Robust Bayesian Analysis of the Linear Model with an Application to a Problem in Psychology

by Yakun Wang

B.S. in Computing and Information Science, May 2014,Taiyuan University of Technology

A Thesis submitted to

The Faculty of
The Columbian College of Arts and Sciences
of The George Washington University
in partial fulfillment of the requirements
for the degree of Master of Science

August 31, 2016

Thesis directed by

Sudip Bose
Associate Professor of Statistics

Acknowledgments

I would like to thank Dr. Sudip Bose and Dr. Subrata Kundu for very valuable comments and suggestions. I am very thankful for the data provided by Dr. George Howe.

Abstract of Thesis

Robust Bayesian Analysis of the Linear Model with an Application to a Problem in Psychology

This thesis discusses the Bayesian analysis method for multi-variable normal linear regression. We use the normal distribution with known variance as the prior for the parameters in the linear model. This is the base prior for the robust Bayesian analysis. The ε-contaminated class and the density ratio class of normal and Cauchy distribution are considered for a robust Bayesian analysis. We also discuss Bayesian methods for hypothesis testing.

Keywords: multivariate linear regression, conjugate priors, ε-contaminated class, density ratio class

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

Bayesian statistical analysis has been widely developed in many fields such as psychology, econometrics, and survival analysis. Berger (1985) discussed the loss function, prior information, and various ways to study Bayesian robustness. Bayesian analysis of the linear regression model is of great appeal. Lindley and Smith (1972) discussed the general Bayesian linear model $E(y) = A\theta$ with a hierarchical normal prior distribution of unknown parameters θ based on exchangeability. Broemeling (1985) introduced a wide variety of linear models including the linear regression model, models for designed of experiments with fixed, mixed or random models, and time series models. Broemeling (1985) considered the conjugate prior density of normal-gamma and improper prior information for the general linear model. In this thesis, we consider the multivariate normal prior distribution of unknown parameters with fixed variance, and we also discuss Bayesian robustness.

Robust Bayesian analysis was first introduced by Good (1950). A robust Bayesian analysis attempts to quantify robustness with respect to the prior. One considers a class of priors to which the prior belongs and calculates the range of posterior quantities of interest. Berger (1990) reviewed Good's thoughts about Bayesian robustness and discussed the main classes used in Bayesian robustness. Berger and Berliner (1986) discussed the ε-contaminated priors with Type II maximum likelihood technique. Sivaganesan and Berger (1989) investigated the robustness of posterior mean, variance and hypothesis testing with ε-contaminated priors. Bose (1994a) considered a modification of the ε-contaminated class with more than one class of contaminating priors.

The density ratio class, mentioned above, is an important class of priors in Bayesian robustness and was first introduced by DeRobertis and Hartigan (1981). Geweke and Petrella (1998) applied the density ratio class to the general linear model and provided a method to calculate the bounds of posterior mean based on the simulated

posterior output. Berger (1990) pointed out that the advantage of the density ratio class is the relatively easy calculation required for a posterior robustness analysis. Bose (1994b) considered a density ratio class with shape and smoothness constraints for multidimensional theta. In this thesis, the ε-contaminated class and the density ratio class are considered for a Bayesian robust analysis of the linear regression model.

Chapter 2: Parameter Estimation

In this thesis, the general linear model $Y = X\theta + \varepsilon$ is considered, where Y is an $n \times 1$ vector of observations, X an $n \times p$ design matrix of observations, θ a $p \times 1$ vector of unknown parameters and ε an $n \times 1$ vector of random errors. Suppose $\varepsilon \sim N(0, \sigma^2 I_n)$ and X is nonsingular.

Bayes Theorem

Bayesian inference combines the prior distribution of the parameter(s) and the density of the data to obtain the posterior probability density function. The likelihood function for the linear model is:

$$f(Y|X, \theta, \sigma^2) = \big(\frac{1}{\sqrt{2\pi\sigma^2}}\big)^n e^{-\frac{1}{2\sigma^2}(Y-X\theta)'(Y-X\theta)}, \tag{1}$$

which gives us the information contained in the sample about unknown parameters and we denote the prior probability density function of θ and σ^2 as $\xi(\theta, \sigma^2)$.

Fixed X v.s. random X

The probability density function of X is $\phi(X|\eta)$, where $\eta \sim \varphi(\eta)$. Thus, according to Bayes theorem, the conditional density of (θ, σ) given Y is:

$$\pi(\theta, \sigma^2|Y, X, \eta) = \frac{f(Y|X, \theta, \sigma^2)\xi(\theta, \sigma^2)\phi(X|\eta)\varphi(\eta)}{\int_0^\infty \int_{R^P} f(Y|X, \theta, \sigma^2)\xi(\theta, \sigma^2)\phi(X|\eta)\varphi(\eta)\ d\theta d\sigma^2}, \tag{2}$$

which is the posterior distribution of θ. Suppose X and η do not involve θ and σ^2, then the term $\phi(x|\eta)\varphi(\eta)$ in both numerator and denominator is constant and can be canceled. The posterior can be rewritten as:

$$\pi(\theta, \sigma^2|Y, X, \eta) = \frac{f(Y|X, \theta, \sigma^2)\xi(\beta, \sigma^2)}{\int_0^\infty \int_{R^P} f(Y|X, \theta, \sigma^2)\xi(\theta, \sigma^2)\ d\theta d\sigma^2}. \tag{3}$$

This shows that the posterior distribution is the same whether X is fixed or random. Thus we can regard X as a fixed design matrix of θ.

Bayesian Inference

The prior information about θ and σ^2 is $\xi(\theta, \sigma^2) = \xi_1(\theta|\sigma^2)\xi_2(\sigma^2)$. Assuming σ^2 is known, the prior density reduces to $\xi_1(\theta)$. Thus the prior information of θ is a normal distribution with mean μ and precision matrix $\Sigma = \tau I_p$, which is:

$$\pi(\theta) = \frac{1}{(2\pi)^{p/2}| \Sigma |^{1/2}} e^{-(\theta-\mu)'\Sigma^{-1}(\theta-\mu)/2} \tag{4}$$

and then the posterior distribution of θ is:

$$\pi(\theta|Y) = \frac{f(Y|\theta)\pi(\theta)}{m(Y)}, \tag{5}$$

where $m(Y) = \int f(Y|\theta)\pi(\theta)d\theta$ is the marginal distribution of Y. It can be easily calculated that $m(Y) \sim N(X\mu, \sigma^2 I_n + X\tau I_p X^T)$. Lindley and Smith (1972) showed that the posterior distribution of θ given Y is $N(Bb, B)$ where

$$B^{-1} = X^T(\sigma^2 I_n)^{-1}X + (\tau^2 I_P)^{-1} \tag{6}$$

and

$$b = X^T(\sigma^2 I_n)^{-1}Y + (\tau^2 I_P)^{-1}\mu. \tag{7}$$

These results will be applied in the Bayesian analysis and robustness study that follows.

