12-2012

OBJECTIVE BAYESIAN INFERENCE ON THE COMMON MEAN OF NORMAL DISTRIBUTIONS

Shiyi Tu
Clemson University, stu@clemson.edu

OBJECTIVE BAYESIAN INFERENCE ON THE COMMON MEAN OF NORMAL DISTRIBUTIONS

A Project
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science
Mathematical Science

---

by
Shiyi Tu
December 2012

Accepted by:
Dr. Xiaoqian Sun, Committee Chair
Dr. Colin Gallagher
Dr. Chanseok Park

Abstract

One of the oldest problems in statistical area is to make inference on a common mean of several different normal populations with unknown and probably unequal variances. There are several different ways to make inference on the common mean. The most common methods are point estimation, hypothesis testing, and interval estimation. Point estimation uses sample data to calculate a single value serving as a best guess for the unknown population mean. Hypothesis testing assumes all populations have the same mean as the null hypothesis. Interval estimation is an interval of possible values of the unknown mean.

In this paper, we focus on point estimation and hypothesis testing and use Bayesian methods to make inference on the common mean of two different normal populations. Since the specification of a subjective prior is often difficult and polemical in scientific communication, information theory may be used to develop a prior, the reference prior, which only depends on the assumed model. We will introduce an invariant loss function – intrinsic loss function to deal with the problem. The combination of the intrinsic discrepancy and appropriately defined reference prior provides an integrated objective Bayesian solution to both estimation and hypothesis testing problems.

Table of Contents

Chapter 1

Introduction

We consider point estimation and hypothesis testing on a common mean of two normal populations with unknown and probably unequal variances based on the Bayesian method. Since both of these inferences could be described as the solution to a specific decision problem which heavily depends on the particular choice of loss functions, the selection of loss function is very important. We will introduce an invariant loss function–intrinsic loss to deal with our problems later. First let's recall some basic ideas about Bayesian inference.

1.1 Posterior Distribution

Given a probability model $M = \{f(\mathbf{x} \mid \boldsymbol{\theta}), \mathbf{x} \in \mathbf{X}, \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, we are interested in making inference about unknown parameter $\boldsymbol{\theta}$ from data $\mathbf{X}$. For example, based on the data, we want to explore the values of $\boldsymbol{\theta}$. Furthermore, exploring $\boldsymbol{\theta}$'s plausible numbers as estimators of different components. Then, the extent of uncertainty associated with such estimators, so the Bayesian needs a distribution for $\boldsymbol{\theta}$ before $\mathbf{X}$ is observed. Such distribution is called a prior distribution or simply a prior because it quantifies uncertainty about $\boldsymbol{\theta}$ prior to seeing data. Then, the Bayesian calculates the conditional probability density of $\boldsymbol{\theta}$ given $\mathbf{X} = \mathbf{x}$ by Bayes formula:

$$\pi(\boldsymbol{\theta} \mid \mathbf{x}) = \frac{\pi(\boldsymbol{\theta}) f(\mathbf{x} \mid \boldsymbol{\theta})}{\int_{\Theta} \pi(\boldsymbol{\omega}) f(\mathbf{x} \mid \boldsymbol{\omega}) d\boldsymbol{\omega}},$$

where $\pi(\boldsymbol{\theta})$ is the prior density function and $f(\mathbf{x} \mid \boldsymbol{\theta})$ is the density of $\mathbf{X}$, interpreted as the conditional

density of $\mathbf{X}$ given $\boldsymbol{\theta}$.

The conditional density $\pi(\boldsymbol{\theta} \mid \mathbf{x})$ of $\boldsymbol{\theta}$ given $\mathbf{X} = \mathbf{x}$ is called the posterior density, a quantification of our uncertainty about $\boldsymbol{\theta}$ in the light of data.

1.2 Point Estimation

We mentioned model M in section 1.1. Now we introduce $l(\boldsymbol{\theta}_0, \boldsymbol{\theta})$ to be the loss function when working with model M, where $\boldsymbol{\theta}$ is an unknown parameter which has generated the available data and $\boldsymbol{\theta}_0$ is a value that is used as a proximation for $\boldsymbol{\theta}$. As mentioned before, point estimation and hypothesis testing may both be appropriately described as specific decision problems using a common prior distribution and a common loss function. The results, which are obviously conditional on the assumed model M, may dramatically depend on the particular choices made for both the prior and the loss function. However, given the available data $\mathbf{x}$, both of the prior and the loss function only depend on those through the corresponding posterior expectation of loss function:

$$L(\boldsymbol{\theta}_0 \mid \mathbf{x}) = \int_{\Theta} l(\boldsymbol{\theta}_0, \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid \mathbf{x}) d\boldsymbol{\theta}. \tag{1.1}$$

To choose an estimator for $\boldsymbol{\theta}$ may be seen as a decision problem where the action space is the class $\boldsymbol{\Theta}$ of all possible $\boldsymbol{\theta}$ values. Foundations of decision theory dictate that the best estimator is the one that minimizes $L(\boldsymbol{\theta}_0 \mid \mathbf{x})$ in (1.1).

