, that we assume we have a sample for which the calculated distance measure is close to one, and for the time being that we do not suspect a misspecified prior, then how could we decide which $x_i$ was most discordant with the model? Well, we could calculate the posterior distribution of $\theta$ given all the sample points except $x_i$, call this $P_i(\theta|x)$, and then calculate the distance between this distribution and the prior. This would give us an ordering of the data points in terms of which were most consonant with the prior, and which were not. Note here that if the distance between the prior and the deleted posterior is small, this would indicate a surprising value. Since this means that the sample would be more consonant with the prior if we deleted that observation. For example, if we consider the normal distribution with the conjugate prior $g(\theta) \sim N(\mu, \tau^2)$, the deleted posterior is

$$P_i(\theta|x) \sim N \left( \frac{(n-1)\tau^2 x_{(i)} + \sigma^2 \mu}{(n-1)\tau^2 + \sigma^2}, \frac{\sigma^2}{\sigma^2 + \tau^2(n-1)} \right),$$

with $x_{(i)} = \frac{1}{n} \sum_{j \neq i} x_j / (n-1)$. Therefore, using (2.2.1) we find that
Note that the distance again is an increasing function of the difference between the prior and the sample means. Thus, if the value we are considering is "far" from our prior estimate, then deleting it will reduce the overall distance.

In conclusion, the robust Bayesian view suggests comparing the sample information with the prior information before using the posterior distribution for inferential purposes. The problem is that in practice it is often quite difficult to find a procedure which will be robust in a wide variety of situations. Berger (1982) has suggested two criteria, these being \( \phi \)-posterior robustness and \( \phi \)-procedure robustness, and has noted that you only need to consider the second type when the first is lacking. But, this second type involves averaging over the sample space, an undesirable quality from the Bayesian viewpoint. Box (1980) suggests using the marginal density as a "predictive check," but this also involves considering the entire sample space. Therefore, an alternative measure suggested is this distance function.

The procedure here being that after one has observed an \( x \), one should calculate the posterior distribution, and then measure the distance between the posterior and prior distributions. If the distance is reasonably small, then one could feel comfortable about making inferences based on this posterior. If the distance is large, then there are two possible information sources which would be in question.
The first is the prior density. In this case, the analyst could perhaps check for posterior robustness, or maybe try varying the prior over some reasonable class of priors, while observing the behavior of the relative probabilities of the distance measures. On the other hand, the sample information could be under suspicion. If the sample space is finite, the analyst could judge how large the distance was, relative to the other x's, by calculating the measure for each x. Note here that we are not averaging over values that did not occur, rather we are making a relative comparison of the distances.

More generally, if we have a sample available to us, we could calculate the distances between the prior and each deleted posterior in order to isolate the more dissonant observations.

Both of these methods then would help distinguish "surprising" sample values, in which case further investigation into the sampling procedure should be undertaken before the posterior is used for inferential purposes.
3. THE USE OF THE DISTANCE MEASURE IN THE
DETECTION OF OUTLIERS

3.1. Introduction

DeFinetti (1961) stated that "According to the Bayesian point of view, there exist no observations to be rejected." For one starts with their prior distribution, observes a sample \( x \), and then calculates the posterior distribution based on all the observations. Now while it is true that Bayesian statistics has not typically concerned itself with the problem of outliers to the extent that classical statistics has, more attention is now being paid to the search for Bayesian criteria for the detection of outliers. For in those instances where after collecting the data, we notice that certain observations seem inconsistent with respect to the other values, it would be useful to have some way of distinguishing and analyzing such incongruities. So now we consider this problem of discerning outlying values, while working under the assumption that both the likelihood function and the prior distribution are correct.

The approach we will use entails calculating the posterior distribution for the unknown parameter, say \( \theta \), given the entire sample, and also the posterior distribution for \( \theta \) given subset deleted data sets. Then, the distance will be calculated between the full posterior and the deleted posterior. So suppose we begin by deleting each \( x_i \) and we calculate the posterior distribution without this observation. Then for each \( x_i \) we can evaluate the distance between the full
posterior and the posterior omitting this value. We could thus order
the observations from least to most consistent in terms of the magnitude
of the distance, i.e. the $x_i$ where the distance is smallest is most
consistent, while the largest distance would indicate a very influential,
perhaps outlying observation. This idea could then be extended to
where we delete a set (i) say, and measure the distance between the full
posterior and the subset deleted posteriors. This would establish
orderings for pairs, triplets, etc. Note here that this is similar to
what was done in the previous chapter, in that if we have a prior, and
a sample $x_1, \ldots, x_n$, then the posterior distribution for $\theta$ given
$x_1, \ldots, x_{n-1}$ is in effect the prior distribution for the next observa-
tion, $x_n$. So we are still contrasting the information in the most
recent prior with the information in the $x$ we now observe.

Because of the technical difficulties involved in the use of other
distributions, we will confine our attention to the case where our
posterior distributions are normal density functions. Therefore,
suppose we take a sample of size $n$ from a normal distribution with
unknown mean $\theta$ and known variance $\sigma^2$. We will use the conjugate
normal prior with mean $\mu$ and variance $\tau^2$. Then, $P(\theta|\bar{x}) \sim N$
\[
\frac{n\tau^2 x + \sigma^2 \mu}{n\tau^2 + \sigma^2}, \frac{\sigma^2}{\sigma^2 + n\tau^2}.
\]
Suppose we delete observation $x_i$, then
\[
P(i|\bar{x}) \sim N \left( \frac{(n-1)\tau^2 x^{(i)} + \sigma^2 \mu}{(n-1)\tau^2 + \sigma^2}, \frac{\sigma^2}{\sigma^2 + (n-1)\tau^2} \right),
\]
where $x^{(i)} = \frac{n}{j \neq i} \sum \frac{j}{n-1}$.

Thus, using (2.2.1) we find that
\[ d^* = 1 - \sqrt{2 \left[ (\sigma^2 + \mu^2)(\sigma^2 + (n-1)\tau^2) \right]^{\frac{1}{2}} \exp \left( -\frac{\tau^2(n(n-1)\tau^2(x_{(i)} - \bar{x}) + \sigma^2(\mu - x_i)^2}{\sigma^2((n-1)\tau^2 + \sigma^2)(n\tau^2 + \sigma^2)(2\sigma^2 + \tau^2(2n-1))} \right)} \]  

(3.1.1)

Since in this case we are not using the distance measure as an overall indication of consonance, we do not need to use the scaled version of the Kullback-Liebler divergence. Therefore, for simplicity, we will calculate \( d_2 \) using formula (2.2.2). Thus,

\[ d_2 = d(P(\theta|\chi_i), P(\theta|\chi)) = \frac{1}{2} \left\{ \ln \frac{\sigma^2 + (n-1)\tau^2}{\sigma^2 + n\tau^2} + \left( \frac{n(n-1)\tau^2(x_{(i)} - \bar{x}) + \sigma^2(\mu - x_i)^2}{\sigma^2((n-1)\tau^2 + \sigma^2)(n\tau^2 + \sigma^2)(2\sigma^2 + \tau^2(2n-1))} \right)^2 \right\} + \frac{\sigma^2 + n\tau^2}{\sigma^2 + (n-1)\tau^2} - 1 \} \]  

(3.1.2)

Notice that both of these measures are increasing functions of \( n(n-1)\tau^2(x_{(i)} - \bar{x}) + \sigma^2(\mu - x_i) \). Therefore, the distances will be minimum when this expression equals zero, i.e. when \( n(n-1)\tau^2(x_{(i)} - \bar{x}) = \sigma^2(x_i^* - \mu) \). Then since \( n(n-1)(x_{(i)} - \bar{x}) = (n-1)(x_{(i)} - x_i^*), \) the value for which the minimum occurs is \( x_i^* = ((n-1)\tau^2x_{(i)} + \sigma^2\mu)/((n-1)\tau^2 + \sigma^2), \)

which is the mean of the deleted posterior distribution. This means that the next observation we sample will be most consistent with the previous values and the prior when it equals the posterior mean of \( \theta \) given those observations. As it moves away from this value, the
distance will increase.

Analogously, if we delete a subset \( i \) of fixed size, say \( k \), the distances become

\[
d^* = 1 - \sqrt{2 \left( \frac{\left( \sigma^2 + n \tau^2 \right) \left( \sigma^2 + (n-k) \tau^2 \right) \left( 1 + \frac{1}{\sigma^2 + \tau^2 (2n-k)} \right)^{\frac{1}{2}} \exp \right. \left( \frac{\tau^2 (n(n-k) \tau^2 (x_{(i)} - \bar{x}) + k \sigma^2 (\mu - \bar{x}_i))^2}{\sigma^2 ((n-k) \tau^2 + \sigma^2)(n \tau^2 + \sigma^2)(2 \sigma^2 + \tau^2 (2n-k))} \right)}}
\]

and,

\[
d_2^* = \frac{1}{2} \left[ \ln \left( \frac{\sigma^2 + (n-k) \tau^2}{\sigma^2 + n \tau^2} \right) + \frac{n(n-k) \tau^2 (x_{(i)} - \bar{x}) + k \sigma^2 (\mu - \bar{x}_i))^2}{\sigma^2 ((n-k) \tau^2 + \sigma^2)} \right] + \frac{\sigma^2 + n \tau^2}{\sigma^2 + (n-k) \tau^2} - 1 \right].
\]

Here, \( \bar{x}_i = \sum_{j \in (i)} x_j / k \) and \( \bar{x}_{(i)} = \sum_{j \notin (i)} x_j / (n-k) \). As before, these distances will be at their minimum when \( n(n-k) \tau^2 (x_{(i)} - \bar{x}) + k \sigma^2 (\mu - \bar{x}_i) \)

= 0, i.e., if \( k(n-k) \tau^2 (x_{(i)} - \bar{x}_i) = k \sigma^2 (\bar{x}_i - \bar{x}) \). This will occur when \( \bar{x}_i = [(n-k) \tau^2 x_{(i)} + \sigma^2 \mu] / [(n-k) \tau^2 + \sigma^2] \), again the mean of the posterior distribution for \( \theta \) with the set \( i \) deleted. Therefore, if we set \( k = 2 \), we could order each pair of observations in terms of their distance from the remaining data set. Utilizing this technique then, we can classify the data in terms of the most inconsistent single values, pairs of values, triplets, etc. For sometimes, two points together will form an influential pair, while singly they will not seem as discrepant. But, if an observation does appear singularly discordant,
while also appearing in the most influential pair or triplet, this
would certainly seem to be an observation we should consider viewing
as an outlier.

3.2. Applications to the Linear Model

We will now consider the linear regression model, \( y = X\beta + \epsilon \),
where \( \epsilon \) is an \( N \times 1 \) random vector distributed as an \( N \)-dimensional
multivariate normal with mean vector zero and covariance matrix
\( \sigma^2 I \), here \( \sigma^2 \) is a scalar constant, \( \beta \) is a \( p \times 1 \) vector of
regression coefficients, \( X \) is an \( N \times p \) matrix of fixed, independent
variables, and \( y \) is the \( N \times 1 \) vector of "dependent" variables.

We will assume that \( y_1, y_2, \ldots, y_n \) are independent, and that for
each \( i \), \( y_i \) is normally distributed with mean \( \sum_{i=1}^{p} X_i \beta \) and
variance \( \sigma^2 \). Therefore, the likelihood function, \( f(y|\beta, \sigma^2) \), will
be of the following form,

\[
f(y|\beta, \sigma^2) = \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left\{ -\frac{1}{2\sigma^2}(y-X\beta)'(y-X\beta) \right\}. \tag{3.2.1}
\]

Suppose now that the joint prior distribution of \( \beta \) and \( W = 1/\sigma^2 \)
is a multivariate normal-gamma distribution. This would mean that the
conditional distribution of \( \beta \) given \( W = w \) is a \( p \)-dimensional
multivariate normal with mean, say \( \mu \), and covariance matrix \( \frac{1}{w} \Sigma \),
while the marginal distribution of \( W \) is a gamma distribution with
parameters \( \alpha \) and \( \Gamma \). In this case, the joint prior distribution of
\( \beta \) and \( W \) would be
\[ g(\beta, w) = \frac{w^{p/2}}{2^{p/2}} \exp\left( -\frac{w}{2} (\beta - \mu)'\Sigma^{-1}(\beta - \mu) \right) \exp(-yw). \] (3.2.2)

Therefore, in order to find the posterior distribution of \( \beta \) and \( w \), given \( y \), we will need to consider \( f(y/\beta, w) \cdot g(\beta, w) \).

Evaluating first the product of the exponential terms, we have

\[ \exp \left( -\frac{w}{2} (\beta - \mu)'\Sigma^{-1}(\beta - \mu) + (y - \chi'\beta)'(y - \chi'\beta) \right). \] (3.2.3)

Now, recall from the theory of least squares that an estimate \( \hat{\beta} \), where \( \hat{\beta} \) is any solution to the normal equations \( \chi'\chi \hat{\beta} = \chi'y \), will minimize the quadratic form \( (y - \chi'\beta)'(y - \chi'\beta) \). And if \( \hat{\beta} \) does solve those equations then

\[ (y - \chi'\beta)'(y - \chi'\beta) = (\hat{\beta} - \beta)'\chi'\chi(\hat{\beta} - \beta) + y'y - \hat{\beta}'\chi'y. \]

Therefore, the above can be expressed as

\[ \exp \left( -\frac{w}{2} (\beta - \mu)'\Sigma^{-1}(\beta - \mu) + (\hat{\beta} - \beta)'\chi'\chi(\hat{\beta} - \beta) + y'y - \hat{\beta}'\chi'y \right). \]

Next, let us define \( \beta_1 = (\Sigma^{-1} + \chi'\chi)^{-1}(\Sigma^{-1}\mu + \chi'y) \). In this case, we can show that

\[ (\beta - \mu)'\Sigma^{-1}(\beta - \mu) + (\hat{\beta} - \beta)'\chi'\chi(\hat{\beta} - \beta) = \beta'\Sigma^{-1}\beta - 2\beta'\Sigma^{-1} \mu + \mu'\Sigma^{-1} \mu \]

\[ + \beta'\chi'\chi\beta - 2\beta'\chi'\chi\hat{\beta} + \hat{\beta}'\chi'\chi\hat{\beta} = \beta'((\Sigma^{-1} + \chi'\chi)^{-1} + \chi'\chi)\beta - 2\beta'((\Sigma^{-1} + \chi'\chi)^{-1})\beta + \hat{\beta}'(\Sigma^{-1} + \chi'\chi)^{-1} \hat{\beta} \]

\[ + \mu'\Sigma^{-1}\mu + \hat{\beta}'\chi'\chi\hat{\beta} = (\beta - \beta_1)'(\Sigma^{-1} + \chi'\chi)(\beta - \beta_1) + \mu'\Sigma^{-1}\mu + \hat{\beta}'\chi'\chi\hat{\beta} \]

\[ - \beta_1'(\Sigma^{-1} + \chi'\chi)\beta_1. \]
Thus, substituting this equality in equation (3.2.3) we see that

\[ P(\beta, w|y) \propto \left[ w^{\alpha/2} \exp\left( -\frac{w}{2}(\beta - \beta_1)'(\Sigma^{-1} + X'X)(\beta - \beta_1) \right) \right] \cdot (w^{\alpha+(n/2)-1} e^{-\gamma_1/w}) \]

where

\[ \gamma_1 = \gamma + \frac{1}{2}(y'Y + \mu'\Sigma^{-1}\mu - \beta_1'\Sigma^{-1}\mu - \beta_1'X'y) \]

\[ = \gamma + \frac{1}{2}(y - X\hat{\beta}_1)'y + (\mu - \beta_1)'\Sigma^{-1}\mu \].