Chapter 3: Robust Bayesian Analysis

In a Bayesian analysis, one combines the likelihood with the prior distribution of the parameter (using Bayes' theorem) to derive one's posterior distribution, which is then used for inference. Inferences of interest include estimation of parameters and tests of hypotheses. In the application to psychological data, that we consider, the parameter of interest is a regression coefficient in a regression equation for depression.

The prior in a Bayesian analysis is supposed to reflect one's initial beliefs about the parameter, prior to observing data, expressed as a probability distribution. However, in practice this is hard to do. First note that the parameter θ is a feature of a probability distribution $f(x|\theta)$ for the data. Unlike the data it is not an observable quantity. This provides an additional challenge in eliciting a person's prior beliefs. Even specifying an accurate probability of a single event involving an observable quantity is not easy. This is succinctly illustrated in a quote from Good (1980) in discussing axiom systems of probability elicitation. "For it would be a joke if you were to say that the probability of rain tomorrow (however sharply defined) is 0.3057876289."

Specifying the probability distribution of a one-dimensional parameter, requires infinitely many probabilities to be specified exactly. Noting that it is enough to specify probabilities of intervals with rational end points, one still has to specify *exactly* a countably infinite number of probabilities. In reality, the best that we can do is to specify one's prior approximately, which leads into Bayesian robustness with a class of priors.

Maximin Posterior Regret in Bayesian Robustness

Minimax approach

The minimax approach is a well-studied approach to choosing a decision rule. The idea is that one considers the worst case performance of one's decision rule, and finds

the rule that has the best worst case performance. In classical decision theory, with parameter $\theta \in \Theta$, the parameter space, the minimax approach is to select the rule δ that minimizes $\sup_{\theta \in \Theta} R(\theta, \delta)$, where $R(.,.)$ is the risk function.

In our problem, we are interested in robustness with respect to uncertainty in the prior $\pi \in \mathcal{C}$ for a class of priors $\mathcal{C}$ rather than robustness with respect to uncertainty for $\theta \in \Theta$. Thus we refer to $\mathcal{C}$-minimax rules. Berger (1985) discusses this in Section 4.7.6, calling it Gamma-minimax or Γ-minimax.

Under fairly general conditions, the minimax rule is also maximin, i.e., it is achieved for the case where one has the worst best-case performance. Thus, one is looking at situations where one is guaranteed a large loss. As an alternative to minimax risk or $\mathcal{C}$-minimax risk, one may find appealing consideration of $\mathcal{C}$-minimax regret.

Posterior regret

Suppose that the prior $\pi \in \mathcal{C}$. The regret of a decision for a prior π is the increase in expected loss incurred by using that decision instead of using the Bayes decision δ_π for that π. After all, one is certain to incur PEL (posterior expected loss) at least $E^\pi[L(\theta, \delta_\pi)|x]$. How much additional PEL does one incur by using δ instead of δ_π? The amount of the additional PEL is the posterior regret.

Consider squared error loss, $L(\theta, a) = (\theta - a)^2$. A standard result is that the Bayes rule is the posterior expectation of θ, i.e. $\delta_\pi(x) = E^\pi(\theta|x)$. Therefore the minimum PEL (posterior expected loss) under prior π is the PEL of δ_π, given by

$$E^\pi[(\theta - E^\pi(\theta|x))^2|x] = Var^\pi(\theta|x),$$

the posterior variance of θ under prior π.

For a decision rule δ, the posterior regret under π is the additional posterior expected loss or additional posterior risk incurred by using δ in place of δ_π. The

posterior regret of δ is the difference, $E^\pi[(\theta - \delta(x))^2|x] - E^\pi[(\theta - \delta_\pi(x))^2|x]$, which equals

$$[\delta(x) - E^\pi(\theta|x)]^2. \tag{8}$$

Estimating a parametric function $\psi(\theta)$

A similar result holds for a real-valued parametric function $\psi(\theta)$. If one uses the squared error loss function $(\psi(\theta) - a)^2$, the posterior regret of δ under π equals

$$[\delta(x) - E^\pi(\psi(\theta)|x)]^2. \tag{9}$$

There is a substantial literature on Bayesian robustness with classes of priors. Much of the work, influenced by Berger (1990) is concerned with finding the range of posterior expectations of parametric functions as the prior varies over a class $\mathcal{C}$. If the range is small, one has robustness. Suppose however, one wants to find a rule that is robust against uncertainty in the prior, which can vary over a class $\mathcal{C}$. The CMPR ($\mathcal{C}$-minimax posterior regret) rule (below) is an appealing answer.

The $\mathcal{C}$-minimax posterior regret (CMPR) rule

Simply put, for a class $\mathcal{C}$ of priors, the CMPR rule is the rule that minimizes the supremum posterior regret as the prior varies over the class $\mathcal{C}$. Let the parametric function of interest be $\psi(\theta)$ and let the prior $\pi \in \mathcal{C}$. Let $a = \inf_{\pi\in\mathcal{C}} E^\pi(\psi(\theta)|x)$ and let $b = \sup_{\pi\in\mathcal{C}} E^\pi(\psi(\theta)|x)$.

Note that a and b depend on x. We shall find the CMPR rule for each fixed x. We shall only consider $\delta(x) \in [a, b]$. The reason is as follows. Suppose that one does not restrict oneself to $\delta(x) \in [a, b]$. If $\delta(x) < a$, the supremum posterior regret is $(b - \delta(x))^2$ which is greater than $(b - a)^2$. On the other hand, for any $\delta(x) \in [a, b]$, the supremum posterior regret does not exceed $(b - a)^2$. Similarly, if $\delta(x) > b$, the supremum posterior regret is $(a - \delta(x))^2$ which is greater than $(b - a)^2$. Therefore, the $\mathcal{C}$-minimax posterior regret rule must be within the interval $[a, b]$.

Now, consider $\delta(x) \in [a, b]$. The supremum posterior regret is

$$\sup_{\pi \in \mathcal{C}} [\delta(x) - E^{\pi}(\psi(\theta)|x)]^2 = \max((\delta(x) - a)^2, (\delta(x) - b)^2),$$

which is at least as large as $(\frac{b-a}{2})^2$. The lower bound $(\frac{b-a}{2})^2$ on the supremum posterior regret is (uniquely) attained at $\delta(x) = (\frac{b+a}{2})$, which is therefore the CMPR rule. As noted above, a and b are respectively the infimum and supremum, over $\pi \in \mathcal{C}$, of $E^{\pi}(\psi(\theta)|x)$.