Definition 1. *The Bayes estimator* $\boldsymbol{\theta}^*(\mathbf{x}) = arg \inf_{\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}} L(\boldsymbol{\theta}_0 \mid \mathbf{x})$ *is one which minimizes the posterior expected loss.*

We should be aware that the Bayes estimator is usually not invariant under one to one transformation.

1.3 Hypothesis Testing

Consider a value $\boldsymbol{\theta}_0$ of the vector of interest, assuming $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ would noticeably simplify the model. Testing the hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ vs $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ may be described as a decision problem where the action space $A = \{a_0, a_1\}$ contains only two elements: to accept (a_0) or to reject (a_1) the null hypothesis under scrutiny. Foundations require to specify a loss function $l(a_i, \boldsymbol{\theta})$ measuring the

consequences of accepting or rejecting H_0 as a function of the actual parameter values. Given data $\mathbf{x}$, the optimal action will be to reject the null hypothesis if and only if the expected posterior loss of acception is larger than that of rejection, that is

$$\int_{\Theta} [l(a_0, \boldsymbol{\theta}) - l(a_1, \boldsymbol{\theta})]\pi(\boldsymbol{\theta} \mid \mathbf{x}) d\boldsymbol{\theta} > 0.$$

Hence, only the loss difference $\Delta l(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = l(a_0, \boldsymbol{\theta}) - l(a_1, \boldsymbol{\theta})$, which measures the advantage of rejecting H_0 as a function of the parameter values, must be specified. Without loss of generality, the function Δl could be written in the form

$$\Delta l(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = l(\boldsymbol{\theta}_0, \boldsymbol{\theta}) - l_0,$$

where $l(\boldsymbol{\theta}_0, \boldsymbol{\theta})$ describes the loss function, which is consistent with the notation in section 1.2, and l_0 is a context dependent positive constant. So we have the following definition

Definition 2. *The Bayes test criterion to decide on the compatibility of $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ with available data $\mathbf{x}$ to reject $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ if and only if $L(\boldsymbol{\theta}_0 \mid \mathbf{x}) > l_0$, where $L(\boldsymbol{\theta}_0 \mid \mathbf{x})$ is the posterior expectation of loss function in (1.1), l_0 is a context dependent positive constant.*

Chapter 2

Research Design and Methods

As we have talked in the previous chapter, our goal is to make inference on the common mean of two different normal populations. Let $N(\mu, \sigma_1^2)$ and $N(\mu, \sigma_2^2)$ be such two normal populations, and assume $x_{i1}, x_{i2}, \cdots, x_{in_i}$ are independent and follow $N(\mu, \sigma_i^2)$, $i = 1, 2$. Our hypothesis test is $H_0 : \mu = \mu_0$ *vs* $H_1 : \mu \neq \mu_0$. We use the intrinsic loss and the reference prior in our calculation.

2.1 Intrinsic Loss Function

The basic idea is to define the discrepancy between two probability densities $p_1(\mathbf{x})$ and $p_2(\mathbf{x})$, denoted by $\min\{k(p_1 \mid p_2), k(p_2 \mid p_1)\}$, where

$$k(p_2 \mid p_1) = \int_{\mathbf{X}} p_1(\mathbf{x}) \ \log \frac{p_1(\mathbf{x})}{p_2(\mathbf{x})} \ d\mathbf{x},$$

is the directed logarithmic divergence of $p_2(\mathbf{x})$ from $p_1(\mathbf{x})$ in [4] and [3].
Then we quote the definition of intrinsic discrepancy from [1].

Definition 3. *The intrinsic discrepancy $\delta(p_1, p_2)$ between two probability densities $p_1(\mathbf{x})$ and $p_2(\mathbf{x})$ for the random quantity $\mathbf{x} \in \mathbf{X}$ is*

$$\delta(p_1(\mathbf{x}), p_2(\mathbf{x})) = \min \left\{ \int_{\mathbf{X}} p_1(\mathbf{x}) \ \log \frac{p_1(\mathbf{x})}{p_2(\mathbf{x})} \ d\mathbf{x}, \ \int_{\mathbf{X}} p_2(\mathbf{x}) \ \log \frac{p_2(\mathbf{x})}{p_1(\mathbf{x})} \ d\mathbf{x} \right\}.$$

The intrinsic discrepancy between two families of probability densities for the random quantity $\mathbf{x} \in \mathbf{X}$,

$M_1 = \{p_1(\mathbf{x} \mid \boldsymbol{\phi}), \boldsymbol{\phi} \in \boldsymbol{\Phi}\}$ *and* $M_2 = \{p_2(\mathbf{x} \mid \boldsymbol{\psi}), \boldsymbol{\psi} \in \boldsymbol{\Psi}\}$*, is given by*

$$\delta(M_1, M_2) = \min_{\boldsymbol{\phi} \in \boldsymbol{\Phi}, \boldsymbol{\psi} \in \boldsymbol{\Psi}} \delta(p_1(\mathbf{x} \mid \boldsymbol{\phi}), p_2(\mathbf{x} \mid \boldsymbol{\psi})).$$

With this definition, we can introduce the intrinsic discrepancy loss.