Thus, the conditional posterior distribution of \( \beta \) given \( W = w \) and \( y \), is a multivariate normal distribution with mean vector \( \beta_1 \) and covariance matrix \( \sigma^2(\Sigma^{-1} + X'X)^{-1} \). Also we note that the marginal distribution of \( W \) is a gamma distribution with parameters \( \alpha + (n/2) \) and \( \gamma_1 \).

Now suppose that \( n > p \), and the matrix \( X'X \) is nonsingular, then we can let \( \Sigma^{-1} \to 0 \), \( \alpha \to n-k/2 \), and \( \Gamma \to 0 \) in equation (3.2.4).

In this case, the conditional posterior distribution of \( \beta \) given \( W = w \) will be multivariate normal with mean vector \( \hat{\beta} = (X'X)^{-1}X'y \) and covariance matrix \( \sigma^2(X'X)^{-1} \). Also, the marginal distribution of \( W \) will be the gamma distribution with parameters \( (n-k)/2 \) and \( (n-k)s^2/2 \) where \( s^2 = \frac{1}{n-k}(y - X\hat{\beta})'(y - X\hat{\beta}) \). This will be useful since this limiting posterior distribution is the same as would be obtained if we use the improper joint prior distribution \( g(\beta, w) = 1/w \).
In order to calculate the distances between the full posterior distributions and the subset deleted posteriors, we need to evaluate the distance between two multivariate normal densities for both measures. Therefore, suppose we have the following two multivariate densities, \( P_1(\beta) \sim MN(\mu_1, \Sigma_1) \) and \( P_2(\beta) \sim MN(\mu_2, \Sigma_2) \). Then to evaluate the scaled Hellinger's distance we need to consider \( \left\{ P_1(\beta) P_2(\beta) \right\}^{1/2} \), where this is equivalent to

\[
(2\pi)^{-p/2} |\Sigma_1|^{1/4} |\Sigma_2|^{1/4} \exp \left\{ - \frac{1}{4} \left[ (\beta - \mu_1)' \Sigma_1^{-1} (\beta - \mu_1) + (\beta - \mu_2)' \Sigma_2^{-1} (\beta - \mu_2) \right] \right\}.
\]

First, let us consider the expression \( (\beta - \mu_1)' \Sigma_1^{-1} (\beta - \mu_1) + (\beta - \mu_2)' \Sigma_2^{-1} (\beta - \mu_2) \). This is equal to

\[
\beta' \Sigma_1^{-1} \beta - 2\beta' \Sigma_1^{-1} \mu_1 + \mu_1' \Sigma_1^{-1} \mu_1 + \beta' \Sigma_2^{-1} \beta - 2\beta' \Sigma_2^{-1} \mu_2 + \mu_2' \Sigma_2^{-1} \mu_2.
\]

Next, let us define \( \mu^* = (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1} (\Sigma_1^{-1} \mu_1 + \Sigma_2^{-1} \mu_2) \). Then the previous expression reduces to

\[
(\beta - \mu^*)' (\Sigma_1^{-1} + \Sigma_2^{-1}) (\beta - \mu^*) - \mu^* (\Sigma_1^{-1} \mu_1 + \Sigma_2^{-1} \mu_2) + \mu_1' \Sigma_1^{-1} \mu_1 + \mu_2' \Sigma_2^{-1} \mu_2.
\]

Then, let \( (\Sigma^*)^{-1} = \frac{1}{2} (\Sigma_1^{-1} + \Sigma_2^{-1}) \). Using these simplifications, we find that
\[ J \left( P_1(\beta)P_2(\beta) \right)^{\frac{1}{2}} d\beta = \left| \Sigma^* \right|^{\frac{1}{4}} \left| \Sigma_1 \right|^{\frac{1}{4}} \left| \Sigma_2 \right|^{\frac{1}{4}} \frac{1}{4} \exp \left( -\frac{1}{4} \mu_1^* \Sigma_1^{-1} \mu_1 + \mu_2^* \Sigma_2^{-1} \mu_2 \right) \]

\[ - \mu^* \left( \Sigma_1^{-1} \mu_1 + \Sigma_2^{-1} \mu_2 \right) \int (2\pi)^{-P/2} \left| \Sigma^* \right|^{\frac{1}{2}} \exp \left( -\frac{1}{2} \left( \beta - \mu^* \right)^* \left( \Sigma^* \right)^{-1} \left( \beta - \mu^* \right) \right) d\beta. \]

Since we know that the last integral will equal one, let us try to first simplify the determinants, i.e.

\[ \frac{\left| \Sigma^* \right|^{\frac{1}{2}}}{\left| \Sigma_1 \right|^{\frac{1}{4}} \left| \Sigma_2 \right|^{\frac{1}{4}}} = \frac{\left| \Sigma_1^{-1} + \Sigma_2^{-1} \right|^{\frac{1}{2}}}{\left| \Sigma_1 \right|^{\frac{1}{4}} \left| \Sigma_2 \right|^{\frac{1}{4}}}. \]

Using the equality \( \left| \Sigma_1^{-1} + \Sigma_2^{-1} \right| = \left| \Sigma_1^{-1} \right| \left| \Sigma_2^{-1} \right| \left| \Sigma_1 + \Sigma_2 \right| \), and the fact that \( \left| \Sigma^{-1} \right| = \left| \Sigma \right|^{-1} \), we find that the expression becomes

\[ \frac{2^P \left| \Sigma_1 \right|^{\frac{1}{2}} \left| \Sigma_2 \right|^{\frac{1}{2}}}{\left| \Sigma_1 + \Sigma_2 \right|^{\frac{1}{2}} \left| \Sigma_1 \right|^{\frac{1}{4}} \left| \Sigma_2 \right|^{\frac{1}{4}}} = \frac{2^P \left| \Sigma_1 \right|^{\frac{1}{4}} \left| \Sigma_2 \right|^{\frac{1}{4}}}{\left| \Sigma_1 + \Sigma_2 \right|^{\frac{1}{2}}}. \]

Next, let us consider the quantity inside the exponent. Here we have,

\[ \mu_1^* \Sigma_1^{-1} \mu_1 + \mu_2^* \Sigma_2^{-1} \mu_2 - \mu^* \left( \Sigma_1^{-1} \mu_1 + \Sigma_2^{-1} \mu_2 \right) = \mu_1^* \Sigma_1^{-1} \mu_1 + \mu_2^* \Sigma_2^{-1} \mu_2 \]

\[ - \left( \Sigma_1^{-1} \mu_1 + \Sigma_2^{-1} \mu_2 \right)^* \left( \Sigma_1^{-1} + \Sigma_2^{-1} \right)^{-1} \left( \Sigma_1^{-1} \mu_1 + \Sigma_2^{-1} \mu_2 \right) \]

\[ = \mu_1^* \Sigma_1^{-1} \mu_1 + \mu_2^* \Sigma_2^{-1} \mu_2 - \mu_1^* \Sigma_1^{-1} \left( \Sigma_1^{-1} + \Sigma_2^{-1} \right)^{-1} \Sigma_1^{-1} \mu_1 \]
Note here that $\Sigma^{-1}_1 - \Sigma^{-1}_1 (\Sigma^{-1}_1 + \Sigma^{-1}_2)^{-1} \Sigma^{-1}_1 = \Sigma^{-1}_2 - \Sigma^{-1}_2 (\Sigma^{-1}_1 + \Sigma^{-1}_2)^{-1} \Sigma^{-1}_2 = \Sigma^{-1}_1 (\Sigma^{-1}_1 + \Sigma^{-1}_2)^{-1} - \Sigma^{-1}_2 = (\Sigma^{-1}_1 + \Sigma^{-1}_2)^{-1}$. Thus, the exponent becomes

$$(\mu_1 - \mu_2)' (\Sigma^{-1}_1 + \Sigma^{-1}_2) (\mu_1 - \mu_2).$$

Therefore, the scaled Hellinger's distance between these two multivariate densities is

$$d^* = 1 - \frac{\exp - \frac{1}{4}[(\mu_1 - \mu_2)' (\Sigma^{-1}_1 + \Sigma^{-1}_2) (\mu_1 - \mu_2)]}{\exp - \frac{1}{4}[(\mu_1 - \mu_2)' (\Sigma^{-1}_1 + \Sigma^{-1}_2) (\mu_1 - \mu_2)]}.$$  

(3.2.5)

Next, we will evaluate the Kullback-Liebler divergence between two multivariate normal densities, i.e. $\int \frac{p_1(\beta)}{p_2(\beta)} d\beta$. In this case, we need to evaluate the following expression,

$$\int (2\pi)^{-p/2} |\Sigma_1|^{-\frac{1}{2}} \exp - \frac{1}{2}[(\beta - \mu_1)' \Sigma^{-1}_1 (\beta - \mu_1)]$$

$$\left[ \ln \frac{|\Sigma_1|^{-\frac{1}{2}} \exp - \frac{1}{2}[(\beta - \mu_1)' \Sigma^{-1}_1 (\beta - \mu_1)]}{|\Sigma_2|^{-\frac{1}{2}} \exp - \frac{1}{2}[(\beta - \mu_2)' \Sigma^{-1}_2 (\beta - \mu_2)]} \right] d\beta.$$
And we find that this is equal to

\[
\frac{1}{2} \ln \frac{|\Sigma_2|}{|\Sigma_1|} + \int (2\pi)^{-p/2} |\Sigma_1|^{-\frac{1}{2}} \exp - \frac{1}{2}[(\beta - \mu_1)' \Sigma_1^{-1} (\beta - \mu_1)] \\
\cdot \left[ - \frac{1}{2}(\beta - \mu_1)' \Sigma_1^{-1} (\beta - \mu_1) + \frac{1}{2}(\beta - \mu_2)' \Sigma_2^{-1} (\beta - \mu_2) \right] d\beta.
\]

Then, using the fact that \( \int (2\pi)^{-p/2} |\Sigma_1|^{-\frac{1}{2}} \exp - \frac{1}{2}[(\beta - \mu_1)' \Sigma_1^{-1} (\beta - \mu_1)] \times (\beta - \mu_1)' \Sigma_1^{-1} (\beta - \mu_1) \) \( d\beta = \text{tr}(I_p) = p \), we find that the above integral becomes,

\[
\frac{1}{2} \ln \frac{|\Sigma_2|}{|\Sigma_1|} - \frac{1}{2} p + \frac{1}{2} \int (2\pi)^{-p/2} |\Sigma_1|^{-\frac{1}{2}} \exp - \frac{1}{2}[(\beta - \mu_1)' \Sigma_1^{-1} (\beta - \mu_1)] \\
\cdot (\beta' \Sigma_1^{-1} \beta - 2\beta' \Sigma_1^{-1} \mu_1 + \mu_1' \Sigma_1^{-1} \mu_1) \) d\beta.
\]

The third part of this expression simplifies to

\[
\{ \text{tr} \Sigma_2^{-1} \Sigma_1 - 2\mu_1' \Sigma_1^{-1} \Sigma_1^{-1} \mu_2 + \mu_1' \Sigma_1^{-1} \Sigma_2 \Sigma_1^{-1} \mu_1 + \mu_2' \Sigma_2^{-1} \mu_2 \}
\]

\[
\cdot \exp\left[ - \frac{1}{2} \mu_1' \Sigma_1^{-1} \Sigma_1^{-1} \mu_1 - \frac{1}{2} \mu_1' \Sigma_1^{-1} \mu_1 \right]
\]

\[
= \text{tr} \Sigma_2^{-1} \Sigma_1 + (\mu_1 - \mu_2)' \Sigma_2^{-1} (\mu_1 - \mu_2).
\]

Therefore, the divergence measure becomes
Notice here that both of these measures reflect differences between the mean vectors as well as the covariance structure.

Recall that when we use the multivariate normal-gamma distribution as the joint prior distribution of $\beta$ and $w = 1/\sigma^2$, we find that the posterior conditional distribution of $\beta$ given $w = w$ is a $p$-dimensional multivariate normal with mean $\mu$ and covariance matrix $1/w \Sigma$. Since we are assuming that $\sigma^2$, or equivalently $1/w$, is known, we will use this conditional density as our posterior distribution.

Therefore, in order to identify outlying, or inconsistent data points, we will calculate the distance between the posterior distribution of $\beta$ given the entire set, and the posterior distribution when the set (i) has been omitted. Note here that (i) is some subset of fixed size, say $k$, of integers between 1 and $n$. Therefore, given the entire data set, the posterior conditional distribution of $\beta$ is, $P(\beta|\sigma^2, y) \sim MN(\beta_1, \sigma^2(\Sigma^{-1} + X'X)^{-1})$. Here $\beta_1 = (\Sigma^{-1} + X'X)^{-1}(\Sigma^{-1} \mu + X'y)$, which would be the Bayes estimate for $\beta$ under squared error loss.

Suppose then that we decide to delete a fixed subset of size $k$. We will denote the remaining values by $(X_{(i)}, y_{(i)})$ and the omitted set is $(X_i, y_i)$. Therefore, the model can be partitioned as follows:

$$y' = (y'_i, y'_{(i)}) = \beta' (X'_i, X'_{(i)}) + (\xi'_i, \xi'_{(i)}).$$
In this case, the posterior conditional density of $\beta$ will again be multivariate normal, i.e. $P_i(\beta | y, \sigma^2) \sim MN_p(\hat{\beta}_i, \sigma^2(\Sigma^{-1} + X'_i X(i))^{-1})$. Here, $\hat{\beta}_i = (\Sigma^{-1} + X'_i X(i))^{-1}(\Sigma^{-1}\mu + i X'_i y(i))$.

If we use the joint uniform prior, then the full conditional posterior distribution will again be multivariate normal, with $P(\beta | \sigma^2, y) \sim MN_p(\hat{\beta}, \sigma^2(X'X)^{-1})$, where $\hat{\beta}$ is the standard least squares estimate for $\beta$, i.e. $(X'X)^{-1}X'y$. When we delete the set $(i)$, the deleted posterior density becomes $P_i(\beta | y, \sigma^2) \sim MN_p(\hat{\beta}(i), \sigma^2(X'_i X(i))^{-1})$, with $\hat{\beta}(i) = (X'_i X(i))^{-1}X'_i y(i)$. We will consider this case first.