We shall show how to find the infimum and supremum of the posterior expectation (a and b) for the ε-contaminated class and the density ratio class. The average of the infimum and supremum gives us an estimate (the CMPR) that has an appealing robustness property with respect to uncertainty in the prior.

ε-contaminated class

For ε-contaminated class, the prior distribution of θ can be written as

$$\pi(\theta) = (1 - \varepsilon)\pi_0(\theta) + \varepsilon q : q \in Q, \tag{10}$$

where $\pi_0(\theta) \sim N(\mu, \tau^2)$ is the base elicited prior, q being a contamination and ε reflecting the uncertainty in π_0. The idea is that instead of a single prior π_0, we only believe with $(1 - \varepsilon)$ probability, that π_0 is the prior and allow ε probability on other distributions known as contaminations. One may allow q to vary within the Q, the class of all distributions or restrict the class Q to distributions having some properties in common with Q, e.g. unimodality and symmetry, which could be appropriate if π_0 is normal. Berger (1990) discussed the choice of Q and he identified four classes of particular interest, stated below:

$Q_A = \{\text{all distributions } q\}$

$Q_U* =$ { all distributions q: $\pi(\theta) = (1-\varepsilon)\pi_0(\theta) + \varepsilon q$ is unimodal}

$Q_U =$ { all unimodal distributions q, with mode μ}

$Q_{SU} =$ { all symmetric unimodal distributions q, with mode μ}.

Since the class of unimodal and symmetric distribution is relatively easy to calculate with and is a reasonably large class of priors, we consider the symmetric unimodal contaminations for q and define:

$$\text{Q= \{all symmetric unimodal distributions with mode } \mu\}$$

which effectively means that the contamination is a nonincreasing function of $\mid \theta - \mu \mid$. The absolute value $\mid \theta - \mu \mid$ is symmetric around μ, and in addition, the nonincreasing requirement makes it unimodal. For $q \in Q$, we can define

$$H^g(z) = \begin{cases} \frac{1}{2z}\int_{\theta-z}^{\theta+z} g(\theta) f_Y(\theta) d\theta & \text{if } z \neq 0 \\ g(\theta) f_Y(\theta) & \text{if } z = 0 \end{cases}$$

where $g(\theta)$ is a function of θ in which we are interested.

Berger and Sellke (1985) pointed out that any unimodal and symmetric distribution can be written as a mixture of uniform distributions. Consider the class of prior distributions Γ defined as:

$$\Gamma = \{\pi = (1-\varepsilon)\pi_0 + \varepsilon q : q \text{ is } U(\theta - z, \theta + z) \text{ for some } z > 0\}.$$

The range of posterior expectations as π varies over Γ is the same as the range of posterior expectations as the prior π varies over the ε-contaminated class with unimodal and symmetric contaminations. From Sivaganesan and Berger (1989), let $\rho^\pi(Y)$ denote the posterior expected value of $g(\theta)$ with respect to the prior π. Then

$$\sup_{\pi \in \Gamma} \rho^\pi(Y) = \sup_z \frac{a_0 + H^g(z)}{a + H_0(z)} \tag{11}$$

and

$$\inf_{\pi\in\Gamma}\rho^{\pi}(Y) = \inf_{z}\frac{a_0 + H^g(z)}{a + H_0(z)}, \tag{12}$$

where $a = (1-\varepsilon)m(Y|\pi_0)/\varepsilon, a_0 = a\rho_0^{\pi}(Y)$ and $H_0(z) = m(Y|q), H^g(z)$ is defined as above.

Now consider $g(\theta) = \theta$, thus denote $H^g(z) = H_1(z) = \frac{1}{2z}\int_{\theta-z}^{\theta+z}\theta\ f_Y(\theta)d\theta$. Let $Y|\theta \sim N(X\theta, \sigma^2 I_n)$, $\pi_0(\theta) \sim N(\mu, \tau^2 I_p)$, where σ, μ and τ are known. We are interested in the range of the posterior mean for θ_1. Then we have

$$a = \big(\frac{1-\varepsilon}{\varepsilon}\big)\frac{e^{\{-\frac{1}{2}(Y-X\mu)^T(\sigma^2 I_n + X\tau I_p X^T)^{-1}(Y-X\mu)\}}}{\sqrt{(2\pi)^n \mid \sigma^2 I_n + X\tau I_p X^T \mid}} \tag{13}$$

and

$$a_0 = a\rho_0^{\pi}(Y), \tag{14}$$

where $\rho_0^{\pi}(Y)$ is the posterior mean of θ_1.

Furthermore,

$$H_0(z) = \frac{1}{2z}\int_{\mu_1-z}^{\mu_1+z}\int\ldots\int\big(\frac{1}{\sqrt{2\pi\sigma^2}}\big)^n e^{(-\frac{(X\theta-Y)^T(X\theta-Y)}{2\sigma^2})}d\theta_n\ d\theta_{n-1}\ \ldots d\theta_1 \tag{15}$$

$$= \frac{1}{2z}\int_{\mu_1-z}^{\mu_1+z}\frac{1}{\sqrt{(2\pi)^n \mid \Sigma_L \mid}}e^{-\frac{1}{2}(Y-X_1\theta_1)^T\Sigma_L^{-1}(Y-X_1\theta_1)}d\theta_1 \tag{16}$$

$$= \frac{1}{2z}C\int_{\mu_1-z}^{\mu_1+z}\frac{1}{\sqrt{2\pi {C_\sigma}^2}}e^{-\frac{1}{2{C_\sigma}^2}(\theta_1-C_\mu)^2}d\theta_1 \tag{17}$$

and

$$H_1(z) = \frac{1}{2z}\int_{\mu_1-z}^{\mu_1+z}\theta_1\int\ldots\int\big(\frac{1}{\sqrt{2\pi\sigma^2}}\big)^n e^{(-\frac{(X\theta-Y)^T(X\theta-Y)}{2\sigma^2})}d\theta_n\ d\theta_{n-1}\ \ldots d\theta_1 \tag{18}$$

$$= \frac{1}{2z}\int_{\mu_1-z}^{\mu_1+z}\theta_1\frac{1}{\sqrt{(2\pi)^n \mid \Sigma_L \mid}}e^{-\frac{1}{2}(Y-X_1\theta_1)^T\Sigma_L^{-1}(Y-X_1\theta_1)}d\theta_1 \tag{19}$$

where

$$\Sigma_L = \sigma^2 I_n + X_{p-1}\tau I_{p-1} X_{p-1}^T, \quad X_{p-1} = (X_2, X_3, ..., X_{p-1})^T, \tag{20}$$

$$C_\mu = \frac{(Y^T\Sigma_L^{-1}X_1)}{X_1^T\Sigma_L^{-1}X_1},\ C_\sigma = \frac{1}{\sqrt{X_1^T\Sigma_L^{-1}X_1}},\ C = \frac{e^{\frac{1}{2}\left(\frac{(Y^T\Sigma_L^{-1}X_1)^2}{X_1^T\Sigma_L^{-1}X_1} - Y^T\Sigma_L^{-1}Y\right)}}{\sqrt{(2\pi)^{(n-1)} \mid \Sigma_L \mid (X_1^T\Sigma_L^{-1}X_1)}}, \tag{21}$$