Definition 4. *The intrinsic discrepancy loss* $\delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\omega})$ *from replacing the probability model* $M = \{p(\mathbf{x} \mid \boldsymbol{\theta}, \boldsymbol{\omega}), \boldsymbol{\theta} \in \boldsymbol{\Theta}, \boldsymbol{\omega} \in \boldsymbol{\Omega}, \mathbf{x} \in \mathbf{X}\}$ *by its restriction with* $\boldsymbol{\theta} = \boldsymbol{\theta}_0$*,* $M_0 = \{p(\mathbf{x} \mid \boldsymbol{\theta}_0, \boldsymbol{\omega}), \boldsymbol{\omega} \in \boldsymbol{\Omega}, \mathbf{x} \in \mathbf{X}\}$ *is the intrinsic discrepancy between the probability density* $p(\mathbf{x} \mid \boldsymbol{\theta}, \boldsymbol{\omega})$ *and the family of probability densities* $\{p(\mathbf{x} \mid \boldsymbol{\theta}_0, \boldsymbol{\omega}), \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$*, that is,*

$$\delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\omega}) = \min_{\boldsymbol{\omega}_0 \in \boldsymbol{\Omega}} \delta\{p(\mathbf{x} \mid \boldsymbol{\theta}, \boldsymbol{\omega}), p(\mathbf{x} \mid \boldsymbol{\theta}_0, \boldsymbol{\omega}_0)\}.$$

It is easy to see that if the support of $p(\mathbf{x} \mid \boldsymbol{\theta}, \boldsymbol{\omega})$ is convex for all $(\boldsymbol{\theta}, \boldsymbol{\omega})$, we could write the intrinsic loss as follows:

$$\delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\omega}) = \min \left\{ \inf_{\boldsymbol{\omega}_0 \in \boldsymbol{\Omega}} k(\boldsymbol{\theta}_0, \boldsymbol{\omega}_0 \mid \boldsymbol{\theta}, \boldsymbol{\omega}), \inf_{\boldsymbol{\omega}_0 \in \boldsymbol{\Omega}} k(\boldsymbol{\theta}, \boldsymbol{\omega} \mid \boldsymbol{\theta}_0, \boldsymbol{\omega}_0) \right\}, \tag{2.1}$$

where $k(\boldsymbol{\theta}_0, \boldsymbol{\omega}_0 \mid \boldsymbol{\theta}, \boldsymbol{\omega})$ is the KL-divergence of $p(\mathbf{x} \mid \boldsymbol{\theta}_0, \boldsymbol{\omega}_0)$ from $p(\mathbf{x} \mid \boldsymbol{\theta}, \boldsymbol{\omega})$, so does $\inf_{\boldsymbol{\omega}_0 \in \boldsymbol{\Omega}} k(\boldsymbol{\theta}, \boldsymbol{\omega} \mid \boldsymbol{\theta}_0, \boldsymbol{\omega}_0)$.

In this case, let $f_i(x \mid \mu, \sigma_i^2)$ denote the probability density function of $N(\mu, \sigma_i^2)$, $i = 1, 2$. Define $\mathbf{x_1} = (x_{11}, x_{12}, \cdots, x_{1n_1})$ and $\mathbf{x_2} = (x_{21}, x_{22}, \cdots, x_{2n_2})$, then the joint density of $(\mathbf{x_1}, \mathbf{x_2})$ is

$$f(\mathbf{x_1}, \mathbf{x_2} \mid \mu, \sigma_1^2, \sigma_2^2) = f_1(\mathbf{x_1} \mid \mu, \sigma_1^2)\ f_2(\mathbf{x_2} \mid \mu, \sigma_2^2),$$

where $f_i(\mathbf{x_i} \mid \mu, \sigma_i^2) = \prod_{j=1}^{n_i} f_i(x_{ij} \mid \mu, \sigma_i^2)$, $i = 1, 2$. From equation (2.1), we could see that the intrinsic loss function is

$$\delta(\mu_0, \mu, \sigma_1^2, \sigma_2^2) = \min \left\{ \inf_{\sigma_3^2 > 0, \sigma_4^2 > 0} k(\mu_0, \sigma_3^2, \sigma_4^2 \mid \mu, \sigma_1^2, \sigma_2^2), \inf_{\sigma_3^2 > 0, \sigma_4^2 > 0} k(\mu, \sigma_1^2, \sigma_2^2 \mid \mu_0, \sigma_3^2, \sigma_4^2) \right\}. \tag{2.2}$$

Before we compute the discrepancy loss function $\delta(\mu_0, \mu, \sigma_1^2, \sigma_2^2)$ in (2.2), we give a lemma which could simplify the calculation.

Lemma 1. *Let $f_1(x \mid \theta_1)$, $f_2(x \mid \theta_2)$ be two probability density functions, assume $x_1 \sim f_1(x \mid \theta_1)$, $x_2 \sim f_2(x \mid \theta_2)$ and x_1, x_2 are independent. Let $f(x_1, x_2 \mid \theta_1, \theta_2)$ be the joint density function of (x_1, x_2), so $f(x_1, x_2 \mid \theta_1, \theta_2) = f_1(x_1 \mid \theta_1)\ f_2(x_2 \mid \theta_2)$. We have*

$$k(\theta_3, \theta_4 \mid \theta_1, \theta_2) = k(\theta_3 \mid \theta_1) + k(\theta_4 \mid \theta_2),$$

where $k(\theta_3, \theta_4 \mid \theta_1, \theta_2)$ is the KL-divergence of $f(x_1, x_2 \mid \theta_3, \theta_4)$ from $f(x_1, x_2 \mid \theta_1, \theta_2)$, so do $k(\theta_3 \mid \theta_1)$ and $k(\theta_4 \mid \theta_2)$.