To evaluate the scaled Bellinger's distance between the full and subset deleted posterior distributions for $\beta$, given the joint uniform prior, we will use formula (3.2.5). Here $\mu_1 = \hat{\beta}, \mu_2 = \hat{\beta}(i), \Sigma_1 = \sigma^2(X'X)^{-1}, \Sigma_2 = \sigma^2(X'_i X(i))^{-1}$. To facilitate the calculation, we define the following expressions:

$$S = X'X, S(i) = (X'_i X(i)), V_i = X_i S^{-1}X'_i.$$ Let us consider the ratio of the determinants, we have,

$$\frac{|\Sigma_1|^{\frac{1}{4}}|\Sigma_2|^{\frac{1}{4}}}{|\Sigma_1 + \Sigma_2|^{\frac{1}{2}}} = \frac{|S^{-1}|^{\frac{1}{4}}|S^{-1}(i)|^{\frac{1}{4}}}{|S^{-1} + S^{-1}(i)|^{\frac{1}{2}}}.$$

In order to simplify this ratio we will use the identity,

$$S^{-1}(i) = S^{-1} + S^{-1}X_i(I - V_i)^{-1}X_i S^{-1}$$ proven by Beckman and Trussell (1974).
Therefore, we find that

\[
\frac{|S^{-1} \frac{1}{4} |S_{(i)}^{-1} \frac{1}{4} |}{|S^{-1} + S_{(i)}^{-1}| \frac{1}{8}} = \frac{|S^{-1} \frac{1}{4} |S^{-1} + S_{(i)}^{-1} X_{i}(I-V_{i})^{-1} X_{i} S^{-1} \frac{1}{4} |}{|2S^{-1} + X_{i}^T(I-V_{i})^{-1} X_{i} S^{-1}| \frac{1}{8}}
\]

\[
= \frac{|S^{-1} \frac{1}{4} |S^{-1} \frac{1}{4} |I+X_{i}^T(I-V_{i})^{-1} X_{i} S^{-1} \frac{1}{4} |}{|S^{-1} \frac{1}{8} 2^{P/2} |I+ \frac{1}{2} X_{i}^T(I-V_{i})^{-1} X_{i} S^{-1}| \frac{1}{8}}
\]

\[
= \frac{|I+X_{i}^T(I-V_{i})^{-1} X_{i} S^{-1} \frac{1}{4} |}{2^{P/2} |I+ \frac{1}{2} X_{i} S^{-1} X_{i}^T(I-V_{i})^{-1} \frac{1}{4} |}
\]

\[
= \frac{|I+X_{i} S^{-1} X_{i}^T(I-V_{i})^{-1} \frac{1}{4} |}{2^{P/2} |I+ \frac{1}{2} X_{i} S^{-1} X_{i}^T(I-V_{i})^{-1} \frac{1}{4} |}
\]

\[
= \frac{|I+V_{i}(I-V_{i})^{-1} \frac{1}{4} |}{2^{P/2} |I+ \frac{1}{2} V_{i}(I-V_{i})^{-1} \frac{1}{4} |}.
\]

(3.2.7)

This form is simpler to work with since we only have to invert the $X'X$ matrix, and do not need to calculate $(X_{(i)}'X_{(i)})^{-1}$ each time.

Next, we consider the quantity within the exponent, i.e.

$(\mu_1 - \mu_2)'(\Sigma_1 + \Sigma_2)^{-1}(\mu_1 - \mu_2) = 1/\sigma^2 (\Sigma_1 + \Sigma_2)^{-1} (\mu_1 - \mu_2)$. Note here that $(S_{(i)}^{-1} + S_{(i)}^{-1})^{-1} = S_{(i)}^{-1} + S_{(i)}^{-1} S = X_{(i)}^{-1} X_{(i)} X_{(i)} + X'X)^{-1}X'X$. We will then use the following identity: $AW'(I+WAW')^{-1} = (A^{-1}+W'W)^{-1}W$, $A$ is any positive definite symmetric matrix, and $W$ is any matrix. Here let $A = (X_{(i)}'X_{(i)})^{-1}$, then we have that
\[(X'_iX(i) + X'X)^{-1}X' = (X'_iX(i))^{-1}X'(I + X(X'_iX(i))^{-1}X')^{-1} = S^{-1}(i)X'(I + XS^{-1}(i)X')^{-1}.\]

This implies that,

\[S(i)(S(i) + S)^{-1}_S = X'(I + XS^{-1}(i)X')^{-1}X'.\]

Next, if we use the equality

\[(I + XS^{-1}(i)X')^{-1} = I - \frac{1}{2} XS^{-1}X' - \frac{1}{2} XS^{-1}X'_i(I - \frac{1}{2} V_i)^{-1}X_i S^{-1}X',\]

we find that,

\[S(i)(S(i) + S)^{-1}S = X'(I - \frac{1}{2} XS^{-1}X' - \frac{1}{2} XS^{-1}X'_i(I - \frac{1}{2} V_i)^{-1}X_i S^{-1}X')X\]

\[= X'X - \frac{1}{2} X'X - \frac{1}{2} X'_i(I - \frac{1}{2} V_i)^{-1}X_i\]

\[= \frac{1}{2}(S - X'_i(I - \frac{1}{2} V_i)^{-1}X_i).\tag{3.2.8}\]

Therefore, combining (3.2.7) and (3.2.8) proves that

\[d^* = 1 - \frac{|I + V_i(I - V_i)^{-1}|^\frac{1}{2}}{|I + \frac{1}{2} V_i(I - V_i)^{-1}|^\frac{1}{2}} \exp - \frac{1}{4}(\frac{1}{2} (\hat{\beta} - \beta(i))'(S - X'_i(I - \frac{1}{2} V_i)^{-1}X_i) \hat{\beta} - \beta(i))}.\tag{3.2.9}\]

If we consider the expression inside the exponent, we see that it is a generalized version of a class of statistics introduced by Cook (1977) to assess the influence of a subset when one estimates \(\beta\).
using the least squares method, i.e. \( D(A) = (\hat{\beta} - \hat{\beta}_{(i)})'A(\hat{\beta} - \hat{\beta}_{(i)}) \). He noted that the usual \((1-\alpha) \times 100\%\) confidence ellipsoid for the parameter \( \beta \) is given by the set of all vectors \( \beta^* \) such that

\[
\frac{(\beta^* - \hat{\beta})'X'X(\beta^* - \hat{\beta})}{ps^2} \leq F(p, n-p, 1-\alpha).
\]

Here \( s^2 = (y - X\hat{\beta})'(y - X\hat{\beta})/(n-p) \) and \( F(p, n-p, 1-\alpha) \) is the \( 1-\alpha \) percentile of the central F-distribution with \( p \) and \( n-p \) degrees of freedom.

Therefore, in order to determine the influence of the \( i^{th} \) data point, he suggested the measure

\[
D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})'X'X(\hat{\beta}_{(i)} - \hat{\beta})}{ps^2}, \quad i = 1, 2, ..., n.
\]

This then would provide some measure of the distance between \( \hat{\beta}_{(i)} \) and \( \hat{\beta} \) in terms of the levels of significance. For example, if \( D_i = F(p, n-p, .5) \), and you remove the \( i^{th} \) data point, the least squares estimate will move to the boundary of the 50% confidence region for \( \beta \). Cook (1977) notes that for "an uncomplicated analysis," we would like each \( \hat{\beta}_{(i)} \) to stay within say a 10% confidence region.

In our situation then, \( D(A) = (\hat{\beta} - \hat{\beta}_{(i)})'A(\hat{\beta} - \hat{\beta}_{(i)}) \) can be explained as a measure of the lack of fit of the data set \( (i) \) relative to some metric \( A \). Therefore, we can rewrite the distance measure as,
\[ d^* = 1 - \frac{|I + V_i (I - V_i)|^{-1}_{1/2}}{|I + \frac{1}{2} V_i (I - V_i)|^{-1}_{1/2}} \exp \left( \frac{1}{8} D \left( \frac{1}{\sigma^2} (S - X_i' (I - \frac{1}{2} V_i)^{-1} X_i) \right) \right). \]

(3.2.10)

Thus, we have shown that this distance measure is an increasing function of Cook's distance function.

Johnson and Geisser (1983) offer further insight into this measure. They suggest the following, let \( e \) denote a vector of ones and define
\[ J = ee'. \]

Then partition the \( X \) matrix in the following manner:

\[ X = (e, \tilde{X}) = \begin{pmatrix} e \\ \tilde{X} \\ (i) \end{pmatrix}. \]

Define \( \tilde{X} = N^{-1} e' X, A = N^{-1} (\tilde{X} - e X)' (\tilde{X} - e X), \) and \( \bar{X}_i = k^{-1} e' \tilde{X}_i. \) Now, consider the expression \( V_i = \tilde{X}_i S^{-1} \tilde{X}_i'. \) If we partition as above we see that,

\[ X_i S^{-1} X_i' = X_i \left( e'e' e'e' \right)^{-1} X_i' = X_i \left( \begin{pmatrix} N & N \tilde{X} \\ \tilde{X}' N & \tilde{X}' \tilde{X} \end{pmatrix} \right)^{-1} X_i'. \]

Next, using the fact that

\[ \begin{pmatrix} N & N \tilde{X} \\ \tilde{X}' N & \tilde{X}' \tilde{X} \end{pmatrix}^{-1} = \begin{pmatrix} N^{-1} + N^{-1} X A^{-1} X' & -N^{-1} X A^{-1} \\ -N^{-1} A^{-1} X' & N^{-1} A^{-1} \end{pmatrix}, \]

we see that
\[ v_i = N^{-1}[ee' + eX'A^{-1}X'e' - X_i'A^{-1}X_i'e' - eX'A^{-1}X_i'i + X_i'A^{-1}X_i'i] \]

\[ = N^{-1}[J + (\hat{X}_i - eX)A^{-1}(\hat{X}_i - eX)'] \]

This shows that the diagonal elements of \( v_i \) measure the distances between vectors of observed independent variables and averaged vectors, relative to an inner product based on the full data set. Therefore, a subset with large diagonal elements of \( v_i \) could be considered to be, say "distantly" observed.

If we consider here the Kullback-Liebler divergence, \( d(P|y, \sigma^2), P(i)\), we find that

\[ d = \frac{1}{2}\left[ \frac{1}{\sigma^2} (\hat{\beta} - \hat{\beta}(i))' S_{(i)} (\hat{\beta} - \hat{\beta}(i)) + \text{tr} [S^{-1}_{(i)}] - \ln |S_{(i)}| - p \right] \]

\[ = \frac{1}{2} [D - \frac{1}{2} (S^{-1} + S^{-1}X_i'(I-V_i)^{-1}X_iS^{-1})] + \text{tr} V_i(I-V_i)^{-1} - \ln |I + S^{-1}X_i'(I-V_i)^{-1}X_i| + (n-k-p)] \quad (3.2.11) \]

So in this case the measure is again a function of Cook's distance function and matrix \( V_i \).

Suppose now we consider the joint multivariate normal-gamma prior introduced earlier. In this case, the conditional posterior distribution of \( \beta \) based on the entire data set is \( P(\beta|\sigma^2, y) \sim MN_p(\beta_1, \sigma^2 (\Sigma^{-1} + X'X)^{-1}) \), where \( \beta_1 = (\Sigma^{-1} + X'X)^{-1}(\Sigma^{-1}u + X'y) \). The conditional
posterior density based on the data set with subset \((i)\) deleted is 
\[ P_{(i)}(\beta_1^*, \sigma^2(\Sigma^{-1} + x'(i)x(i))^{-1}), \text{ with } \beta_1^* = (\Sigma^{-1} + x'(i)x(i))^{-1} \]
\[ (\Sigma^{-1} + x'(i)y(i)). \]

To calculate the scaled Hellinger's distance between the full and subset deleted posteriors, we will use (3.2.5) with \(\mu_1 = \beta_1, \mu_2 = \beta_1^*, \Sigma_1 = \sigma^2(\Sigma^{-1} + x'x)^{-1}, \Sigma_2 = \sigma^2(\Sigma^{-1} + x'(i)x(i))^{-1}. \) Again, let us begin by considering the ratio of the determinants, here we have

\[
\frac{|\Sigma_1^{\frac{1}{4}}|\Sigma_2^{\frac{1}{4}}}{|\Sigma_1^{-1} + \Sigma_2^{-1}|^{\frac{1}{2}}} = \frac{|(\Sigma^{-1} + S)^{-1}|^{\frac{1}{4}}|\Sigma^{-1} + S(i)^{-1}|^{\frac{1}{4}}}{|(\Sigma^{-1} + S)^{-1} + (\Sigma^{-1} + S(i))^{-1}|^{\frac{1}{2}}}
\]

For simplicity, let us define the following quantities:

\[ \widetilde{S} = \Sigma^{-1} + S, \quad \widetilde{S}(i) = \Sigma^{-1} + S(i), \quad \widetilde{V}_i = x_i \widetilde{S}^{-1} x_i. \]

Thus, we find that the above can be rewritten as,

\[
\frac{|\widetilde{S}^{-1}|^{\frac{1}{2}}|\widetilde{S}^{-1}(i)|^{\frac{1}{2}}}{|\widetilde{S}^{-1} + \widetilde{S}^{-1}(i)|^{\frac{1}{2}}}
\]

In order to simplify this expression we need to establish identities analogous to those previously used. To this end, we consider the following propositions:
Proposition 1:

\[ S_{(i)}^{-1}X_1' = S_{(i)}^{-1}X_1'(I - \tilde{V}_{1})^{-1}. \]

Proof.

Clearly,

\[ S_{(i)}^{-1}(S_{(i)} - S_{1}) = S_{(i)}^{-1}(S_{(i)}^{-1}X_1'X_1'X_1') = S_{(i)}^{-1}(S_{(i)}^{-1}X_1'X_1') = I, \]

If we expand this expression, we have

\[ S_{(i)}^{-1}S_{1} = S_{(i)}^{-1}S_{1} = I. \]

Then, postmultiplying both sides by \( S_{(i)}^{-1}X_1' \) we obtain,

\[ S_{(i)}^{-1}X_1' - S_{(i)}^{-1}S_{1}S_{(i)}^{-1}X_1' = S_{(i)}^{-1}X_1', \]

which is equivalent to,

\[ S_{(i)}^{-1}X_1'(I - X_1'S_{1}^{-1}X_1') = S_{(i)}^{-1}X_1', \]

and this implies that,

\[ S_{(i)}^{-1}X_1' = S_{(i)}^{-1}X_1'(I - \tilde{V}_{1})^{-1}. \]
Proposition 2:

\[ S^{-1}_{(i)} = S^{-1} + S^{-1}_{(i)} (I - V_i) ^{-1} X_i S^{-1} . \]

Proof.