Define t, u, v as follows:

$$t = \phi\big(\frac{\mu_1 + z - C_\mu}{C_\sigma}\big) + \phi\big(\frac{\mu_1 - z - C_\mu}{C_\sigma}\big), \tag{22}$$

$$u = \phi\big(\frac{\mu_1 + z - C_\mu}{C_\sigma}\big) - \phi\big(\frac{\mu_1 - z - C_\mu}{C_\sigma}\big), \tag{23}$$

$$v = \Phi\big(\frac{\mu_1 + z - C_\mu}{C_\sigma}\big) - \Phi\big(\frac{\mu_1 - z - C_\mu}{C_\sigma}\big). \tag{24}$$

Then we differentiate $\frac{a_0+H_1(z)}{a+H_0(z)}$to find out the extreme value of z which can minimize and maximize the expected posterior mean. The value of z is the solution of the equation

$$z = \frac{(vC_\mu - C_\sigma u)(Ct + 2C_\sigma a) - v(Ct\mu + 2a_0C_\sigma)}{2(auz + t(a\mu - a_0) + Cvu/2)}. \tag{25}$$

We can suppose the initial value of z larger or smaller than $\rho_0^\pi(Y)$ when maximizing or minimizing the value respectively and then do an iterative calculation for z.

Clearly, the range of the posterior mean grows as ε increases. In particular, when $\varepsilon = 0$, the posterior mean is a single value and when $\varepsilon = 1$, the range of posterior mean equals to the range of posterior expectations corresponding to $\pi(\theta) = q$ as q varies over Q.

Adjustment of the calculation of extreme values

There appears to be a mistake in the article of Sivaganesan and Berger (1989). The example 2.3.1 showed the range of posterior mean of θ when $X|\theta \sim N(\theta, \sigma^2)$ and $\pi_0(\theta) \sim N(\mu, \tau^2)$. When calculating the value of z to get the extreme values of

posterior mean, the author obtained the following equation involving z:

$$z = \frac{(vx - \sigma u)(2\sigma a + t) - v(t\mu + 2a_0\sigma)}{2(auz + t(a\mu - a_0) - \frac{uv}{2})}. \tag{26}$$

However, based on my calculation of the solution of $(\frac{a_0+H_1(z)}{a+H_0(z)})' = 0$, the equation about z should be:

$$z_{adjust} = \frac{(vx - \sigma u)(2\sigma a + t) - v(t\mu + 2a_0\sigma)}{2(auz + t(a\mu - a_0) + \frac{uv}{2})}. \tag{27}$$

We employed R to do the value of z satisfying the above equation for a certain x and substitute it into the $\frac{a_0+H_1(z)}{a+H_0(z)}$. The results are consistent with the authors' results shown in figure of the article and we can also verify the range of posterior mean by directly calculating the value of $\frac{a_0+H_1(z)}{a+H_0(z)}$ for a wide range of z with a certain value of x and plot it using the programming language R.

For example, let $\mu = 0$, $\sigma^2 = 1$, $\tau^2 = 1$ and $\varepsilon = 0.1$. When $x = 4$, we compute the value of $\frac{a_0+H_1(z)}{a+H_0(z)}$ for all $z \in R$. Figure 1 shows the result that the maximum value is 3.2395 obtained when $z = 6.15$ and the minimum value is 1.9574 when $z = 1.86$.

Also, note that when $z = 6.15$, the value of function $f = \frac{(vx-\sigma u)(2\sigma a+t)-v(t\mu+2a_0\sigma)}{2(auz+t(a\mu-a_0)-\frac{uv}{2})}$ is -14.09591 which does not satisfy the equation stated in Sivaganesan and Berger (1989). While, according to the adjusted function $f_{adujst} = \frac{(vx-\sigma u)(2\sigma a+t)-v(t\mu+2a_0\sigma)}{2(auz+t(a\mu-a_0)+\frac{uv}{2})}$, the value is 6.1548 and we can also prove that local minimum value of the error function $e = z - f_{adjust}$ is obtained at the point $z = 6.15$.

Density ratio class

The density ratio class was first proposed by DeRobertis and Hartigan (1981). Calculating the range of posterior value is relatively easy with the density ratio class. Berger (1990) gave the definition

$$\Gamma_{DR} = \{\pi : L(\theta) \leq \alpha\pi(\theta) \leq U(\theta) \text{ for some } \alpha > 0\},$$

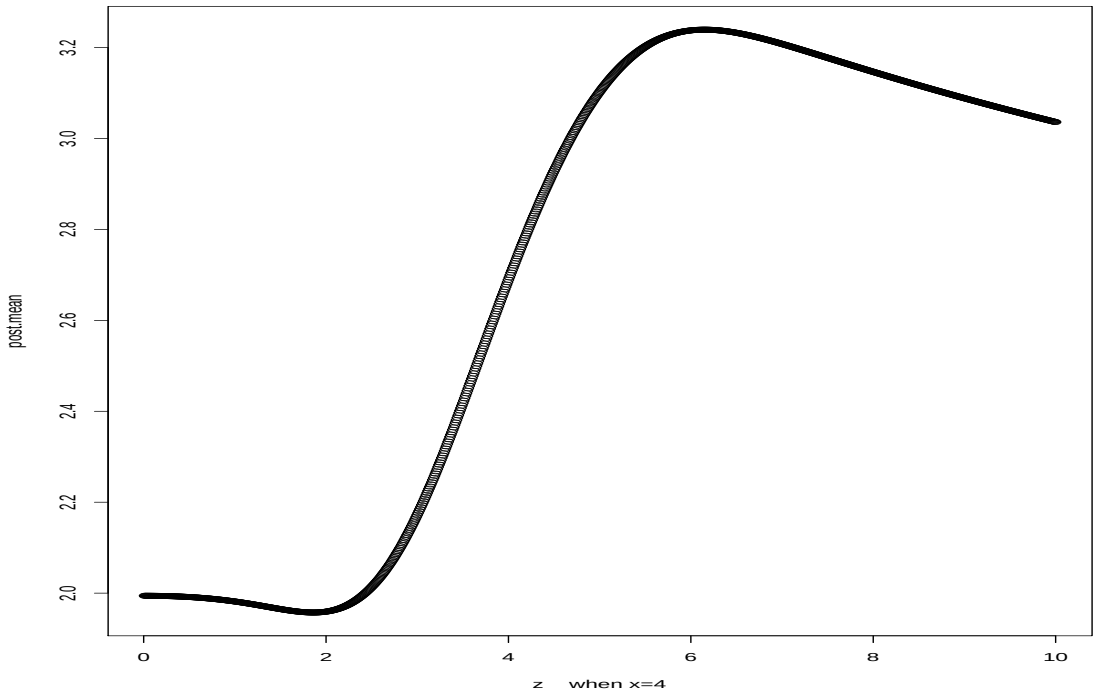

Figure 1. Plot of Posterior mean v.s. z when $x = 4$

where L and U are known non-negative density functions that do not integrate to 1. An alternative definition of this class is