Proof.

$$\begin{aligned}
k(\theta_3, \theta_4 \mid \theta_1, \theta_2) &= \int_{\mathbf{X_1}} \int_{\mathbf{X_2}} f(x_1, x_2 \mid \theta_1, \theta_2)\ \log \frac{f(x_1, x_2 \mid \theta_1, \theta_2)}{f(x_1, x_2 \mid \theta_3, \theta_4)}\ dx_1 dx_2 \\
&= \int_{\mathbf{X_1}} \int_{\mathbf{X_2}} f_1(x_1 \mid \theta_1)\ f_2(x_2 \mid \theta_2)\ \log \frac{f_1(x_1 \mid \theta_1)\ f_2(x_2 \mid \theta_2)}{f_1(x_1 \mid \theta_3)\ f_2(x_2 \mid \theta_4)}\ dx_1 dx_2 \\
&= \int_{\mathbf{X_1}} \int_{\mathbf{X_2}} f_1(x_1 \mid \theta_1)\ f_2(x_2 \mid \theta_2)\ \Big(\log \frac{f_1(x_1 \mid \theta_1)}{f_1(x_1 \mid \theta_3)} + \log \frac{f_2(x_2 \mid \theta_2)}{f_2(x_2 \mid \theta_4)} \Big)\ dx_1 dx_2 \\
&= \int_{\mathbf{X_1}} f_1(x_1 \mid \theta_1)\ \log \frac{f_1(x_1 \mid \theta_1)}{f_1(x_1 \mid \theta_3)}\ \int_{\mathbf{X_2}} f_2(x_2 \mid \theta_2)\ dx_2 dx_1 \\
&\quad + \int_{\mathbf{X_2}} f_2(x_2 \mid \theta_2)\ \log \frac{f_2(x_2 \mid \theta_2)}{f_2(x_2 \mid \theta_4)}\ \int_{\mathbf{X_1}} f_1(x_1 \mid \theta_1)\ dx_1 dx_2 \\
&= \int_{\mathbf{X_1}} f_1(x_1 \mid \theta_1)\ \log \frac{f_1(x_1 \mid \theta_1)}{f_1(x_1 \mid \theta_3)}\ dx_1 + \int_{\mathbf{X_2}} f_2(x_2 \mid \theta_2)\ \log \frac{f_2(x_2 \mid \theta_2)}{f_2(x_2 \mid \theta_4)}\ dx_2 \\
&= k(\theta_3 \mid \theta_1) + k(\theta_4 \mid \theta_2). \quad \square
\end{aligned}$$

Back to our situation, by the using of Lemma 1, it is easy to see that

$$\begin{aligned} k(f(\mathbf{x_1}, \mathbf{x_2} \mid \mu, \sigma_1^2, \sigma_2^2) \mid f(\mathbf{x_1}, \mathbf{x_2} \mid \mu_0, \sigma_3^2, \sigma_4^2)) &= k(f(\mathbf{x_1} \mid \mu, \sigma_1^2) \mid f(\mathbf{x_1} \mid \mu_0, \sigma_3^2)) \\ &\quad + k(f(\mathbf{x_2} \mid \mu, \sigma_2^2) \mid f(\mathbf{x_2} \mid \mu_0, \sigma_4^2)). \end{aligned} \tag{2.3}$$

So what we need is just $k(f(\mathbf{x_1} \mid \mu, \sigma_1^2) \mid f(\mathbf{x_1} \mid \mu_0, \sigma_3^2))$, we use $k(\mu, \sigma_1^2 \mid \mu_0, \sigma_3^2)$ to represent it for simplicity. Then we have