First, we have that

\[ \tilde{S} = (\Sigma + X'X) = (\Sigma + X'X)^{(i)(i)} + (I - V_i) = \tilde{S} (i) + S_i . \]

Then, if we premultiply by \( S^{-1}_{(i)} \) and postmultiply by \( S^{-1} \) we find,

\[ \tilde{S}^{-1}_{(i)} = \tilde{S}^{-1} + \tilde{S}^{-1}_{(i)} S^{-1} = \tilde{S}^{-1} + \tilde{S}^{-1}_{(i)} X_i S^{-1} . \]

Using Proposition 1, we have

\[ \tilde{S}^{-1}_{(i)} = \tilde{S}^{-1} + \tilde{S}^{-1}_{(i)} (I - V_i) ^{-1} X_i S^{-1} . \]

Returning then to the ratio of the determinants, we now see that

\[ \frac{|S^{-1}|^{\frac{1}{4}} |S^{-1}_{(i)}|^{\frac{1}{4}}}{|S^{-1} + S^{-1}_{(i)}|^{\frac{1}{2}}} = \frac{|\tilde{S}^{-1}|^{\frac{1}{4}} |\tilde{S}^{-1} + \tilde{S}^{-1}_{(i)} (I - V_i) ^{-1} X_i S^{-1}|^{\frac{1}{4}}}{|2 \tilde{S}^{-1} + \tilde{S}^{-1}_{(i)} (I - V_i) ^{-1} X_i S^{-1}|^{\frac{1}{2}}} \]

\[ = \frac{|I + X_i (I - V_i) ^{-1} X_i S^{-1}|^{\frac{1}{4}}}{2^{D/2} |1 + \frac{1}{2} X_i (I - V_i) ^{-1} X_i S^{-1}|^{\frac{1}{2}}} \]
Next, we consider the expression in the exponent of formula (3.2.5), i.e. $(\mu_1 - \mu_2)'(\Sigma_1 + \Sigma_2)^{-1}(\mu_1 - \mu_2) = \frac{1}{\sigma^2} (\beta_1 - \beta_1^*)'(\hat{S}^{-1} + \hat{S}^{-1}_s(1))^{-1}$

$(\beta_1 - \beta_1^*)$. If we use Proposition 2 here, we see that

$$(\hat{S}^{-1} + \hat{S}^{-1}_s(1))^{-1} = (2\hat{S}^{-1} + \hat{S}^{-1}_s)'(\hat{S}^{-1}_s(1))^{-1} = \frac{1}{2}(I + \frac{1}{2} X_i' (I - \hat{V}_i)^{-1} X_i \hat{S}^{-1})^{-1} \hat{S}^{-1/2}.$$

Thus, the exponent can be expressed as

$$\frac{1}{2\sigma^2} (\beta_1 - \beta_1^*)'(I + \frac{1}{2} X_i' (I - \hat{V}_i)^{-1} X_i \hat{S}^{-1})^{-1} \hat{S}^{1/2} (\beta_1 - \beta_1^*).$$

Now let us define a Bayesian distance function analogous to the generalized version of Cook's measure, i.e. $\tilde{D}(A) = (\beta_1 - \beta_1^*)'A(\beta_1 - \beta_1^*)$. Recall that Cook motivated this measure by considering a 100(1-\alpha)% confidence region for $\beta$ using the least squares method. A Bayesian, on the other hand, considers what is known as a 100(1-\alpha)% credible region for the unknown parameter. By definition, if $\theta$ say is the
parameter of interest, a 100(1-\alpha)\% credible region for \theta is a subset C of the parameter space, such that

\[ 1 - \alpha \leq P(C|X) = \int_C P(\theta|X) \, d\theta. \]

Here we have that the conditional posterior density of \beta given y is, \( P(\beta|\sigma^2, y) \sim MN(\beta_1, \sigma^2(\Sigma^{-1} + X'X)^{-1}) \). Therefore, the posterior density of \( \beta - \beta_1 \sim MN(0, \sigma^2(\Sigma^{-1} + X'X)^{-1}) \). This implies that 

\[ \frac{1}{\sigma^2} (\beta - \beta_1)'(\Sigma^{-1} + X'X)(\beta - \beta_1) \]

has a chi-square distribution with \( p \) degrees of freedom. Thus, if we wanted to measure the degree of influence of the \( i^{th} \) observation, we could calculate the Bayes estimate using all the points, call this \( \beta_1 \), and the estimate using all points except \((X_i, y_i), \beta_1^*\), evaluate 

\[ \frac{1}{\sigma^2} (\beta_1 - \beta_1^*)'(\Sigma^{-1} + X'X)(\beta_1 - \beta_1^*), \]

and compare this to a chi-squared value with \( p \) degrees of freedom and significance level \( 1-\alpha \). Therefore, a generalized Bayesian measure of the influence of the set \((i)\) on the estimation of \( \beta_1 \) is given by 

\[ \tilde{D}(A) = (\beta_1 - \beta_1^*)'A(\beta_1 - \beta_1^*), \]

where this can be viewed in a sense as a measure of the Bayesian lack of fit of the data set \((i)\) relative to the metric \( A \).

We see that if we use the multivariate normal-gamma distribution as the joint prior density, the distance between the full posterior and the subset deleted posterior is,
The Kullback-Liebler divergence in this case becomes,

\[ d^* = 1 - \frac{|I + \frac{1}{2} \tilde{V}_i (I - \tilde{V}_i)_{-1}^{\frac{1}{2}}|}{|I + \frac{1}{2} \tilde{V}_i (I - \tilde{V}_i)_{-1}^{\frac{1}{2}}|} \exp \left\{ -\frac{1}{2} \bar{D} \left( \frac{1}{\sigma^2} (I + \frac{1}{2} X_i^T (I - \tilde{V}_i)_{-1} X_i^{-1}) - \frac{1}{\sigma^2} \tilde{V}_i \right) \right\}. \]

(3.2.13)

Both again are increasing functions of the Bayesian generalized version of Cook's distance function.

Johnson and Geisser (1983) considered a similar technique in which they used the Kullback-Liebler divergences between the predictive densities, full and subset deleted, in order to decide which sets were most influential in predicting a future y vector. Specifically, they considered the model, \( y = X\beta + \epsilon \), with \( \epsilon \sim MN(0, \sigma^2 I) \), and used the noninformative prior, \( g(\beta, \sigma^2) \propto \frac{1}{\sigma^2} \). Then, the full predictive density of a future y vector is \( MN_n(X\hat{\beta}, \sigma^2(I + X(X'X)^{-1}X')) \), and the subset deleted density is \( MN_n(X_{(i)}\hat{\beta}(i), \sigma^2(I + X_{(i)}(X_{(i)}X_{(i)})^{-1}X_{(i)}')) \). Johnson and Geisser define the Kullback-Leibler divergences
between these densities to be their "predictive influence functions."
In the case of the noninformative prior, with \( \sigma^2 \) known, they showed that

\[
2d(\mathcal{P}(\beta|\sigma^2, y), \mathcal{P}_{(i)}(\beta|\sigma^2, y)) = \frac{1}{2} D\left(\frac{1}{\sigma^2}(S - \frac{1}{2} X_i (I - \frac{1}{2} V_i)^{-1} X_i^t)\right) \\
+ \ln |I + \frac{1}{2} V_i (I - V_i)^{-1}| - \frac{1}{2} \text{tr} V_i (I - \frac{1}{2} V_i)^{-1}.
\]

Again, this is a function of Cook's distance and the matrix \( V_i \).

In order to calculate these predictive densities though, we must average over the sample space of \( X \), a procedure previously noted as being undesirable. Although Johnson and Geisser (1983) are specifically addressing the prediction problem, it seems that the more general approach would be to use the distances between the full and deleted posterior densities. Then, if it is decided that a certain value, or set of values, is suspected of being outliers, the analyst could decide to predict \( y \) using the deleted data set.

Therefore, using the distances between the full and subset deleted posterior distributions of \( \beta \) as a way of analyzing the effect of a subset seems to be more in keeping with the Bayesian viewpoint. For, it gives a way of discerning the sensitivity of an estimate to potentially outlying observations. Then, depending on the specific problem being considered, the estimates could be adjusted accordingly.
3.3. Modeling Outliers: An Open Topic

An alternative approach to the problem of outliers that has been investigated by a number of authors, involves building the possibility of such an occurrence into the statistician's model. Although most procedures are derived under the assumption that each value in a given set of observations is generated by a specific stochastic model, the very fact that the analyst makes provisions for the rejection of "outliers" shows that this assumption is not in total accordance with their beliefs.

One example of this type of thinking was presented by Box and Tiao (1968). While considering the case of the general linear model, they viewed the data as a series of runs made by some experimental apparatus. While most of the runs would be considered "good," there was the possibility of a few "bad" runs, i.e. the observations may have been contaminated in some way, perhaps recording errors, temporary changes in experimental conditions, or the use of abnormal experimental units. Therefore, for any given run there will be two alternative models, the first a standard one appropriate when the observation comes from a "good" run, and an alternative model for when it comes from a "bad" run. Then, a prior probability would be assigned to the chance of a "good" or "bad" run.

Specifically, they considered the linear model \( y = X\beta + \epsilon \), where they supposed that the errors could independently have come from one of two distributions, a standard distribution \( f(\epsilon|\xi_1) \), or alternatively
Then, let $a(r)$ be the event that a particular set of $r$ of the $n$ $\xi$'s are from $g(\xi|\xi_2)$, while the remaining $s = n-r$ are from $f(\xi|\xi_1)$. Then, if the matrix $X$, and the vectors $y$ and $\xi$ are partitioned according to this dichotomy, the likelihood function of $(\beta, \xi_1, \xi_2, a(r))$ is

$$L(\beta, \xi_1, \xi_2, a(r)) = \tilde{f}(\xi(s)|\xi_1) \tilde{g}(\xi(r)|\xi_2)$$

where $\tilde{f}(\xi(s)|\xi_1)$ is the product of the density functions of the elements of $\xi(s)$, and $\tilde{g}$ that of $\xi(r)$. Therefore, the entire likelihood function would consist of $2^n$ expressions of this type, corresponding here to the $2^n$ possible combinations of the $\xi$'s.

Next, they assign the prior probability $p(r)$ to the event $a(r)$, where $p(r) > 0$ and $\sum_r p(r) = 1$. Then, if the prior distribution is $p(\beta, \xi_1, \xi_2)$, the posterior density of $(\beta, \xi_1, \xi_2, a(r))$ given $y$ will be

$$p(\beta, \xi_1, \xi_2, a(r)|y) = \frac{p(r)p(\beta, \xi_1, \xi_2)\tilde{f}(\xi(s)|\xi_1)\tilde{g}(\xi(r)|\xi_2)}{\sum_r p(r)h(y(r) \sim g; y(s) \sim f)}.$$

(3.3.1)

Here, $h(y(r) \sim g; y(s) \sim f)$ is the marginal distribution of $y$ under the assumption that the elements of $\xi(r)$ are from $g(\xi|\xi_2)$ and the elements of $\xi(s)$ are from $f(\xi|\xi_1)$. If we then integrate (3.3.1) with respect to $(\xi_1, \xi_2)$ the distribution of $(\beta, a(r))$ becomes,
\[ p(\beta, a_{(r)} | y) = p(a_{(r)} | y) \ p(\beta | a_{(r)}, y), \]

where

\[ p(a_{(r)} | y) = \frac{\sum_{(r)} p^r h(y_{(r)} \sim g; y_{(s)} \sim f)}{\sum_{(r)} p^r h(y_{(r)} \sim g; y_{(s)} \sim f)} , \]

and,

\[ p(\beta | a_{(r)}, y) = \frac{\int p(\beta, \xi_1, \xi_2 \sim g; \xi_{(r)}) \ \tilde{p}(\epsilon | \xi_1, \xi_2) \ d\xi_1 \ d\xi_2}{h(y_{(r)} \sim g; y_{(s)} \sim f)} . \]

This last density is the conditional posterior distribution of \( \beta \) given that a particular \( r \) combination of the \( \epsilon \)'s are from \( g(\epsilon | \xi_2) \), while the previous expression is the marginal posterior distribution of \( a_{(r)} \). Therefore, the marginal posterior distribution of \( \beta \) can be expressed as a weighted average of \( 2^n \) such conditional distributions, i.e.

\[ p(\beta | y) = \sum_{(r)} p(a_{(r)} | y) \ p(\beta | a_{(r)}, y) . \]

And, if we consider the weights, \( p(a_{(r)} | y) \), we see that they can be expressed as,

\[ p(a_{(r)} | y) = \frac{c \ p^r h(y_{(r)} \sim g; y_{(s)} \sim f)}{p^o h(y_{(r)} \sim f; y_{(s)} \sim f)} , \]
where \( c \) is the normalizing constant,

\[
c = \frac{p^{(0)} h(y(r) \sim g; Y(s) \sim f)}{\sum_{(r)} p^{(r)} h(y(r) \sim g; Y(s) \sim f)} .
\]

Using this expression, the weights are seen to be proportional to the ratio of the posterior probabilities of two alternative hypotheses.

The first is that the \( \varepsilon_r \) are from \( g(\varepsilon | \xi_2) \) and the second is that the \( \varepsilon_r \) are from \( f(\varepsilon | \xi_1) \), both given that the remaining \( \varepsilon \)'s are from \( f(\varepsilon | \xi_1) \).

Box and Tiao (1968) then consider the case where the error \( \varepsilon \) associated with each observation could have come from two different models, a 'central' model \( N(0, \sigma^2) \) and an alternative model, \( N(0, k^2 \sigma^2) \), with probabilities \((1-\alpha)\) and \(\alpha\) respectively. Here \( k \) and \( \alpha \) are assumed to be known, and the improper prior is used. They show that the posterior distribution of \( \beta \) can be expressed as a weighted average of \( 2^n \) multivariate t-distributions, with weights given by,

\[
p(\alpha^{(r)} | y) = w^{(r)} = c \left( \frac{\alpha}{1-\alpha} \right) \frac{1}{\phi} \frac{1}{\sqrt{X'X}} \frac{1}{\sqrt{X'X + \phi X^{(r)} X^{(r)\prime}}} \left( \frac{S^2}{s^2} \right)^{-\nu / 2},
\]

where \( \phi = 1-k^{-2} \) and \( \nu = n-p \).

Another use of this technique was investigated by Abraham and Box (1978). The model they considered was \( y = X\beta + \delta Z + \varepsilon \), where
Z is an n x 1 vector consisting of r ones and (n-r) zeros. Here the probability of \( Z_i \) being one is \( \alpha \), and equalling zero, is \( 1-\alpha \). Letting \( \alpha \) be known and assuming a uniform prior, they show that the posterior distribution can be expressed again as a weighted average of multivariate t-distributions, here with weights

\[
W(r) \propto \left( \frac{\alpha}{1-\alpha} \right)^r \frac{1}{u^n \Gamma \left( \frac{n}{2} \right)} x^{\frac{r}{2}} \left( 1 + \frac{x}{v} \right)^{-\frac{n+1}{2}},
\]

where \( v(r) = 1 - r^{-1}z^t(z) \).