$$\Gamma_{DR} = \{\pi\text{: } \tfrac{L(\theta)}{U(\theta')} \leq \tfrac{\pi(\theta)}{\pi(\theta')} \leq \tfrac{U(\theta)}{L(\theta')} \text{ for all } \theta,\, \theta'\}.$$

The posterior mean ranges between (δ^L,δ^U), where δ^L and δ^U denote the lower and upper bounds. δ^L and δ^U are the unique solutions of

$$\int_{\theta<\delta^L} (\theta - \delta^L) f(x|\theta) U(\theta) d(\theta) + \int_{\theta>\delta^L} (\theta - \delta^L) f(x|\theta) L(\theta) d(\theta) = 0 \tag{28}$$

$$\int_{\theta<\delta^U} (\theta - \delta^U) f(x|\theta) L(\theta) d(\theta) + \int_{\theta>\delta^U} (\theta - \delta^U) f(x|\theta) U(\theta) d(\theta) = 0 \tag{29}$$

respectively. Now consider $g(\theta) = \theta$. The minimum posterior expected value is given by:

$$\underline{E}^{\pi}(\theta|x) = inf_{\delta^L} \frac{\int_{\theta>\delta^L} \theta\ f(x|\theta) L(\theta) d(\theta) + \int_{\theta<\delta^L} \theta\ f(x|\theta) U(\theta) d(\theta)}{\int_{\theta>\delta^L} f(x|\theta) L(\theta) d(\theta) + \int_{\theta<\delta^L} f(x|\theta) U(\theta) d(\theta)} \tag{30}$$

and the maximum posterior expected value of θ is:

$$\overline{E}^{\pi}(\theta|x) = sup_{\delta^U} \frac{\int_{\theta>\delta^U} \theta \ f(x|\theta)U(\theta)d(\theta) + \int_{\theta<\delta^U} \theta \ f(x|\theta)L(\theta)d(\theta)}{\int_{\theta>\delta^U} f(x|\theta)U(\theta)d(\theta) + \int_{\theta<\delta^U} f(x|\theta)L(\theta)d(\theta)}. \tag{31}$$

Now, we will study robustness with respect to the prior distribution of θ_1. We will consider two classes of priors. First, we consider a density ratio class, instead of a single normal prior π_0.

We have $\pi_0(\theta_1) \sim N(\mu_1, \tau^2)$, and $f(Y|\theta) \sim \ N(X\theta, \sigma^2 I_n)$. let $L(\theta_1) = \pi_0$ and $U(\theta_1) = k\pi_0$, we can define the class of priors $\mathcal{C} = \{L(\theta) \leq \alpha\pi(\theta) \leq U(\theta)$ for some $\alpha > 0\}$.

Define $g(\theta_1|Y)$, the marginal likelihood function of θ_1 as follows:

$$g(\theta_1|Y) = \int ... \int (\frac{1}{\sqrt{2\pi\sigma^2}})^n e^{(-\frac{(X\theta-Y)^T(X\theta-Y)}{2\sigma^2})} \pi(\theta_p)\pi(\theta_{p-1})...\pi(\theta_2)d\theta_p \ d\theta_{p-1} \ ...d\theta_2 \tag{32}$$

$$= \frac{1}{\sqrt{(2\pi)^n \mid \Sigma_L \mid}} e^{-\frac{1}{2}(Y-X_1\theta_1)^T\Sigma_L^{-1}(Y-X_1\theta_1)} \tag{33}$$

$$= C\frac{1}{\sqrt{2\pi {C_\sigma}^2}} e^{-\frac{1}{2C\sigma^2}(\theta_1-C_\mu)^2}, \tag{34}$$

where $\mid \Sigma_L \mid$, C_σ and C_μ are the same as defined in (20) and (21).

Thus, the maximum value of posterior mean is:

$$\overline{E}^{\pi}(\theta_1|x) = sup_{\delta^U} \frac{k\int_{\theta_1>\delta^U} \theta_1 \ g(\theta_1|Y)\pi_0 d(\theta_1) + \int_{\theta_1<\delta^U} \theta_1 \ g(\theta_1|Y)\pi_0 d(\theta_1)}{k\int_{\theta_1>\delta^U} g(\theta_1|Y)\pi_0 d(\theta_1) + \int_{\theta_1<\delta^U} g(\theta_1|Y)\pi_0 d(\theta_1)}. \tag{35}$$

We employ R to do the integration and get the maximum value of posterior mean. The minimum value of posterior mean can be obtained in the same way.

Then, we use a multiple of the Cauchy distribution as the upper bound on the DR class and the normal distribution as the lower bound on the DR class. Thus let $L(\theta_1) = \pi_0$, and $U(\theta_1) = k\frac{\gamma_0}{\pi\gamma_0^2+\pi(\theta_1-\mu_1)^2}$. Thus, the upper bound of the posterior

mean is:

$$\overline{E}^{\pi} = sup_{\delta^U} \frac{\int_{\theta_1<\delta^U} \theta_1 g(\theta_1|Y)\pi_0 d\theta_1 + k\int_{\theta_1>\delta^U} \theta_1 g(\theta_1|Y)\frac{\gamma_0}{\pi\gamma_0^2+\pi(\theta_1-\mu_1)^2}d\theta_1}{\int_{\theta_1<\delta^U} g(\theta_1|Y)\pi_0 d\theta_1 + k\int_{\theta_1>\delta^U} g(\theta_1|Y)\frac{\gamma_0}{\pi\gamma_0^2+\pi(\theta_1-\mu_1)^2}d\theta_1}$$

$$= \frac{\int_{\theta_1<\delta^U} \theta_1 g(\theta_1|Y)\pi_0 d\theta_1 + kC\int_{\theta_1>\delta^U} \theta_1 \frac{1}{\sqrt{2\pi{C_\sigma}^2}} e^{-\frac{1}{2C_\sigma{}^2}(\theta_1-C_\mu)^2}\frac{\gamma_0}{\pi\gamma_0^2+\pi(\theta_1-\mu_1)^2}d\theta_1}{\int_{\theta_1<\delta^U} g(\theta_1|Y)\pi_0 d\theta_1 + kC\int_{\theta_1>\delta^U} \frac{1}{\sqrt{2\pi{C_\sigma}^2}} e^{-\frac{1}{2C_\sigma{}^2}(\theta_1-C_\mu)^2}\frac{\gamma_0}{\pi\gamma_0^2+\pi(\theta_1-\mu_1)^2}d\theta_1}.$$