$$\begin{aligned} k(\mu, \sigma_1^2 \mid \mu_0, \sigma_3^2) &= \int_{\mathbf{X}} f_1(\mathbf{x_1} \mid \mu_0, \sigma_3^2) \ \log \frac{f_1(\mathbf{x_1} \mid \mu_0, \sigma_3^2)}{f_1(\mathbf{x_1} \mid \mu, \sigma_1^2)} \ d\mathbf{x_1} \\ &= n_1 \int_{\mathbf{X}} f_1(x \mid \mu_0, \sigma_3^2) \ \log \frac{f_1(x \mid \mu_0, \sigma_3^2)}{f_1(x \mid \mu, \sigma_1^2)} \ dx \\ &= n_1 \int_{\mathbf{X}} f_1(x \mid \mu_0, \sigma_3^2) \ \log \frac{\frac{1}{\sqrt{2\pi\sigma_3^2}} \exp\{-\frac{(x-\mu_0)^2}{2\sigma_3^2}\}}{\frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\{-\frac{(x-\mu)^2}{2\sigma_1^2}\}} \ dx \\ &= n_1 \int_{\mathbf{X}} f_1(x \mid \mu_0, \sigma_3^2) \left(-\frac{1}{2} \log \frac{\sigma_3^2}{\sigma_1^2} + \frac{(x-\mu)^2}{2\sigma_1^2} - \frac{(x-\mu_0)^2}{2\sigma_3^2} \right) \ dx \\ &= n_1 \left(-\frac{1}{2} \log \frac{\sigma_3^2}{\sigma_1^2} + \frac{\mu^2}{2\sigma_1^2} - \frac{\mu_0^2}{2\sigma_3^2} + (\frac{1}{2\sigma_1^2} - \frac{1}{2\sigma_3^2}) E(x^2) + (\frac{\mu_0}{\sigma_3^2} - \frac{\mu}{\sigma_1^2}) E(x) \right) \\ &= n_1 \left(-\frac{1}{2} \log \frac{\sigma_3^2}{\sigma_1^2} + \frac{\mu^2}{2\sigma_1^2} - \frac{\mu_0^2}{2\sigma_3^2} + (\frac{1}{2\sigma_1^2} - \frac{1}{2\sigma_3^2})(\mu_0^2 + \sigma_3^2) + (\frac{\mu_0}{\sigma_3^2} - \frac{\mu}{\sigma_1^2}) \mu_0 \right) \\ &= \frac{n_1}{2} \left(\frac{(\mu - \mu_0)^2}{\sigma_1^2} + \frac{\sigma_3^2}{\sigma_1^2} - \log \frac{\sigma_3^2}{\sigma_1^2} - 1 \right). \end{aligned} \tag{2.4}$$

Above equation is minimized when σ_3^2 equals σ_1^2, that is

$$\inf_{\sigma_3^2 > 0} k(\mu, \sigma_1^2 \mid \mu_0, \sigma_3^2) = \frac{n_1}{2} \frac{(\mu - \mu_0)^2}{\sigma_1^2}. \tag{2.5}$$

$k(\mu, \sigma_2^2 \mid \mu_0, \sigma_4^2)$ has a similar form to (2.4), so it is easy to get

$$\inf_{\sigma_4^2 > 0} k(\mu, \sigma_2^2 \mid \mu_0, \sigma_4^2) = \frac{n_2}{2} \frac{(\mu - \mu_0)^2}{\sigma_2^2}. \tag{2.6}$$

Combining (2.5) and (2.6) with (2.3), we could get

$$\inf_{\sigma_3^2>0,\sigma_4^2>0} k(\mu,\sigma_1^2,\sigma_2^2 \mid \mu_0,\sigma_3^2,\sigma_4^2) = \frac{n_1}{2}\frac{(\mu-\mu_0)^2}{\sigma_1^2} + \frac{n_2}{2}\frac{(\mu-\mu_0)^2}{\sigma_2^2}. \tag{2.7}$$

It is similarly for the case $k(\mu_0,\sigma_3^2 \mid \mu,\sigma_1^2)$, we could find

$$k(\mu_0,\sigma_3^2 \mid \mu,\sigma_1^2) = \frac{n_1}{2}\left(\frac{(\mu-\mu_0)^2}{\sigma_3^2} + \frac{\sigma_1^2}{\sigma_3^2} - \log\frac{\sigma_1^2}{\sigma_3^2} - 1\right).$$

It is minimized by choosing σ_3^2 as $(\mu-\mu_0)^2+\sigma_1^2$, so

$$\inf_{\sigma_3^2>0} k(\mu_0,\sigma_3^2 \mid \mu,\sigma_1^2) = \frac{n_1}{2}\log\left(1+\frac{(\mu-\mu_0)^2}{\sigma_1^2}\right). \tag{2.8}$$

Similarly,

$$\inf_{\sigma_4^2>0} k(\mu_0,\sigma_4^2 \mid \mu,\sigma_1^2) = \frac{n_2}{2}\log\left(1+\frac{(\mu-\mu_0)^2}{\sigma_2^2}\right). \tag{2.9}$$

Also combining (2.8) and (2.9) with (2.3), it comes to

$$\inf_{\sigma_3^2>0,\sigma_4^2>0} k(\mu_0,\sigma_3^2,\sigma_4^2 \mid \mu,\sigma_1^2,\sigma_2^2) = \frac{n_1}{2}\log\left(1+\frac{(\mu-\mu_0)^2}{\sigma_1^2}\right) + \frac{n_2}{2}\log\left(1+\frac{(\mu-\mu_0)^2}{\sigma_2^2}\right). \tag{2.10}$$

Since we have $x > \log(1+x),\ \forall x > 0$, so from (2.2), (2.7) and (2.10), we could obtain the intrinsic loss function

$$\delta(\mu_0,\mu,\sigma_1^2,\sigma_2^2) = \frac{n_1}{2}\log\left(1+\frac{(\mu-\mu_0)^2}{\sigma_1^2}\right) + \frac{n_2}{2}\log\left(1+\frac{(\mu-\mu_0)^2}{\sigma_2^2}\right). \tag{2.11}$$