For a nice review of the various applications of this idea of expressing the posterior distributions as a weighted average of densities allowing for outliers, see Freeman (1980).

Now in practical application, what these authors suggest doing is to only consider the first few terms in the summation of the densities. That is, we can write the posterior distribution in the following form:

\[
p(\beta|y) = w_o p_o(\beta|y) + \sum_{i=1}^{n} w_i p_i(\beta|y) + \sum_{i<j}^{n} w_{ij} p_{ij}(\beta|y) + \ldots.
\]

Here \( p_o(\beta|y) \) would be the appropriate density if all the observations were from the central model, the distributions \( p_i(\beta|y) \) correspond to allowing the possibility of each observation being an outlier, those in next summation allow for the possibility of two outliers, etc. Likewise, \( w_o \) is the posterior probability that no
observations are bad, \( w_i \) that only the \( i^{th} \) observation is bad, etc.

In actual applications it has been found to be sufficient to include the leading term, terms in the first summation, and perhaps those in the second summation. The weights are then normalized accordingly. Thus, considering the magnitude of the weights associated with each density should give the analyst insight into the most likely configuration of outliers, i.e. the largest weight would correspond to the most plausible possibility. We should note here that once the weights have been established, any estimates drawn from the model will be a weighted average of the estimates from each density. For example, \( E(\beta|y) = w_0 E(p_0(\beta|y)) + \sum w_i E(p_i(\beta|y)) \) and this is in accordance with the intuitive idea for detecting outliers first suggested by DeFinetti (1961). Whereby, he hypothesized that if you could express an estimate as a weighted average of your data, then the weights should relate to the degree to which the particular value could be considered an outlier.

Now, after considering this technique of modeling outliers we conjectured that there should be a natural connection between our distance notion and the weights given to each density. For, suppose you consider using a posterior density for the parameter of interest which allows for the possibility of each value being an outlier, i.e.

\[
p(\theta|x) = w_0 p_0(\theta|x) + \sum_{i=1}^{n} w_i p_i(\theta|x).
\]

Then it would seem that if we measured the distance between each of the densities, \( p_0, p_1, p_2, \ldots, p_n \)
and the posterior \( p(\theta|\mathbf{X}) \), that the density which was closest to
\( p(\theta|\mathbf{X}) \) should be the one with the largest weight.

The problem here though is that it was not theoretically possible, using either distance function, to show such a connection. And even when specific distributions were considered, it turned out to be impossible to evaluate any of the integrals with the posterior expressed as a weighted sum. Some insight could be gained perhaps by numerically approximating the distances using a computer, evaluating the weights, and then checking the results to see how the two methods correspond. But as yet this remains an open question.
4. A MEASURE OF INFORMATION

4.1. Introduction

To this point we have been discussing various methods for comparing and contrasting the different sources of information available to a statistician. But, what exactly does the term "information" mean? As Basu (1975) states, "No other concept in statistics is more elusive in its meaning and less amenable to a generally agreed definition." But, while there does not seem to be a general consensus about the meaning of "statistical information," there does seem to be agreement about the notion of a statistical experiment. So this is where we begin our consideration of this topic.

Formally then, a statistical experiment, E, can be expressed as a triple \( (X, \Omega, p) \). Here \( X \), the sample space, is the set of all possible samples, \( x \), that could occur in a particular performance of \( E \). \( \Omega \), the parameter space, is the set of all possible values of a state of nature \( \theta \). And \( p = p(x|\theta) \), the probability function, is a map \( p: X \times \Omega \to [0,1] \), such that \( \sum_{x \in X} p(x|\theta) = 1 \) for all \( \theta \in \Omega \).

When an experiment \( E = (X,\Omega,p) \) has been performed, resulting in the value \( x \in X \), then the function \( \theta \mapsto p(x|\theta) \) is called the likelihood function generated by the data \( (E,x) \), we will denote this by \( L(\theta|x) \). Note, that each likelihood function, \( L(\theta|x) \) in \( \Omega \), gives rise to an equivalent standardized likelihood function, \( \tilde{L}(\theta|x) \) on \( \Omega \), where
\[ L(\theta|x) = \frac{L(\theta|x)}{\sum_{\theta \in \Omega} L(\theta|x)}. \]

Here, equivalency for likelihood functions, say \( L_1(\theta|x) \) and \( L_2(\theta|x) \), means that there exists a constant \( c \) such that \( L_1(\theta|x) = c \cdot L_2(\theta|x) \) for all \( \theta \in \Omega \). There is one additional definition we need before proceeding, and that is for the notion of similar experiments. The experiments \( E_1 = (X_1, \Omega, p_1) \) and \( E_2 = (X_2, \Omega, p_2) \), with the same parameter space \( \Omega \) are said to be "similar" if there exists a one to one map \( g: X_1 \rightarrow X_2 \) such that

\[ p_1(x_1|\theta) = p_2(gx_1|\theta), \]

for all \( x_1 \in X_1 \) and \( \theta \in \Omega \).

Using these concepts, we are now ready to consider the guidelines that must be followed when trying to quantify the notion of statistical information. Basu (1975) states four principles concerning the recognition of the equivalence of two different bits of information. We shall denote the information in experiment, \( E \), when we observe \( x \) as \( \text{Inf}(E,x) \).

1. **The Invariance Principle:** If \( E_1 = (X_1, \Omega, p_1) \) and \( E_2 = (X_2, \Omega, p_2) \) are similar experiments with \( g: X_1 \rightarrow X_2 \) as a similarity map of \( E_1 \) into \( E_2 \), then

\[ \text{Inf}(E_1,x_1) = \text{Inf}(E_2,x_2) \]
if \( q x_1 = x_2 \).

What this principle really means, is that if two experiments are identical in every respect except the manner in which they label points, then they should both yield the same amount of information.

2. The Sufficiency Principle: If, in the context of an experiment, the statistic \( T \) is sufficient then, for all \( x \in X \) and \( t = T(x) \),

\[
\text{Inf}(E,x) = \text{Inf}(E_{T},t).
\]

Here, \( E_T = (\xi,\Omega,P_T) \), and \( p_T(t|\theta) = \sum_{x \in X_T} p(x|\theta) \). Operationally this means, perform \( E \) and then observe only \( t = T(x) \). This principle then implies that, if the statistic \( T \) is sufficient for the experiment, then observing just \( t \) will give us the same amount of information as if we observed \( x \).

Before we state the next principle, consider an experiment \( E = (X,\Omega,p) \) which is a mixture of two experiments, \( E_i = (X_i,\Omega,P_i) \), \( i = 1,2 \), with known mixture probabilities \( \alpha \) and \( 1 - \alpha \). An outcome of such an experiment \( E \) then would be represented as \( x = (i,x_i) \). Therefore, once \( E \) has been performed and the value \( x = (i,x_i) \) collected, we then have the choice of presenting the data as \( (E,x) \) or \( (E_i,x_i) \). This is the problem the third principle addresses.

3. The Weak Conditionality Principle: If \( E \) is a mixture of \( E_1, E_2 \) as described above, then for any \( i \in \{1,2\} \) and \( x_i \in X_i \),
The point here is that while the sufficiency principle warned against any "post-randomization," this form of the conditionality principle concerns itself with any "pre-randomization" that may be built into the experiment.

Lastly, we have the most important principle, and perhaps the quintessential guideline for defining information.

4. The Likelihood Principle: If the data \((E_1, x_1)\) and \((E_2, x_2)\) generate equivalent likelihood functions on \(\Omega\), then

\[
\text{Inf}(E_1, x_1) = \text{Inf}(E_2, x_2).
\]

What this principle is saying, is that if two experiments lead to similar likelihood functions for \(\theta\), then they should provide us with the same amount of information. It should be noted here that it has been proven that the likelihood principle implies the other three principles. So in defining our measure, it is only essential to show that it satisfies this principle.

Now for a Bayesian, the likelihood principle becomes almost a truism. Because, the Bayesian views the data, or rather its information content as a sort of operator which transforms the pattern of their prior beliefs, \(p(\theta)\), into a new posterior pattern, \(p(\theta|x)\). Recall that
\[ p(\theta | x) = \frac{p(\theta)L(\theta | x)}{\int p(\theta)L(\theta | x)d\theta} . \]

The problem here though is that apart from identifying the information in \((E, x)\) with the likelihood function, the likelihood principle does not tell us how to precisely define such a measure. Rather, it should be viewed as a guideline to follow when attempting to quantify the idea of "statistical information."

Lindley (1956) presented one possible measure of information content. His idea entails comparing the knowledge we have prior to an experiment with that we have after observing our sample data.

Specifically, the amount of information in the prior distribution \(p(\theta)\) with respect to \(d\theta\) is defined as

\[ I_0 = \int p(\theta) \log p(\theta) d\theta . \]

Note here, that for any \(\theta\) such that \(p(\theta) = 0\), we define \(p(\theta) \log p(\theta)\) to be zero.

What this expression actually evaluates is the degree of concentration of the prior distribution. As Lindley (1956) states, "the maximum information, in a statistician's sense, will be obtained when the probability distribution is concentrated on a single value of \(\theta\), and the information will be reduced as the distribution of \(\theta\) spreads."

After the experiment has been performed then, and the value of \(x\) is observed, we obtain the posterior density, \(p(\theta | x)\). The information in this distribution then becomes,
\[ I_1(x) = \int p(\theta|x) \log p(\theta|x) \, d\theta. \]

Again, this measures the concentration of the posterior density.

Therefore, Lindley (1956) defines the amount of information provided by an experiment \( E \), with prior knowledge \( p(\theta) \), when we observe \( x \), as

\[ I(E, p(\theta), x) = I_1(x) - I_0. \]

Note here, that this measure does satisfy Basu's principles. Since having equivalent likelihood functions implies that the posterior densities will be identical, i.e. if \( L(\theta|x_1) = cL(\theta|x_2) \) for all \( \theta \in \Theta \), then

\[ \frac{L(\theta|x_1)g(\theta)}{\int L(\theta|x_1)g(\theta)d\theta} = \frac{cL(\theta|x_2)g(\theta)}{\int cL(\theta|x_2)g(\theta)d\theta} = p(\theta|x_2). \]

Thus, with equivalent likelihoods, \( I_1(x_1) = I_1(x_2) \), and since \( I_0 \) is constant, the likelihood principle will be satisfied.

While on the surface this seems to be a quite reasonable measure of information, the idea of making information synonymous with concentration can sometimes lead to serious problems.

Consider, for example, the case where \( X \) is a normal random variable with unknown mean \( \theta \) and known variance \( \sigma^2 \). Let us use the conjugate normal prior, i.e. \( p(\theta) \sim N(\mu, \tau^2) \). Here,
\[ I_0 = \int p(\theta) \log p(\theta) d\theta = \int \frac{1}{\sqrt{2\pi}} \exp - \frac{1}{2} \left( \frac{(\theta - \mu)^2}{\tau^2} \right) \]

\[ \cdot \left[ - \frac{1}{2} \log 2\pi \tau^2 - \frac{1}{2} \left( \frac{\theta - \mu}{\tau} \right)^2 \right] d\theta \]

\[ = - \frac{1}{2} \log 2\pi \tau^2 - \frac{1}{2} \left( \frac{\tau^2}{\tau^2} \right) = - \frac{1}{2} \log 2\pi \tau^2 - \frac{1}{2} \cdot \]

Also, we have that \( p(\theta|x) \sim N \left( \frac{\sigma^2 + \tau^2 x}{\sigma^2 + \tau^2}, \frac{\sigma^2}{\sigma^2 + \tau^2} \right) \). Therefore,

\[ I_1(x) = \int p(\theta|x) \log p(\theta|x) d\theta = \int \sqrt{\frac{\sigma^2 + \tau^2}{2\pi}} \exp - \frac{\sigma^2 + \tau^2}{2\sigma^2} \left( \frac{\sigma^2 + \tau^2}{\sigma^2} \right)^2 \frac{(\theta - \frac{\sigma^2 + \tau^2 x}{\sigma^2 + \tau^2})^2}{2\sigma^2 \tau} \]

\[ \cdot \left[ - \frac{1}{2} \log \frac{2\pi \sigma^2 \tau^2}{\sigma^2 + \tau^2} - \frac{\sigma^2 + \tau^2}{2\sigma^2 \tau} (\theta - \frac{\sigma^2 + \tau^2 x}{\sigma^2 + \tau^2})^2 \right] d\theta \]

\[ = - \frac{1}{2} \log \frac{2\pi \sigma^2 \tau^2}{\sigma^2 + \tau^2} - \frac{\sigma^2 + \tau^2}{2\sigma^2 \tau^2} \left( \frac{\sigma^2}{\sigma^2 + \tau^2} \right) = - \frac{1}{2} \log \frac{2\pi \sigma^2 \tau^2}{\sigma^2 + \tau^2} - \frac{1}{2} \cdot \]

Thus,

\[ I(E, p(\theta), x) = - \frac{1}{2} \log \frac{2\pi \sigma^2 \tau^2}{\sigma^2 + \tau^2} + \frac{1}{2} \log 2\pi \tau^2 = \frac{1}{2} \log (1 + \tau^2/\sigma^2). \]

The point of this example is that the function used to measure the information in the outcome \( x \) obtained from the experiment, does not involve the \( x \) we have observed. The information function
turns out to be a constant, implying that any one \( x \) is as informative as any other. This seems to contradict our intuitive feeling that observing an \( x \) closer to our prior mean, in which case both our sources of information are in agreement, should be more informative than observing a value which is quite dissonant with regard to our prior beliefs.

Therefore, in the next section we introduce a new measure of information which satisfies Basu's principles, incorporates the concentration of the densities, but is also sensitive to the consonance between our observed \( x \) and prior beliefs.

4.2. A New Measure of Information

Quantification of the amount of statistical information inherent in a sample observation, \( x \), derived from an experiment, \( E \), involves dissection of the experiment into its relevant features.

For a Bayesian, adhering strictly to the likelihood principle, the posterior distribution of the parameter of interest is the primary source of information concerning the experiment. This is because it combines the statistician's prior beliefs with the observed sample in order to arrive at the most current "pattern of beliefs" about the parameter. But, there also is information in the \( x \) itself, regardless of which prior density is used, and this should be reflected in the information measure.

Therefore, we decided that an information measure should combine the information in the posterior density with the information in the
x itself, represented by the standardized likelihood function. If our model seems to be a valid representation of the situation we are considering, i.e. our prior beliefs and sample results are in accordance, then we would consider the posterior density to be the richest source of information. Whereas, if the prior and sample information seem to be in conflict, then we would want to concentrate on the information provided by the standardized likelihood function. Therefore, we decided that incorporating our measure of consonance, i.e. the distance between the posterior and prior densities, into the information function would be a natural way of balancing the information in the posterior and the standardized likelihood.