Denote

$$S(\theta_1) = \frac{1}{\sqrt{2\pi{C_\sigma}^2}} e^{-\frac{1}{2C_\sigma{}^2}(\theta_1-C_\mu)^2}\frac{\gamma_0}{\pi\gamma_0^2+\pi(\theta_1-\mu_1)^2} \tag{36}$$

$$F(\theta_1) = \theta_1\frac{1}{\sqrt{2\pi{C_\sigma}^2}} e^{-\frac{1}{2C_\sigma{}^2}(\theta_1-C_\mu)^2}\frac{\gamma_0}{\pi\gamma_0^2+\pi(\theta_1-\mu_1)^2}, \tag{37}$$

we use the Monte Carlo method to approximate the integrals in the above formula.

The Monte Carlo approximation to an integral is as follows:

$$\int_\Theta h(x)d(F(x)) = E^X h(X) \approx \frac{1}{m}\sum_{i=1}^{m} h(X_i), \tag{38}$$

where $Eh(X) < \infty$ and $X_1, ..., X_m$ is a random sample from distribution F. Therefore

$$\mathcal{F} = kC\int_{\theta_1>\delta_U} F(\theta_1) \approx kC\frac{1}{m}\sum_{i=1}^{m}\theta_1\frac{1}{\sqrt{2\pi{C_\sigma}^2}} e^{-\frac{1}{2C_\sigma{}^2}(\theta_1-C_\mu)^2} \tag{39}$$

$$\mathcal{S} = kC\int_{\theta_1>\delta_U} S(\theta_1) \approx kC\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\sqrt{2\pi{C_\sigma}^2}} e^{-\frac{1}{2C_\sigma{}^2}(\theta_1-C_\mu)^2}, \tag{40}$$

where $\theta_1 \sim Cauchy(\gamma_0, \mu_1)$ and $\theta_1 > \delta^U$. Thus, we can employ R to do the calculation and the minimum value of posterior can be obtained as well with the same method.

Berger (1990) stated that when $L(\theta)$ and $U(\theta)$ are both normal distributions, the prior classes have the same tail behavior as π_0 and thus they may not include all the reasonable prior distributions. However, employing the Cauchy distribution can

overcome this limitation. The Cauchy distribution has a flat tail which allows for greater variation in tail behavior.

Chapter 4: Hypothesis Testing

For data X following a model $f(x|\theta)$, one specifies a null hypothesis $H_0 : \theta \in \Theta_0$ and an alternative hypothesis $H_1 : \theta \in \Theta_1$, where Θ_0 and Θ_1 are subsets of the parameter space Θ. In the frequentist approach, a test procedure is evaluated via the probabilities of type I error and Type II error. Type I and Type II error probabilities are the probabilities that the test chooses the wrong hypothesis, assuming that the null is true and that the alternative is true, respectively.

In other words, the key probabilities are probabilities of parts of the sample space, for a specified value of the parameter (or more precisely, that it falls in a subset of the parameter space, as specified in a hypothesis). However, many users of statistics, such as in psychology and medicine, just to name a few, routinely misinterpret frequentist error probabilities as the probabilities of the null and alternative hypotheses being true, given the observed data.

In Bayesian analysis, the process of choosing between the null and alternative hypotheses H_0 and H_1 is more direct. One calculates the posterior probabilities,

$$\alpha_0 = P(\theta \in \Theta_0|x), \quad \alpha_1 = P(\theta \in \Theta_1|x) = 1 - \alpha_0.$$

In other words, α_0 and α_1 are the posterior probabilities of the null and alternative hypothesis, respectively, given the data. One chooses between the hypotheses based on the value of α_0 and α_1. This illustrates the conceptual advantage of the Bayesian approach, since it calculates the probabilities of the hypotheses based on the data and prior knowledge.

The quantities α_0 and α_1 are functions of the data x and one might properly denote them by $\alpha_0(x)$ and $\alpha_1(x)$. However in Bayesian inference, one makes inferences conditional on the data. Thus the notation often does not show the dependence on the data, which having been observed, are considered fixed.

Besides posterior probabilities, there are additional quantities of interest that arise in Bayesian hypothesis testing. To define them, let us denote by π_0 and π_1, the prior probabilities of Θ_0 and Θ_1 respectively.

Definition The ratio of α_0/α_1 is called the posterior odds ratio of H_0 to H_1. Similarly, the prior odds ratio H_0 to H_1 is π_0/π_1, (Some authors prefer to drop the word "ratio" and call them simply "posterior odds" and "prior odds".). The Bayes factor in favor of Θ_0 is the ratio

$$B = \ BF_{01} \ = \ \frac{\alpha_0/\alpha_1}{\pi_0/\pi_1}.$$

In other words, the Bayes factor is the ratio of the posterior odds to the prior odds. It is the factor by which odds change as a result of observing the data, i.e., by using the data to update prior probabilities to posterior probabilities.

Odds are widely used; many people find them easier to think about than probabilities. If one says that $\alpha_0/\alpha_1 = 3$, a user readily understands that the null is three times as likely to be true as the alternative (based on both the prior knowledge as well as observed data). Thus the Bayes factor is sometimes interpreted as the odds for H_0 to H_1 that are given by the data. Consider the special case of testing simple hypotheses, i.e. $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$. Then one has,

$$\alpha_0 = \ \frac{\pi_0 f(x|\theta_0)}{\pi_0 f(x|\theta_0) + \pi_1 f(x|\theta_1)}, \qquad \alpha_1 = \ \frac{\pi_1 f(x|\theta_1)}{\pi_0 f(x|\theta_0) + \pi_1 f(x|\theta_1)},$$

and

$$B = \ BF_{01} = \ \frac{f(x|\theta_0)}{f(x|\theta_1)}.$$

Thus $B = BF_{01}$ is just the likelihood ratio of H_0 to H_1.