2.2 Reference Prior and Posterior Distribution

The main of objective in this paper is the parameter, μ. Considering the prior in the form of $\pi(\mu,\sigma_1^2) = p(\sigma_2^2 \mid \sigma_1^2,\mu)\ p(\sigma_1^2 \mid \mu)\ p(\mu)$, which leads to the reference prior. Simple calculation

shows that the Fisher Information matrix of $(\mu, \sigma_1^2, \sigma_2^2)$ is

$$I(\mu, \sigma_1^2, \sigma_2^2) = \begin{pmatrix} \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} & 0 & 0 \\ 0 & \frac{1}{2\sigma_1^4} & 0 \\ 0 & 0 & \frac{1}{2\sigma_2^4} \end{pmatrix}.$$

Since $I(\mu, \sigma_1^2, \sigma_2^2)$ has a diagonal form, it is easy to see the reference prior

$$\pi(\mu, \sigma_1^2, \sigma_2^2) \propto \frac{1}{\sigma_1^2} \frac{1}{\sigma_2^2}.$$

Then the posterior distribution

$$\begin{aligned} \pi(\mu, \sigma_1^2, \sigma_2^2 \mid \mathbf{x}) &\propto \left(\frac{1}{\sqrt{2\pi\sigma_1^2}}\right)^{n_1} \exp\left\{-\frac{\sum_{i=1}^{n_1}(x_{1i}-\mu)^2}{2\sigma_1^2}\right\} \left(\frac{1}{\sqrt{2\pi\sigma_2^2}}\right)^{n_2} \exp\left\{-\frac{\sum_{j=1}^{n_2}(x_{2j}-\mu)^2}{2\sigma_2^2}\right\} \frac{1}{\sigma_1^2}\frac{1}{\sigma_2^2} \\ &\propto \left(\frac{1}{\sigma_1^2}\right)^{\frac{n_1}{2}+1} \left(\frac{1}{\sigma_2^2}\right)^{\frac{n_2}{2}+1} \exp\left\{-\frac{\sum_{i=1}^{n_1}(x_{1i}-\mu)^2}{2\sigma_1^2} - \frac{\sum_{j=1}^{n_2}(x_{2j}-\mu)^2}{2\sigma_2^2}\right\}. \end{aligned} \tag{2.12}$$

This posterior distribution will be used for our Bayesian inference.

2.3 Posterior Expected Loss

Denote $L(\mu_0 \mid \mathbf{x})$ to be the posterior expectation of loss function in this case. From (1.1), we have

$$L(\mu_0 \mid \mathbf{x}) = \iiint \delta(\mu_0, \mu, \sigma_1^2, \sigma_2^2)\ \pi(\mu, \sigma_1^2, \sigma_2^2 \mid \mathbf{x})\ d\mu\ d\sigma_1^2\ d\sigma_2^2. \tag{2.13}$$

Based on (2.13), we can calculate $L(\mu_0 \mid \mathbf{x})$ for both hypothesis testing and point estimation. By observing the form of the loss function (2.11) and the posterior distribution (2.12), we find it may not be an easy task to compute $L(\mu_0 \mid \mathbf{x})$ by using numerical integration. Consequently, we use MCMC algorithm to generate a sequence of $(\mu^{(1)}, \sigma_1^{2(1)}, \sigma_2^{2(1)}, \cdots, \mu^{(n)}, \sigma_1^{2(n)}, \sigma_2^{2(n)})$. Then we can apply the theorem, Law of Large Numbers for Markov chains [2], to approach $L(\mu_0 \mid \mathbf{x})$ by

$$\iiint \delta(\mu_0, \mu, \sigma_1^2, \sigma_2^2)\ \pi(\mu, \sigma_1^2, \sigma_2^2 \mid \mathbf{x})\ d\mu\ d\sigma_1^2\ d\sigma_2^2 \approx \frac{1}{n}\sum_{i=1}^{n} \delta(\mu_0, \mu_i, \sigma_{1i}^2, \sigma_{2i}^2),$$

where n is the number of iterations.

From equation (2.12), we know the posterior distribution is

$$\pi(\mu, \sigma_1^2, \sigma_2^2 \mid \mathbf{x}) \propto \left(\frac{1}{\sigma_1^2}\right)^{\frac{n_1}{2}+1} \left(\frac{1}{\sigma_2^2}\right)^{\frac{n_2}{2}+1} \exp\left\{-\frac{\sum_{i=1}^{n_1}(x_{1i}-\mu)^2}{2\sigma_1^2} - \frac{\sum_{j=1}^{n_2}(x_{2j}-\mu)^2}{2\sigma_2^2}\right\}.$$

It is hard to reach a simple form like $\pi(\mu, \sigma_1^2, \sigma_2^2 \mid \mathbf{x}) = \pi(\sigma_1^2 \mid \sigma_2^2, \mu, \mathbf{x})\ \pi(\sigma_2^2 \mid \mu, \mathbf{x})\ \pi(\mu \mid \mathbf{x})$, so we use the Gibbs Sampler Method instead. From (2.12), we could see the full conditional distributions are:

$$\mu \mid \mathbf{x}, \sigma_1^2, \sigma_2^2 \quad \sim \quad \mathrm{N}\left(\frac{n_1\bar{x_1}/\sigma_1^2 + n_2\bar{x_2}/\sigma_2^2}{n_1/\sigma_1^2 + n_2/\sigma_2^2},\ \frac{1}{n_1/\sigma_1^2 + n_2/\sigma_2^2}\right), \tag{2.14}$$

$$\sigma_1^2 \mid \mathbf{x}, \mu \quad \sim \quad \mathrm{Inv-Gamma}\left(\frac{n_1}{2},\ \frac{\sum(x_{1i}-\mu)^2}{2}\right), \tag{2.15}$$

$$\sigma_2^2 \mid \mathbf{x}, \mu \quad \sim \quad \mathrm{Inv-Gamma}\left(\frac{n_2}{2},\ \frac{\sum(x_{2j}-\mu)^2}{2}\right), \tag{2.16}$$

where $\bar{x_j}$ is the sample mean of population $N(\mu, \sigma_j^2)$, $j = 1, 2$.