After investigating numerous possible arrangements of these various factors, we decided that the information in an experiment, E, with prior density \( p(\theta) \), and sample value \( x \), is best represented by,

\[
I(x) = I(E,p(\theta),x) = I(p(\theta|x))(1-d(x)) + I(\tilde{L}(\theta|x))d(x). \tag{4.2.1}
\]

In this expression, \( I(p(\theta|x)) = \int p(\theta|x) \log p(\theta|x)d\theta \), \( I(\tilde{L}(\theta|x)) = \int \tilde{L}(\theta|x) \log \tilde{L}(\theta|x)d\theta \), and \( d(x) = d(p(\theta|x), p(\theta)) \). This is the distance between the prior and posterior densities for \( \theta \), with \( 0 \leq d(x) \leq 1 \).

The idea here is that if the prior and sample information are in accordance, in which case \( d(x) \) will be small, then we consider the posterior density to be a valid representation of our most current information about \( \theta \). Therefore, we would want the information in the posterior to carry the most weight. Antithetically, if the prior and
sample information seem to be in conflict then we would want to down-play the information in the posterior and concentrate on the information in \( x \) that is provided by the likelihood function.

If it turns out that \( d(x) = 0 \), this means that the prior distribution is identical to the posterior distribution, implying that there is no information gained by observing \( x \). In this case, our measure reduces to \( I(P(\theta)) \).

If the other extreme occurs, i.e. \( d(x) = 1 \), then there is probably something incorrect about the way we have specified our prior, in which case we would only consider the information in our likelihood function. And, thus, for any other \( d(x) \), we would use the convex combination defined in (4.2.1).

Three important features of this measure should be mentioned. First of all, it satisfies Basu's principles. Recall that it is only necessary to show that the likelihood principle applies, since the other three will then also be true. To reiterate then, the likelihood principle states that if the results of two experiments, say \((E_1, x_1)\) and \((E_2, x_2)\) generate equivalent likelihood functions, then they should both supply equal amounts of information. We have already shown that if two likelihood functions are equivalent, i.e. \( L(\theta|x_1) = cL(\theta|x_2) \), then the two corresponding posterior densities will be identical. This implies that \( I(p(\theta|x_1)) = I(p(\theta|x_2)) \). Likewise, the standardized likelihood functions will be identical, i.e.
This means that all of Basu’s principles will be satisfied.

Secondly, notice that when $p(\theta)$ is the uniform prior, the posterior distribution of $\theta$ is identical to the standardized likelihood function. In this case, the information measure reduces to the information in the likelihood function, i.e. the distance function cancels out. But, this seems reasonable since if we assume that our prior is noninformative, measuring the degree to which the sample information agrees or conflicts with it does not give rise to any meaningful interpretation.

Lastly, if it turns out that the information functions for the two densities are constants not dependent upon $x$, as in the previously described normal example, then our information measure will become an
inverse function of our measure of consonance. That is, the more informative of two x's will be the one with the minimum d(x) value.

For example, consider the normal distribution with unknown mean \( \theta \) and known variance \( \sigma^2 \), i.e. \( f(x|\theta) \sim N(\theta, \sigma^2) \). Let us again use the conjugate normal prior, \( p(\theta) \sim N(\mu, \tau^2) \). Then, the posterior density will be \( P(\theta|x) \sim N\left(\frac{x^2 + \sigma^2 \mu}{\tau^2 + \sigma^2}, \frac{\sigma^2 \tau^2}{\tau^2 + \sigma^2}\right) \). We have previously shown that \( I(p(\theta|x)) = -\frac{1}{2} \log \frac{2\pi \sigma^2}{\tau^2} - \frac{1}{2} \). And, in this case, the likelihood function and the standardized likelihood function are identical. Therefore, \( I(L(\theta|x)) = -\frac{1}{2} \log 2\pi \sigma^2 - \frac{1}{2} \). Using the Hellinger's scaled distance given in formula (2.2.4), we have

\[
d(x) = 1 - \frac{1}{\sqrt{2\pi} \sigma^2} \frac{(\sigma^2 + \tau^2)^{\frac{1}{4}}}{\tau^2 + \sigma^2} \exp \left( -\frac{1}{4} \left[ \frac{\tau^2 (\mu - x)^2}{(\tau^2 + \sigma^2)^2} \right] \right).
\]

Therefore,

\[
I(E, p(\theta), x) = \left[ -\frac{1}{2} \ln \frac{2\pi \sigma^2}{\tau^2} - \frac{1}{2} \right] \cdot \left[ \frac{1}{\sqrt{2\pi} \sigma^2} \frac{(\sigma^2 + \tau^2)^{\frac{1}{4}}}{\tau^2 + \sigma^2} \exp \left( -\frac{1}{4} \left[ \frac{\tau^2 (\mu - x)^2}{(\tau^2 + \sigma^2)^2} \right] \right) \right] + \left[ -\frac{1}{2} \ln 2\pi \sigma^2 - \frac{1}{2} \right]
\]

\[
[1 - \frac{1}{\sqrt{2\pi} \sigma^2} \frac{(\sigma^2 + \tau^2)^{\frac{1}{4}}}{\tau^2 + \sigma^2} \exp \left( -\frac{1}{4} \left[ \frac{\tau^2 (\mu - x)^2}{(\tau^2 + \sigma^2)^2} \right] \right)]
\]
Unlike Lindley's measure then, this function distinguishes between the various $x$'s in the sample space. The implication here is that if we observe an $x$ that is the same as our prior mean for $\theta$, then all our sources of information are in total agreement, and the information measure will be maximum. And, as $x$ moves farther away from $\mu$, denoting a degree of discrepancy in our model, the information measure decreases.

Therefore, while our measure recognizes the importance of the concentration of the densities involved, it is also sensitive to the location of the densities. That is, if the prior and posterior densities are both highly concentrated, but about different regions of the parameter space, then this will be signaled by a large value for the distance function.

Until now, we have considered experiments which yield a sample observation, $x$. But obviously, formula (4.2.1) can be generalized for the case of observing a sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. That is, we define the information from an experiment $E$, with prior density $p(\theta)$, which generates a sample $\mathbf{x}$ to be:

$$
-\frac{1}{2} \ln 2\pi \sigma^2 - \frac{1}{2} + \frac{1}{2} \ln \left( \frac{\sigma^2 + \tau^2}{\sigma^2 + \tau^2} \right) \left( \frac{\sigma^4 + \sigma^2 \tau^2}{\sigma^4 + \sigma^2 \tau^2} \right) x
$$

$$
\exp -\frac{1}{4} \left( \frac{\tau (u-x)^2}{\tau^2 + \sigma^2} \right) \left( \frac{2 \sigma^2 \tau^2 + \tau^4}{\tau^2 + \sigma^2} \right).
$$
\[ I(x_i) = I(E, p(\theta), x_i) = I(p(\theta|x_i))(1-d(x_i)) + I(\bar{L}(\theta|x_i))d(x_i). \quad (4.2.2) \]

As before, \( I(p(\theta|x_i)) = \int p(\theta|x_i) \log p(\theta|x_i) d\theta, I(\bar{L}(\theta|x_i)) = \int \bar{L}(\theta|x_i) \log \bar{L}(\theta|x_i) d\theta, \) and \( d(x_i) = d(p(\theta|x_i), p(\theta)). \)

Obviously, this measure also satisfies Basu's principles. And, all motivational discourse and salient properties previously discussed are still relevant.

One further question of information that came to mind was how we could, given a sample \( x_i, \) evaluate the amount of information in each \( x_i, \) relative to our \( x. \) That is, could we use our information measure to establish some sort of ordering among the \( x_i's, \) in terms of which was most informative, which was least informative, etc.

To this end, let us define the amount of information in an observation \( x_i, \) from an experiment \( E, \) with prior density \( p(\theta), \) and sample observations \( x, \) to be

\[ I(x_i) = I(E, p(\theta), x_i) = I(p(\theta|x_i))(1-d(x_i)) + I(\bar{L}(\theta|x_i))d(x_i). \quad (4.2.3) \]

Here, \( I(p(\theta|x)) \) and \( I(\bar{L}(\theta|x_i)) \) are as defined above, but \( d(x_i) = d(p(\theta|x_i), p(\theta|x)) \) is the distance between the posterior distribution of \( \theta \) given the full sample, \( x, \) and the posterior density given all observations except \( x_i. \) This measure is essentially a generalization of formula (4.2.1) since here the deleted posterior distribution represents our most current beliefs about \( \theta \) prior to observing \( x_i. \)
The idea here is that if \( d(x_i) \) is very large, implying that the observation \( x_i \) is incongruous with the rest of the sample, then we would want to discount the posterior density which incorporates this value, and focus primarily upon the standardized likelihood function. If \( x_i \) does seem to be in accordance with the rest of the sample, then we would want our information measure to be principally a function of the full posterior distribution.

As in our previous discussion, when the information in the full posterior and standardized likelihood is constant, then formula (4.2.3) will be a reflection of the behaviour of the distance function.

By way of example, if each \( x_i \) in our sample \( x \) is normally distributed, i.e. \( f(x_i|\theta) \sim N(\theta, \sigma^2) \), and \( p(\theta) \sim N(\mu, \tau^2) \), then

\[
p(\theta|x) \sim N\left(\frac{n \tau x + \sigma^2 \mu}{n \tau^2 + \sigma^2}, \frac{\sigma^2}{n \tau^2 + \sigma^2}\right), \quad \text{and} \quad p_i(\theta|x) \sim N\left(\frac{(n-1) \tau x_i + \sigma^2 \mu}{(n-1) \tau^2 + \sigma^2}, \frac{\sigma^2}{(n-1) \tau^2 + \sigma^2}\right),
\]

where \( \bar{x}_{(i)} = \frac{1}{n} \sum_{j \neq i} x_j/n-1 \). Here, we find that

\[
I(p(\theta|x)) = -\frac{1}{2} \log \frac{2\sigma^2 \tau^2}{\sigma^2 + n \tau^2} - \frac{1}{2} \quad \text{and} \quad I(p(\theta|x)) = -\frac{1}{2} \log \frac{2\sigma^2 \tau^2}{\sigma^2 + n \tau^2} - \frac{1}{2}.
\]

Also, using equation (3.1.1) we have,

\[
d(x_i) = 1 - \frac{\sqrt{2[(\sigma^2 + n \tau^2)(\sigma^2 + (n-1) \tau^2)]^2}}{\sqrt{2(\sigma^2 + \tau^2)(2n-1)}} \exp -\frac{1}{4}\left[\frac{\tau^2(n-1)(\bar{x}_{(i)} - x_i)}{\sigma^2 + (n-1) \tau^2 + \sigma^2[(\mu - x_i)^2]}\right].
\]
For convenience, let us define the following constants:

\[ c_1 = \frac{1}{\sqrt{2\pi \tau^2 (2n-1)}} \left( \frac{\sigma^2 + (n-1)\tau^2}{\sigma^2 + \tau^2} \right)^{\frac{1}{2}}, \quad c_2 = \frac{\tau^2}{\sigma^2 (n-1)\tau^2 + \sigma^2 (n\tau^2 + \sigma^2)} \left( \frac{\sigma^2 + (n-1)\tau^2}{\sigma^2 + \tau^2} \right) \]

Therefore, we find that,

\[ I(x_i) = \left[ -\frac{1}{2} \log \frac{2\pi \sigma^2}{2} - \frac{1}{2} \right] \sqrt{2} c_1 \exp \left\{ -\frac{1}{4} c_2 (\tau^2 (n-1) - \tau^2 - \tau^2 (2n-1)) \right\} \]

\[ + \left[ -\frac{1}{2} \log \frac{2\pi \sigma^2}{2} - \frac{1}{2} \right] [1 - \sqrt{2} c_1 \exp \left\{ -\frac{1}{4} c_2 (\tau^2 (n-1) - \tau^2 (2n-1)) \right\}] \]

\[ + \exp \left\{ -\frac{1}{4} c_2 (\tau^2 (n-1) - \tau^2 (2n-1)) \right\} \]

This implies that the information function attains its maximum when the next observation sampled is equal to the posterior mean of \( \theta \) given all previous observations, i.e. when \( x_i = \frac{(\tau^2 (n-1)-\tau^2 + \tau^2)}{2c_1} \). And, as \( x_i \) moves away from this mean, the function decreases.

In conclusion, we have introduced a new measure of the amount of statistical information given by an observation \( x \). This measure is a convex combination of the information in the posterior distribution and the information in \( x \) itself. The corresponding weights are given...
by an appropriate measure of the consonance between the sample observations and prior beliefs. Thus, while emphasizing the concentration of the densities, as Lindley's measure does, this function also takes into account the position of the densities relative to one another. This insures that even when the information in the densities is constant, the measure will still be a function of the $x$ we observe. And lastly, it satisfies the guidelines presented by Basu for defining equivalent pieces of information.
5. A PREDICTION PROBLEM

In this final chapter, we consider an interesting application of our distance measure. Suppose we have the model \( y = X\beta + \epsilon \), where as in Chapter 3, \( \epsilon \) is an \( Nx1 \) random vector following a multivariate normal distribution, i.e. \( \epsilon \sim \mathcal{N}(0,\sigma^2 I) \). We assume \( \sigma^2 \) is known. As before, \( \beta \) is a \( px1 \) vector of regression coefficients, \( X \) is an \( Nxp \) matrix of fixed independent variables, and \( y \) is an \( Nx1 \) vector of dependent variables. Suppose that our goal is to predict a future \( y \) value where we assume we are sampling this new \( y \) from the same model as generated our original data set.

We begin by having available to us \( N \) observations on \( p \) independent variables. What we want to decide is which, if not all, of these variables do we need to include in order to best predict a future \( y \).

For convenience, let \( K \) denote a subset of the integers \( 1,2,\ldots,p \) containing \( s \) members and let \( J \) denote its complement. Then let \( X_K \) denote the \( Nxs \) matrix consisting of those \( x_i \)'s whose indices are contained in \( K \). Likewise, \( X_J \) contains those whose indices are in \( J \).

What we wish to do, is to decide which subset \( K \) of our regression coefficients we should use in the prediction process. The method we suggest involves calculating first the predictive density of \( y \) using all available variables. Then, for each fixed \( s \) value,
calculate the predictive density of \( y \) using each subset \( K \) of parameters. Next, measure the distances between each of these "reduced" predictive densities, and the predictive density which includes the full data set \((X,y)\).

The idea here is that for a fixed number of variables, we would decide to predict a future \( y \) value using the subset \( K \) whose predictive density is "closest" to the full predictive density, i.e. whose distance measure is minimum.