When one or both hypotheses are composite, the Bayes factor is a function of the prior input. The prior probabilities π_0 and π_1 cancel out, but the expression

involves integrals with respect to the (conditional) prior density on the hypotheses. For example, for simple null and composite alternative, one has prior probability π_0 of θ_0 and density $\pi_1 g(\theta)$ on Θ_1, where $\int_{\Theta_1} g(\theta)d\theta = 1$.

$$\alpha_0 = \frac{\pi_0 f(x|\theta_0)}{\pi_0 f(x|\theta_0) + \pi_1 \int_{\Theta_1} f(x|\theta)g(\theta)d\theta}, \qquad \alpha_1 = \frac{\pi_1 \int_{\Theta_1} f(x|\theta)g(\theta)d\theta}{\pi_0 f(x|\theta_0) + \pi_1 \int_{\Theta_1} f(x|\theta)g(\theta)d\theta},$$

and

$$B = \ BF_{01} = \ \frac{f(x|\theta_0)}{\int_{\Theta_1} f(x|\theta)g(\theta)d\theta}.$$

Decision theoretic interpretation

As stated above, one could choose between the hypotheses based on the posterior probabilities or posterior odds. A simple choice is to prefer the hypothesis that has the larger posterior probability. This can be given a simple decision theoretic interpretation. Suppose that one has a loss function that is zero if one chooses the correct hypothesis and is $K > 0$ if one chooses the wrong hypothesis. It is easy to see that rule of choosing the hypothesis with the larger posterior probability has the smallest risk. (If the posterior probabilities are both equal to 0.5, then one has the same risk, whichever hypothesis is chosen.)

Suppose instead, one has loss $K_0 > 0$ for selecting H_1 when H_0 is true and loss $K_1 > 0$ for selecting H_0 when H_1 is true, and the loss is still zero if the correct hypothesis is chosen. Then, one chooses H_0 if

$$\alpha_0 > \frac{K_1}{K_0 + K_1},$$

which is equivalent to (in terms of the posterior odds ratio)

$$\alpha_0/\alpha_1 > \frac{K_1}{K_0}.$$

Prior probabilities both equal to 0.5

A common choice is to take $\pi_0 = \pi_1 = 0.5$. In that case the Bayes factor equals the posterior odds α_0/α_1. One could state the above decision theoretic interpretation in terms of the Bayes factor.

For the case of $0, K$ loss, one would choose H_0 if $B = BF_{01} > 1$. For the $0, K_0, K_1$ loss, one would choose H_0 if $B = BF_{01} > K_1/K_0$.

We shall now discuss hypothesis testing for θ_1. Suppose

$$H_0 : \theta_1 = \theta_{10} \tag{41}$$

$$H_1 : \theta_1 \neq \theta_{10} \tag{42}$$

Suppose the prior probability of H_0 is π_0 and the prior probability of H_1 is $(1 - \pi_0)$. This is equivalent to comparing two models

$$H_0 : Y = \theta_{10}X_1 + \theta_2 X_2 \tag{43}$$

$$H_1 : Y = \theta_1 X_1 + \theta_2 X_2 \tag{44}$$

to find which one better fits the data. We partition $X = (X_1, X_2)$ and $\theta^T = (\theta_1^T, \theta_2^T)$. Consider $\theta_{10} = 0$ (remember that θ_{10} is a value of θ_1, as specified in H_0). We refer to H_0 as the reduced model and H_1 as the full model. The Bayes factor for selecting between two models H_0 and H_1 can be written as:

$$BF = \frac{f(x|\theta_0)}{\int_{\Theta_1} f(x|\theta)g(\theta)d\theta} = \frac{P(Y|H_0)}{P(Y|H_1)} \tag{45}$$

where

$$P(Y|H_0) = m(Y|H_0) = \frac{1}{\sqrt{2\pi \mid D_{p-1} \mid}} e^{-\frac{1}{2}(Y-X\theta)^T D_{p-1}^{-1}(Y-X\theta)}$$
$$P(Y|H_1) = m(Y|H_1) = \frac{1}{\sqrt{2\pi \mid D_p \mid}} e^{-\frac{1}{2}(Y-X\theta)^T D_p^{-1}(Y-X\theta)}.$$

Here $m(Y|H_0)$ is the marginal distribution of Y under H_0 and $m(Y|H_1)$ is the marginal distribution of Y under H_1.

Chapter 5: An Example: Linear Regression model of the psychological problem

The function of the psychological problem is:

$$\begin{aligned} DEP3 =& \beta_0 + \beta_1 DEP2 + \beta_2 FS2 + \beta_3 FS1 + \beta_4 REEMP \\ &+ \beta_5 (NEO \times FS2) + \beta_6 (NEO \times REEMP) + \epsilon_1, \end{aligned}$$

where $\epsilon_1 \sim N(0,1)$. The data set contains 280 observations and seven variables. There are seven unknown parameters in the equation including the intercept term. We use independent standard normal distributions as the priors on the parameters. Hence $\mu = 0$,$\tau^2 = 1$, $p = 7$, $n = 280$ and $\sigma^2 = 1$. The posterior expected values of the unknown parameters are:

$$\mu_{pos} = (-0.0302, 0.6497, 0.1783, -0.0362, -0.1142, -0.0821, 0.0947)$$

DEP3 and DEP2 are the depressive symptoms measured twice six weeks apart. The relationship of DEP2 and DEP3 is what we are interested in. Then we will do the robustness analysis for β_1.

ε-contaminated class

Suppose $\varepsilon = 0.1$, we applied the data on equation (11), (12), (25), and found that when $n = 285$, range of posterior mean is $(0.6434, 0.6502)$. Figure 2 depicts the value of posterior mean for z from 0.01 to 10. When $z = 0.64$, we have the minimum posterior expected value 0.6434. At the point $z = 0.9$, we have the maximum posterior mean, 0.6502.

It can be easily observed that the upper bound of the range is very close to the estimated posterior mean. Thus, we change the value of ε to look at the variation of the range. Table 1 shows the results for ε set to 0.1, 0.3, 0.5, 0.7 and 0.9.

Density ratio class

(i) Assume $L(\beta_1) = \pi_0 = \phi(\beta_1)$ and $U(\beta_1) = k\pi_0 = k\phi(\beta_1)$, Table 2 shows the

range for different k values.