Let $p(\mu \mid \mathbf{x}, \sigma_1^2, \sigma_2^2)$, $p(\sigma_1^2 \mid \mathbf{x}, \mu)$ and $p(\sigma_2^2 \mid \mathbf{x}, \mu)$ denotes the full conditional distributions from (2.14) to (2.16) respectively. With these formulations, we can propose the following algorithm:

Step 1: Choose an initial value $(\mu^{(0)}, \sigma_1^{2(0)}, \sigma_2^{2(0)})$.

Step 2: Given the i^{th} sample $(\mu^{(i)}, \sigma_1^{2(i)}, \sigma_2^{2(i)})$, generating $\mu^{(i+1)}$ from $p(\mu^{(i+1)} \mid \mathbf{x}, \sigma_1^{2(i)}, \sigma_2^{2(i)})$.

Step 3: Generate $\sigma_1^{2(i+1)}$ from $p(\sigma_1^{2(i+1)} \mid \mathbf{x}, \mu^{(i)})$.

Step 4: Generate $\sigma_2^{2(i+1)}$ from $p(\sigma_2^{2(i+1)} \mid \mathbf{x}, \mu^{(i)})$.

Step 5: Return to Step 2.

The sequence $\{(\mu^{(i)}, \sigma_1^{2(i)}, \sigma_2^{2(i)}), i = 0, 1, 2, \cdots\}$ constitutes a Markov chain, and the stationary distribution of this Markov chain is just the joint posterior distribution of $\pi(\mu, \sigma_1^2, \sigma_2^2 \mid \mathbf{x})$ in (2.12). However, it may take a while for that stationary distribution to be reached, so samples from the beginning of the chain (the burn-in period) may not accurately represent the desired distribution,

and must be thrown away. We use the remaining samples to approach our posterior expectation of loss function $L(\mu_0 \mid \mathbf{x})$ in (2.13) by applying the theorem of the Law of Large Numbers for Markov chains. With the calculation of $L(\mu_0 \mid \mathbf{x})$, we could continue our inference on hypothesis testing and point estimation.

Chapter 3

Simulation Study

In this chapter, we illustrate the performance of our proposed Bayesian method by hypothesis testing and point estimation. We set μ at several different values. For each μ, we repeat the process of hypothesis testing 10000 times to find the probability of rejections. Then we compare the Bayesian estimators obtained during the process to their theoretical values.

3.1 Data generating process

In the simulation, we assign the parameters to various values. The parameter μ_0 is the hypothesized population mean which is set at 0 in the studies for simplicity. And the parameter μ is the actual mean of distributions where we sample from. When $\mu_0 = 0$, we are testing $H_0 : \ \mu = 0$ *vs* $H_1 : \ \mu \neq 0$.

Now, we set μ at 0, 0.1, 0.15, 0.2, and without loss of generality picking $\sigma_1^2 = 0.04$ and $\sigma_2^2 = 0.25$, where σ_1^2 and σ_2^2 are the population variances. Also, we define n_j to be the sample size for population j, $j = 1, 2$. In this case, we fix $n_1 = 30$ and $n_2 = 50$.

Aa we mentioned in section 2.3, when we calculate the posterior expectation of loss function $L(\mu_0 \mid \mathbf{x})$, we need to drop the beginning samples of sequence $\{(\mu^{(i)}, \sigma_1^{2(i)}, \sigma_2^{2(i)}), i = 0, 1, 2, \cdots\}$ which is obtained by applying Gibbs Sampler method. Figure (3.1) shows the trace plots of three parameters.