We begin by considering the case where the prior distribution of \( \beta \) and \( \sigma^2 \) is the noninformative uniform prior, i.e. \( p(\beta, \sigma^2) = \frac{1}{\sigma^2} \). Then, as derived in Chapter 3, the marginal posterior distribution of \( \beta \), given \( \sigma^2 \) and \( y \), will be \( p(\beta | \sigma^2, y) \sim MN_p(\hat{\beta}, \sigma^2(X'X)^{-1}) \), with \( \hat{\beta} = (X'X)^{-1}X'y \). Therefore, the predictive density of the vector \( y \) is, \( y \sim MN_m(X\hat{\beta}, \sigma^2(I+X(X'X)^{-1}X')) \), where \( X \) is the original matrix of observed independent variables.

We need to derive the predictive density of the reduced model estimate. But first, we must find the posterior distribution of \( \beta_K \), i.e. the subset of regression coefficients whose indices are contained in the set \( K \). Therefore, we will integrate the posterior distribution of \( \beta \) with respect to the remaining coefficients, \( \beta_J \).

Thus, we have

\[
p(\beta_K | \sigma^2, y) = \int (2\pi \sigma^2)^{-P/2} |X'X|^{1/2} \exp \left( -\frac{1}{2\sigma^2}((\beta - \hat{\beta})'X'X(\beta - \hat{\beta})) \right) d\beta_J
\]
Here, we have partitioned the vectors $\beta$ and $\hat{\beta}$, and the matrix $X'X$ as follows:

$$
\beta = \begin{pmatrix} \beta_k \\ \beta_J \end{pmatrix}, \quad \hat{\beta} = \begin{pmatrix} \hat{\beta}_k \\ \hat{\beta}_J \end{pmatrix}, \quad \text{and} \quad X'X = \begin{pmatrix} X'X_k & X'X_J \\ X'X_k & X'X_J \end{pmatrix}.
$$

Therefore, (5.1.1) reduced to

$$(2\pi\sigma^2)^{-s/2} \exp \left\{-\frac{1}{2\sigma^2} [ (\beta_k - \hat{\beta}_k)'X_k'X_k(\beta_k - \hat{\beta}_k) + \hat{\beta}_J'X_J'X_J\hat{\beta}_J - 2(\beta_k - \hat{\beta}_k)'X_k'X_J(\hat{\beta}_J'X_J'X_J)^{-1} \hat{\beta}_J ] \right\}$$

$$
\times \int (2\pi\sigma^2)^{-\frac{p-s}{2}} |X'X|^\frac{1}{2} \exp \left\{-\frac{1}{2\sigma^2} [\hat{\beta}_J'X_J'X_J\hat{\beta}_J - 2(\hat{\beta}_J - (\beta_k - \hat{\beta}_k)'X_k'X_J(\hat{\beta}_J'X_J'X_J)^{-1}) \hat{\beta}_J ] \right\} d\beta_J.
$$

Next, let $\hat{\gamma}' = \hat{\beta}_J - (\beta_k - \hat{\beta}_k)'X_k'X_J(\hat{\beta}_J'X_J'X_J)^{-1}$, this implies that the above is equivalent to,

$$(2\pi\sigma^2)^{-s/2} \exp \left\{-\frac{1}{2\sigma^2} [ (\beta_k - \hat{\beta}_k)'X_k'X_k(\beta_k - \hat{\beta}_k) + \hat{\beta}_J'X_J'X_J\hat{\beta}_J \right.$$}

$$
- 2(\beta_k - \hat{\beta}_k)'X_k'X_J\hat{\beta}_J - \hat{\gamma}'X_J'X_J\hat{\beta}_J \right\}.$$
Noting then that the integral is equal to unity, and

\[ |X'X| = |X'X_j| \cdot |X'X_k - X'X_j (X'X_j)^{-1}X'X_k|, \]

after some simplification we find that formula (5.1.1) becomes,

\[
(2\pi \sigma^2)^{-s/2} \left| X'X_k - X'X_j (X'X_j)^{-1}X'X_k \right|^{1/2} \exp \left\{ -\frac{1}{2\sigma^2} \left[ (\beta_k - \hat{\beta}_k)'(X'X_k - X'X_j (X'X_j)^{-1}X'X_k)(\beta_k - \hat{\beta}_k) \right] \right\}
\]

\[
X'X_j (\beta_k - \hat{\beta}_k) \right\} . \tag{5.1.2}
\]

Therefore, if we define \( v_K = X'X_k - X'X_j (X'X_j)^{-1}X'X_k \), we see that \( p(\beta_k, \sigma^2, y) \sim N_S(\hat{\beta}_k, \sigma^2 v_K^{-1}) \). This means that our predicted vector under the reduced model is \( \hat{y}_K = X' \hat{\beta}_K \).

Note here that \( \hat{\beta}_K \) is not the least squares estimate for the reduced model, rather it is a subset of the \( \hat{\beta} \) vector calculated under the full model. Following Bayesian procedures, with squared error loss, we use as our estimate of \( \beta_K \) the posterior mean of the marginal posterior distribution.

In order to calculate the distance measure we must evaluate the predictive density of \( y \) under the reduced model. This will be accomplished by considering the following integral:

\[
\int p(y/\beta_K, \sigma^2) \cdot p(\beta_K/\sigma^2, y) \, d\beta_K.
\]
This becomes,

\[
\int p(y/x_k, \sigma^2) p(\beta_k/\sigma^2, y) \, d\beta_k \propto \int \exp - \frac{1}{2\sigma^2} \left[ (y - x_k \beta_k)' (y - x_k \beta_k) \right. \\
\left. + (\beta_k - \hat{\beta}_k)' V_k (\beta_k - \hat{\beta}_k) \right] \, d\beta_k . \tag{5.1.3}
\]

Simplifying the quantity in the exponent, we have

\[
(y - x_k \beta_k)' (y - x_k \beta_k) + (\beta_k - \hat{\beta}_k)' V_k (\beta_k - \hat{\beta}_k) = y'y - 2\beta_k'X_k y + \beta_k'X_k \beta_k + V_k \beta_k - 2\beta_k'V_k \hat{\beta}_k + \hat{\beta}_k'V_k \hat{\beta}_k
\]

\[
= y'y + \hat{\beta}_k'V_k \hat{\beta}_k + \beta_k'(X_k'X_k + V_k) \beta_k - 2\beta_k'(X_k'X_k + V_k) \hat{\beta}_k .
\]

Let us define \( \Gamma = (X_k'X_k + V_k)^{-1} (X_k'X_k + V_k) \), then the previous formula can be reexpressed as,

\[
y'y + \hat{\beta}_k'V_k \hat{\beta}_k - \Gamma'(X_k'X_k + V_k) \Gamma + (\beta_k - \Gamma)'(X_k'X_k + V_k)(\beta_k - \Gamma) .
\]

Thus, (5.1.3) reduces to

\[
\exp - \frac{1}{2\sigma^2} [y'y + \hat{\beta}_k'V_k \hat{\beta}_k - \Gamma'(X_k'X_k + V_k) \Gamma] \cdot \int \exp - \frac{1}{2\sigma^2} [(\beta_k - \Gamma)'(X_k'X_k + V_k) \Gamma] \cdot (\beta_k - \Gamma)] \, d\beta_k \propto \exp - \frac{1}{2\sigma^2} [y'y + \hat{\beta}_k'V_k \hat{\beta}_k - \Gamma'(X_k'X_k + V_k) \Gamma] . \tag{5.1.4}
\]

Expanding the exponent in formula (5.1.4), we have
\[
\begin{align*}
y'\gamma + \beta_k^* V_k^* \beta_k & = (y'X_k^* + \beta_k^* V_k^*) (X_k^* X_k + V_k)^{-1} (X_k^* y + V_k^* \beta_k) \\
& = y'\gamma + \beta_k^* V_k^* \beta_k - y'X_k (X_k^* X_k + V_k)^{-1} x_k^* \gamma - 2 \beta_k^* V_k (X_k^* X_k + V_k)^{-1} x_k^* y \\
& \quad - \beta_k^* V_k (X_k^* X_k + V_k)^{-1} y_k^* \beta_k \\
& \quad = y'(I - X_k (X_k^* X_k + V_k)^{-1} x_k^*) \gamma - 2 \beta_k^* V_k (X_k^* X_k + V_k)^{-1} x_k^* y \\
& \quad + \beta_k^* V_k (I - (X_k^* X_k + V_k)^{-1} V_k) \beta_k^* \\
& \quad = y'(I - X_k (X_k^* X_k + V_k)^{-1} x_k^*) \gamma \\
& \quad - 2 \beta_k^* V_k (X_k^* X_k + V_k)^{-1} x_k^* y \\
& \quad + \beta_k^* V_k (I - (X_k^* X_k + V_k)^{-1} V_k) \beta_k^* .
\end{align*}
\]

We now need the identity given in Chapter 3 which states that,

\[(X_k^* X_k + V_k)^{-1} x_k^* = V_k^{-1} x_k^* (I + X_k V_k V_k^{-1} x_k^*)^{-1} .\]  

This implies that

\[I - X_k (X_k^* X_k + V_k)^{-1} x_k^* = I - X_k V_k^{-1} x_k^* (I + X_k V_k V_k^{-1} x_k^*)^{-1} .\]  

And, since this second quantity is identical to \((I + X_k V_k V_k^{-1} x_k^*)^{-1}\), the first term in the expansion becomes, \(y'(I + X_k V_k V_k^{-1} x_k^*)^{-1} \gamma\).

Premultiplying both sides of the previous identity by \(V_k\), we find that \(V_k (X_k^* X_k + V_k)^{-1} x_k^* = x_k^* (I + X_k V_k V_k^{-1} x_k^*)^{-1} .\) Therefore, the second terms reduces to \(2 \beta_k^* V_k (I + X_k V_k V_k^{-1} x_k^*)^{-1} .\) And lastly, noting that

\[(I - (X_k^* X_k + V_k)^{-1} V_k) = (X_k^* X_k + V_k)^{-1} x_k^* X_k,\]  

we see that \(V_k (I - (X_k^* X_k + V_k)^{-1} V_k)\)

\[= V_k (X_k^* X_k + V_k)^{-1} x_k^* X_k = x_k^* (I + X_k V_k V_k^{-1} x_k^*)^{-1} x_k.\]  

Thus, equation (5.1.5) can be written as

\[(y - X_k \beta_k) (I + X_k V_k V_k^{-1} x_k^*)^{-1} (y - X_k \beta_k^*).\]
Therefore, the predictive density of \( y \) using a subset \( K \) of the parameters is \( MN_n(X_K \hat{\beta}_K, \sigma^2(I + X_K X_K^{-1} X_K')) \).

Before proceeding to calculations of the distances, let us consider how evaluating the significance of subsets of independent variables is typically handled. The classical statistician considers what is known as the reduction in sum of squares due to \( \beta_J \) after \( \beta_K \) has been included. We denote this by \( R(\beta_J/\beta_K) \). This is the difference between the residual sum of squares under the reduced model, i.e. \((y-X_K y)^\prime(y-X_K y)\) with \( y = (X_K X_K) X_K^{-1} y \), and the residual sum of squares using the full model, i.e. \((y-X y)^\prime(y-X y)\) with \( \hat{\beta} = (X X)^{-1} X y \).

It is well known that \( R(\beta_J/\beta_K) \) can be expressed as follows:

\[
R(\beta_J/\beta_K) = \hat{\beta}' X' X \hat{\beta} - \hat{\gamma}' X' X \hat{\gamma}
\]

\[
= \hat{\beta}' X' X \hat{\beta} - y' X' (X' X)^{-1} X' y .
\]

(5.1.6)

This difference is then divided by the residual sum of squares for the full model, and the ratio is compared to an F statistic with appropriate degrees of freedom. The point here being that if the amount of variation in \( y \) explained by including the variables in set \( J \) is minimal, then we would be justified in considering those variables to be nonsignificant.

If we partition the vector \( \hat{\beta} \) and the matrix \( X'X \) as before, i.e.
\[ \hat{\beta} = \begin{pmatrix} \hat{\beta}_K \\ \hat{\beta}_J \end{pmatrix}, \quad x'x = \begin{pmatrix} x'x_K & x'x_J \\ x'x_K^T & x'x_J^T \end{pmatrix}, \]

we see that equation (5.1.6) can be written as,

\[ R(\hat{\beta}_J/\hat{\beta}_K) = \hat{\beta}_J'x'x\hat{\beta}_K + \hat{\beta}_J'x'x\hat{\beta}_J + \hat{\beta}_K'x'x\hat{\beta}_J + \hat{\beta}_J'x'x\hat{\beta}_J \\
- y'x_K(x'x_K)^{-1}x'_Ky. \quad (5.1.7) \]

Recall the classical normal equations, \( x'\hat{\beta} = x'y \). Partitioning the components of this identity in an analogous manner gives us the following identities:

\[ x'x\hat{\beta}_K' + x'x\hat{\beta}_J = x'y \]

\[ x'x\hat{\beta}_K' + x'x\hat{\beta}_J = x'_Ky. \]

Substituting the first identity into the last term of equation (5.1.7) gives us

\[ R(\hat{\beta}_J/\hat{\beta}_K) = \hat{\beta}_J'x'x\hat{\beta}_K + \hat{\beta}_J'x'x\hat{\beta}_J + \hat{\beta}_K'x'x\hat{\beta}_J + \hat{\beta}_J'x'x\hat{\beta}_J \\
- (\hat{\beta}_K'x'x\hat{\beta}_K + x'x_ \hat{\beta}_J''_K + x'x_ \hat{\beta}_J''_J) \]

\[ = \hat{\beta}_J'(x'x_J - x'x_K(x'x_K)^{-1}x'_Ky) \hat{\beta}_J \\
= \hat{\beta}_J'(I - x'x_K(x'x_K)^{-1}x'_Ky) \hat{\beta}_J. \quad (5.1.8) \]
The reason for expressing \( R(\beta_j/\beta_K) \) in this form will soon become evident.