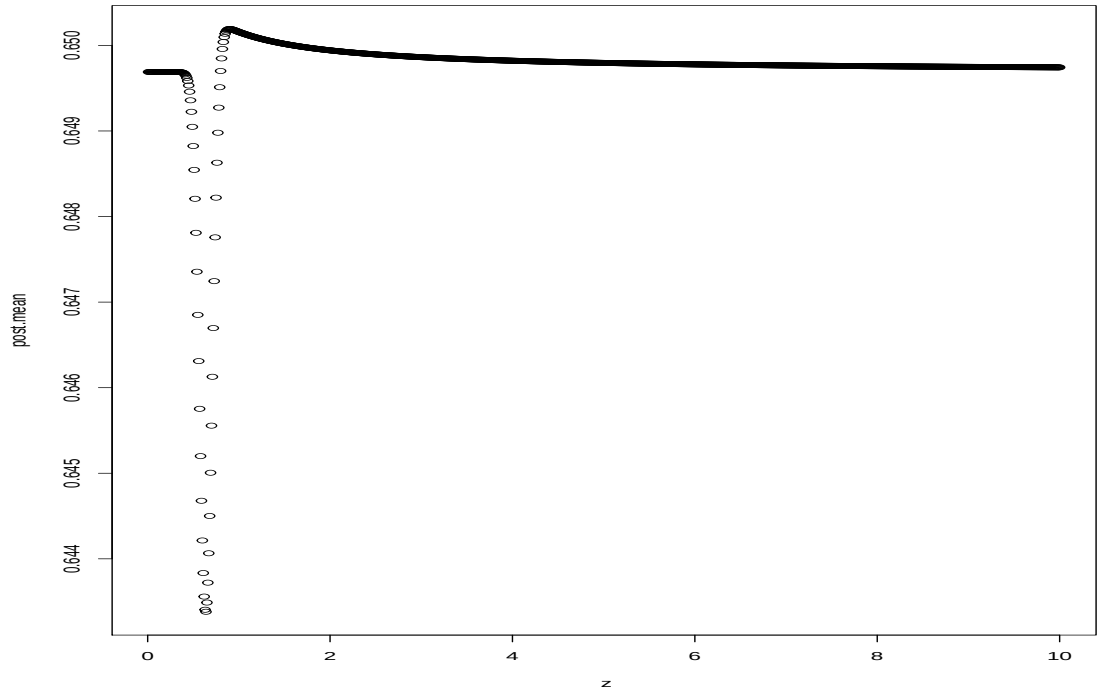

Figure 2. Plot of posterior mean v.s. z

Table 1

Range of posterior mean for various ε

ε	0.1	0.3	0.5	0.7	0.9
$\underline{E}^{\pi}$	0.6434	0.6305	0.6161	0.5977	0.5644
$\overline{E}^{\pi}$	0.6502	0.6510	0.6517	0.6522	0.6527

(ii) Assume $L(\beta_1) = \pi_0 = \phi(\beta_1)$ and $U(\beta_1) = kCauchy(0,1) = k\frac{1}{\pi+\pi\theta_1^2}$. The extreme value of posterior mean can be obtained for various k values. In the density ratio class, $U(\theta_1)$ must be greater or equal to $L(\theta_1)$. Therefore k must be chosen large enough so that $kCauchy(0,1)$ is pointwise at least as large as the standard normal density. By examining the ratio of the Cauchy to the normal density, we can show that k should be greater than or equal to $\frac{\sqrt{2\pi}}{\sqrt{e}} = 1.52$. Thus, we choose $k = 2, 3, 5, 7, 10$ to observe the impact of k value on the variation of the range of the posterior mean

Table 2

Range of posterior mean for various k when $U(\theta)$ is normal distribution

k	1.25	1.5	2	3	5	10
$\underline{E}^{\pi}$	0.6446	0.6400	0.6322	0.6204	0.6054	0.5878
$\overline{E}^{\pi}$	0.6548	0.6593	0.6671	0.6789	0.6939	0.7116

Table 3

Range of posterior mean for various k when $U(\theta)$ is Cauchy distribution

k	2	3	5	7	10
$\underline{E}^{\pi}$	0.6413	0.6303	0.6152	0.6054	0.5957
$\overline{E}^{\pi}$	0.6567	0.6675	0.6824	0.6923	0.7021

and Table 3 shows the result.

Chapter 6: Summary and Discussion

In this thesis, we discuss Bayesian inference on one of the unknown parameters of the general linear regression model. We considered two classes of priors, the ε-contaminated class and the density ratio class. In robust Bayesian analysis, one recognizes possible uncertainty in the prior, and considers a class of priors instead of a single prior.

We choose the symmetric and unimodal ε-contaminated class to study the robustness since it not only contains most of the reasonable prior distributions but also, the computations are not too difficult to carry out. From the results shown above, the range of posterior mean becomes larger as ε increases. However the range is relatively small and does not change much with ε. It is possible that the analysis is truly robust to prior uncertainty. However, one may be concerned that the class of priors is too small in the sense that it does not properly represent uncertainty in the prior. A wider prior class may need to be considered for the analysis. A possibility is to consider the class Q of all unimodal distributions. It is possible to go further and take Q to be the class of all distributions, but that may be too large a class.

With the density ratio class, we get a larger range of posterior means than with the ε-contaminated classes that we studied. This indicates that we have a greater variety of priors. For the two cases of the upper prior bound U, multiples of Cauchy and multiples of normal, the results are similar.

It would be interesting to calculate the range of posterior expectation for variation in the prior for multiple parameters. Thus in the density ratio class, L and U would be functions of more than one θ.

Reference

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag New York.

Berger, J. (1990). Robust bayesian analysis: sensitivity to the prior. *Journal of Statistical Planning and Inference*, 25:303–328.

Berger, J. and Berliner, L. M. (1986). Robust bayes and empirical bayes analysis with ε-contaminated priors. *The Annals of Statistics*, 14:461–486.

Berger, J. and Sellke, T. (1985). Testing a point null hypothesis: The irreconcilability of p-values and evidence. *Journal of the American Statistical Association*, pages 112–139.

Bose, S. (1994a). Bayesian robustness with more than one class of contaminations. *Journal of statistical planning and inference*, 40:177–187.

Bose, S. (1994b). Bayesian robustness with mixture classes of priors. *The Annals of Statistics*, 22:652–667.

Broemeling, L. (1985). *Bayesian analysis of linear models*. M. Dekker, New York.

DeRobertis, L. and Hartigan, J. A. (1981). Bayesian inference using intervals of measures. *The Annals of Statistics*, 9(2):235–244.

Geweke, J. and Petrella, L. (1998). Prior density-ratio class robustness in econometrics. *Journal of Business & Economic Statistics*, 16:469–478.

Good, I. J. (1950). *Probability and the weighing of evidence*. Charles Griffin, New York.

Good, I. J. (1980). Some history of the hierarchical bayesian methodology. *Trabajos de Estadistica Y de Investigacion Operativa*, 31:489–519.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistics Society. Series B (Methodological)*, 34:1–41.

Sivaganesan, S. and Berger, J. (1989). Ranges of posterior measures for priors with unimodal contaminations. *The Annals of Statistics*, 17:868–889.