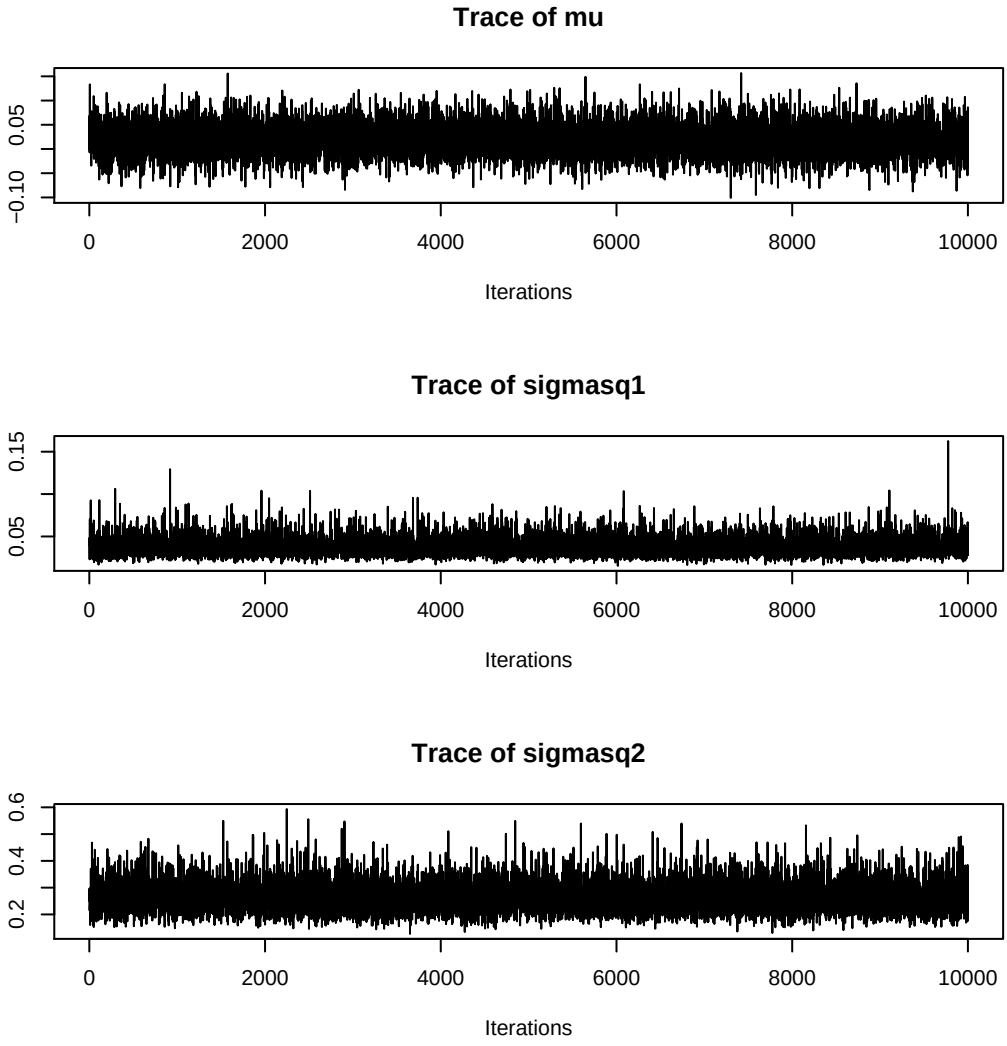

Figure 3.1: trace plots of parameters

Figure (3.1) shows that the chain mixes well and converges very rapidly. Therefore, it is reasonable for us to regard the first 1000 samples of sequence $\{(\mu^{(i)}, \sigma_1^{2(i)}, \sigma_2^{2(i)}), i = 0, 1, 2, \cdots\}$ as burn-in period.

3.2 Results

From previous discussion we anticipate the results reject $H_0 : \ \mu = 0$ if $L(\mu_0 \mid \mathbf{x}) > l$ where l is a constant. For the selection of l, we have the criterion quoted from [1], the values $l = 1.0$ (no evidence against the null), $l = 2.5$ (mild evidence against the null) and $l > 5$ (significant evidence against the null). In this case, we adopt $l = 2.5$ in our simulation.
For the hypothesis test, we have a summation, which is shown in the table below,

μ	number of rejections	percentage of rejections(%)
0	447	4.47
0.1	8404	84.04
0.15	9925	99.25
0.2	10000	100

Table 3.1: summation of hypothesis test

By observing the data from above table, when $\mu = 0$, the probability we reject H_0 is 0.0447, which is quite small. Now, given $\mu = 0.1,\ 0.15,\ 0.2$, we reject H_0 with very high probability.

From Definition 1, we know the Bayesain estimator $\hat{\mu}$ is the one which minimizes the posterior expected loss $L(\mu_0 \mid \mathbf{x})$, that is,

$$\hat{\mu} = arg \inf_{-\infty < \mu_{est} < \infty} L(\mu_0 \mid \mathbf{x}).$$

For each μ, we obtained 10000 Bayesian estimators (each set of generated data could result one estimator), we calculate their mean and variance for each case which is shown in Table (3.2)

μ	Bayesian estimator(sd)
0	0.000251(0.032861)
0.1	0.100666(0.033066)
0.15	0.149914(0.033191)
0.2	0.200427(0.032764)

Table 3.2: summation of Bayesian estimators

Based on table (3.2), we could see the Bayesian estimators are very close to μ and the variances of these estimators are quite small.

Consider the formula of loss function (2.11), since we always have $\log\left(1+\frac{(\mu-\mu_0)^2}{\sigma_i^2}\right) \geq 0$ (with equality $\mu_0 = \mu$), so the theoretical value for Bayesian estimator is μ. That means the values obtained from simulations are very close to theoretical value. Therefore, both hypothesis test and point estimation perform well under the intrinsic loss.

Bibliography

[1] José M. Bernardo and Raúl Rueda. Bayesian hypothesis testing: a reference approach. *International Statistical Review,*, 70(3):351–372, 2002.

[2] J.K. Ghosh, Mohan. Delampady, and Tapas. Samanta. *An introduction to Bayesian analysis*. Springer New York, 2006.

[3] S. Kullback. *Information theory and statistics*. Dover Pubns, 1997.

[4] S. Kullback and R.A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.