Now we calculate the Kullback-Liebler distance between the "full" predictive density and the "reduced" predictive density using formula (3.2.6). Here \( \mu_1 = \hat{\gamma}_F = X\hat{\beta}, \mu_2 = \hat{\gamma}_R = X_{K}\hat{\beta}_K, \Sigma_1 = \sigma^2(I + X(x'x)^{-1}x') \Sigma_2 = \sigma^2(I + X_{K}X_{K}^{-1}x'). \) Therefore, we have

\[
2\hat{d}(p_F, p_R) = \frac{1}{\sigma^2} (\hat{\gamma}_F - \hat{\gamma}_R)' (I + X_{K}X_{K}^{-1}x')^{-1} (\hat{\gamma}_F - \hat{\gamma}_R) + \text{tr}[(I + X(x'x)^{-1}x')(I + X_{K}X_{K}^{-1}x')^{-1}]
\]

\[- \ln \frac{|I + X(x'x)^{-1}x'|}{|I + X_{K}X_{K}^{-1}x'|} - N. \tag{5.1.9}\]

First, we consider the initial term in expression (5.1.9), i.e.

\[
\frac{1}{\sigma^2} (\hat{\gamma}_F - \hat{\gamma}_R)' (I + X_{K}X_{K}^{-1}x')^{-1} (\hat{\gamma}_F - \hat{\gamma}_R). \text{ Note here that } \hat{\gamma}_F - \hat{\gamma}_R = X\hat{\beta} - X_{K}\hat{\beta}_K = X_{K}\hat{\beta}_K + x\hat{\beta}_J - X_{K}\hat{\beta}_K = x\hat{\beta}_J. \text{ Therefore, this quantity may be written as } \frac{1}{\sigma^2} x\hat{\beta}_J (I + X_{K}X_{K}^{-1}x')^{-1} x\hat{\beta}_J. \]

Next, recall that in simplifying equation (5.1.5) we used the fact that \((I + X_{K}X_{K}^{-1}x')^{-1} = (I - X_{K}(X_{K}X_{K} + V_{K})^{-1}x'). \) Thus, we can express the leading factor as,

\[
\frac{1}{\sigma^2} x\hat{\beta}_J (I - X_{K}(X_{K}X_{K} + V_{K})^{-1}x') x\hat{\beta}_J. \tag{5.1.10}\]
Next, let us substitute the identity 

\[(X'K'K + V)^{-1} = (X'K'K)^{-1} - (X'K'K + V)^{-1}v'_{\text{K}}(X'K'K)^{-1}
\]

into (5.1.10). This gives,

\[
\frac{1}{\sigma^2} \beta^J J^\top (I - X'K'K)^{-1}x' + X'K'K + V)^{-1}v'_{\text{K}}(X'K'K)^{-1}x' + X'K'K)^{-1}x' J^\top \beta J.
\]

Using then the form of \(R(\beta / J^\top \beta^J K)^{-1}\) given by (5.1.8), the previous expression becomes

\[
\frac{1}{\sigma^2} [R(\beta / J^\top \beta^J K) + \beta^J J^\top (I + X'K'K + V)^{-1}v'_{\text{K}}(X'K'K)^{-1}x' + X'K'K)^{-1}x' J^\top \beta J]
\]

\[
= \frac{1}{\sigma^2} [R(\beta / J^\top \beta^J K) + \beta^J J^\top (I + X'K'K + V)^{-1}x' + X'K'K)^{-1}x' J^\top \beta J].
\]

(5.1.11)

This last simplification was accomplished by noting that

\[X'K'K + V)^{-1} = (I + X'K'K)^{-1}x' + X'K'K)^{-1}x'.
\]

To complete our analysis of the distance function given by formula (5.1.9), let us consider

\[\ln |I + X(X'X)^{-1}x' + X'K'K + V)| - \ln |I + X'K'K + V|.
\]

Note here that

\[|I + X(X'X)^{-1}x' + X'K'K + V)| = |I + X'X(X'X)^{-1}| = |2I| = 2^p,\]

therefore, this term becomes, \(p \ln 2 - \ln |I + X'K'K + V|\). Thus, equation (5.1.9) simplifies to,

\[
\frac{1}{\sigma^2} [R(\beta / J^\top \beta^J K) + \beta^J J^\top (I + X'K'K + V)^{-1}x' + X'K'K)^{-1}x' J^\top \beta J]
\]

\[+ \text{tr}((I + X(X'X)^{-1}x')(I + X'K'K + V)^{-1}) + \ln |I + X'K'K + V|\]
If we calculate the Hellinger's distance between the full and predictive densities, using formula (3.2.5) with \( \mu_1, \mu_2, \Sigma_1, \) and \( \Sigma_2 \) as defined before, we find that,

\[
\tilde{d} = 1 - \frac{2^{N/2} |I + X_kK^{-1}x_k^i|^4 |I + X(X'X)^{-1}x'|^2}{|2I + X(X'X)^{-1}x' + xK^{-1}x_k^i|^2} \exp \left\{ -\frac{1}{4} (\tilde{y}_P - \tilde{y}_R)' (I + X(X'X)^{-1}x' + I + xK^{-1}x_k^i)^{-1} (\tilde{y}_P - \tilde{y}_R) \right\}.
\]

(5.1.13)

Considering first the exponent, and noting that \((\Sigma_1 + \Sigma_2)^{-1} = \Sigma_2^{-1} - (\Sigma_1 + \Sigma_2)^{-1} \Sigma_1 \Sigma_2^{-1}\), we have,

\[
\hat{\beta}_J^i X_k^i ((I + x_kK^{-1}x_k^i)^{-1} - (2I + X(X'X)^{-1}x' + xK^{-1}x_k^i)^{-1} (I + X(X'X)^{-1}x'))
\]

\[
(I + xK^{-1}x_k^i)^{-1} x_k^i \hat{\beta}_J^i
\]

\[
= \hat{\beta}_J^i X_k^i (I - X_k^i (X'X)^{-1}x_k^i) ^{-1} x' - (2I + X(X'X)^{-1}x' + xK^{-1}x_k^i)^{-1} (I + X(X'X)^{-1}x')
\]

\[
(I + xK^{-1}x_k^i)^{-1} x_k^i \hat{\beta}_J^i
\]

\[
= R(\beta_J / \beta_k) - \hat{\beta}_J^i X_k^i (2I + X(X'X)^{-1}x' + xK^{-1}x_k^i)^{-1} (I + X(X'X)^{-1}x')
\]

\[
(I + xK^{-1}x_k^i)^{-1} x_k^i \hat{\beta}_J^i.
\]
Then since $|I + X(X'X)^{-1}X'| = 2^p$, formula (5.1.13) becomes,

$$1 - \frac{2^{N/2+p/4}|I+X_K V_{K-K}^{-1} X|^{1/4}}{|2I+X_K V_{K-K}^{-1} X + X(X'X)^{-1} X'|^{1/2}} \exp - \frac{1}{4\sigma^2} [R(\beta_J/\beta_K)]$$

$$- \beta_J X_J' (2I+X(X'X)^{-1}X'+X V_{K-K}^{-1} X_{K-K})^{-1} (I+X(X'X)^{-1} X') (I+X V_{K-K}^{-1} X_{K-K})^{-1}$$

$$\times X_J \beta_J]. \quad (5.1.14)$$

Therefore, we have shown that in the case of the uniform prior both distance functions, i.e. (5.1.12) and (5.1.14), are functions of the classical statistic, $R(\beta_J/\beta_K)$, used for deciding the significance of subsets of the independent variables. We conjecture that the distance measures are monotonic functions of the reduction in sum of squares, but as yet this has not been proven. Were this to be true though, then in the case of the uniform prior, we would have a Bayesian justification for the use of the classical statistic, $R(\beta_J/\beta_K)$, in deciding the significance of parameters in the prediction problem, i.e. this value could thus be viewed as an approximation to the distance function.

If we use the conjugate normal prior, then as discussed in Chapter 3, the posterior distribution of $\beta$ will be, $\beta \sim MN_p(\betahat, \sigma^2(\Sigma^{-1} + X'X)^{-1})$, where $\betahat = (\Sigma^{-1} + X'X)^{-1}(\Sigma^{-1}u + X'y)$. Therefore, the predictive distribution of a future $y$ value, where we include the entire data set, will be $y \sim MN_N(x_{\betahat}, \sigma^2(I + X(\Sigma^{-1} + X'X)^{-1}X'))$. 
Then, to find the marginal posterior density of $\beta_K$, we again integrate the prior covariance matrix as follows:

$$
\Sigma = \begin{pmatrix}
\Sigma_{KK} & \Sigma_{KJ} \\
\Sigma_{JK} & \Sigma_{JJ}
\end{pmatrix}
$$

If we define $\tilde{V}_K = (\Sigma_{KK} + X'KX) - (\Sigma_{JK} + X'JX)^{-1}(\Sigma_{JJ} + X'JX)^{-1}$, then it follows that the predictive density under the reduced model is $y \sim MN_{N}(X'K\beta_K', \sigma^2(I + X'K\tilde{V}_K'^{-1}X')^{-1})$.

The Kullback-Liebler distance between the full and reduced model predictive densities is,

$$
\frac{1}{2} \ln \left| \frac{I + X'(\Sigma^{-1} + X'X)^{-1}X'}{I + X'K\tilde{V}_K'^{-1}X'} \right| - \frac{N}{2}
$$

$$
= \frac{1}{2} \ln \left( I + X'K\tilde{V}_K'^{-1}X' \right) - \frac{N}{2}
$$

$$
= \frac{1}{2} \ln \left( I + X'(\Sigma^{-1} + X'X)^{-1}X' \right) - \frac{N}{2}
$$

$$
= \frac{1}{2} \ln \left| \frac{I + X'K\tilde{V}_K'^{-1}X'}{I + X'K\tilde{V}_K'^{-1}X'} \right| - \frac{N}{2}.
$$

(5.1.15)
The Hellinger's distance between the two densities will be,

\[
1 - \frac{2^{N/2} |I + \mathbf{x}' \mathbf{V}_K^{-1} \mathbf{x}_K | |I + \mathbf{x}' (\Sigma^{-1} + \mathbf{x}' \mathbf{x})^{-1} \mathbf{x}'|^{1/2}}{|2I + \mathbf{x}' (\Sigma^{-1} + \mathbf{x}' \mathbf{x})^{-1} \mathbf{x}' + \mathbf{x}_K \mathbf{V}_K^{-1} \mathbf{x}_K |^{1/2}}
\]

\[x \exp \left\{ - \frac{1}{4\sigma} (\tilde{\beta}' \mathbf{x}_J' (2I + \mathbf{x}' (\Sigma^{-1} + \mathbf{x}' \mathbf{x})^{-1} \mathbf{x}' + \mathbf{x}_K \mathbf{V}_K^{-1} \mathbf{x}_K)^{-1} \mathbf{x}_J \tilde{\beta}_J \right\}. \quad (5.1.16)\]

There is no analogue here for the classical statistic \( R(\beta_J/\beta_K) \).

For in the least squares model, the total sum of squares can be factored into the model sum of squares and residual sum of squares, i.e.

\[y'y = (y - \mathbf{x}\hat{\beta})'(y - \mathbf{x}\hat{\beta}) + \hat{\beta}'x'y.\]

But, the Bayesian model does not lend itself to a similar identity. Therefore, we will use the distances as given in formulas (5.1.15) and (5.1.16).

To recapitulate, we have derived distance measures between the full and reduced predictive densities in the cases of the uniform and conjugate priors. The procedure we suggest then is that after an appropriate prior density and distance function have been decided upon, calculate the distance between the predictive density using all the variables, and each possible subset of variables. Now, note that the distances will decrease as the number of variables included increases. Therefore, in order to make comparisons between models of different sizes we will need to include an additional criterion.
One possible solution would be to assign a cost $c_i$ to each independent variable we consider including. In some cases, these values may follow naturally from the experiment being performed. For example, in an industrial situation, we may have to expend a specific amount of money in order to observe each variable. But, if this is not the case, we could assign relative costs to the variables in terms of the degree of difficulty involved in obtaining them. Or finally, we could consider including a complicating cost for each additional variable.

Therefore, for each subset $K$ of variables we calculate the quantity $d_K + c_K$. Here $d_K$ is the distance between the predictive density including the variables in set $K$ and the predictive density including all the variables. And, $c_K$ is the sum of the costs of the variables in $K$. The analyst should be aware though, that the costs should be scaled so that the total cost of including all the variables does not exceed the maximum distance, i.e. the distance calculated when no variables are included. Because in this case, you would always decide that it did not pay to observe any variables. This guideline noted then, what we are suggesting is to compute $d_K + c_K$ for every subset $K$ of sizes $s = 0, 1, ..., p$, then to predict a future $y$ value using the appropriate estimate from that subset, $K$, which minimizes this sum.

This idea of assigning costs to variables was introduced by Lindley (1968) as a key factor in a loss function useful in deciding
which subset of the independent variables should be included in the prediction problem. Specifically, if $y$ is the true value of the dependent variable, $f(X_k)$ is the predictor of $y$ using the set $X_k$, and $c_k$ is the cost of observing the variables in $K$, then the loss function is $[y-f(X_k)]^2 + c_k$. Lindley then showed that under certain common assumptions, and using the uniform prior, minimizing this loss function is equivalent to choosing the set $K$ which minimizes $N^{-1}R(\beta_j/\beta_k) + c_k$. Again, giving a strong Bayesian justification for using the statistic $R(\beta_j/\beta_k)$.

In conclusion, we have shown that in the case of the uniform prior, the distance between the full predictive density and the predictive density derived from a set $K$, $d_K$, is a function of $R(\beta_j/\beta_k)$. And, we conjecture that $d_K$ is a monotonic function of $R(\beta_j/\beta_k)$, but this is yet to be proven. Nevertheless, these distances are useful in deciding upon which variables to include when predicting a future $y$ value. For, suppose we have decided upon a fixed number of parameters to include, say $s$, then we would calculate the distances for each subset of this size, and then use that set which minimized the distance. The question of how to make comparisons between models of differing sizes is not as simple to answer. One suggestion we have made is to consider the cost of each variable, and use the set $K$ which minimizes $d_K + c_K$. Note, this is similar to Lindley's (1968) idea of minimizing $[y-f(X_k)]^2 + c_k$. But, this is only one suggestion and this question certainly deserves further consideration.
6. REFERENCES


7. ACKNOWLEDGMENTS

I wish to thank Dr. Glen Meeden for all the guidance he has given me while working on this dissertation. His patient prodding and gentle encouragement helped in many subtle ways. His willingness to share his knowledge, his insights, and his sense of humor was a gift I will always deeply appreciate.

I especially want to thank my Mother, without whose help I could never have done any of this. She has given me years of endless support, love, and caring, and this paper is as much a credit to her as it is to me. I will also be eternally grateful for her generous belief in long distance phone calls.

I also want to thank my Father, whose memory has sustained me through many moments of self-doubt. His faith in me provided me with the confidence I needed to do this work. He always tried to teach me to have respect for my own abilities and those of others, and I only hope that I have lived up to his expectations.

On the Jersey front, I want to thank my best friends Kathy Thurston and Holly Gera, and my sister and brother-in-law, Sandy and Robert Otto. All of them have always been so supportive and encouraging, even though they never really understood why I would want to go to Iowa.

In Iowa, I have three very special friends to thank, Mike Rogers, Jan van Schaik, and Robert Stephenson. Mike is the most considerate friend I have ever had, and his thoughtfulness and unique sense of humor have brightened many long days. I also want to thank him for whisking
me away in his wonderful car every time I started to go crazy. I want
to thank Jan for teaching me about strength and independence, and for
helping me to learn to appreciate myself. And I especially want to
thank him for going to lunch with me every Friday during these last
trying months. I want to thank Robert for being a gentle and caring
friend, and for always being an exceptionally patient listener. I
also want to thank him for continually trying, usually successfully,
to make me take a break.

Lastly, but most importantly, I thank God for all the strength
and love He has given me, and for blessing my life with all of these,
and so many other, special people.