Multiple Hypothesis Testing and Multiple Outlier Identification Methods

A Thesis Submitted to the
College of Graduate Studies and Research
in Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy
in the
Department of Mathematics and Statistics
University of Saskatchewan
Saskatoon, Saskatchewan

By
Yaling Yin

March 2010

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Mathematics and Statistics
University of Saskatchewan
Saskatoon, Saskatchewan
Canada
S7N 5E6

Abstract

Traditional multiple hypothesis testing procedures, such as that of Benjamini and Hochberg, fix an error rate and determine the corresponding rejection region. In 2002 Storey proposed a fixed rejection region procedure and showed numerically that it can gain more power than the fixed error rate procedure of Benjamini and Hochberg while controlling the same false discovery rate (FDR). In this thesis it is proved that when the number of alternatives is small compared to the total number of hypotheses, Storey's method can be less powerful than that of Benjamini and Hochberg. Moreover, the two procedures are compared by setting them to produce the same FDR. The difference in power between Storey's procedure and that of Benjamini and Hochberg is near zero when the distance between the null and alternative distributions is large, but Benjamini and Hochberg's procedure becomes more powerful as the distance decreases. It is shown that modifying the Benjamini and Hochberg procedure to incorporate an estimate of the proportion of true null hypotheses as proposed by Black gives a procedure with superior power.

Multiple hypothesis testing can also be applied to regression diagnostics. In this thesis, a Bayesian method is proposed to test multiple hypotheses, of which the ith null and alternative hypotheses are that the ith observation is not an outlier versus it is, for $i = 1, \cdots, m$. In the proposed Bayesian model, it is assumed that outliers have a mean shift, where the proportion of outliers and the mean shift respectively follow a Beta prior distribution and a normal prior distribution. It is proved in the thesis that for the proposed model, when there exists more than one outlier, the marginal distributions of the deletion residual of the ith observation under both null and alternative hypotheses are doubly noncentral t distributions. The "outlyingness" of the ith observation is measured by the marginal posterior probability that the ith observation is an outlier given its deletion residual. An importance sampling method is proposed to calculate this probability. This method requires the computation of the density of the doubly noncentral F distribution and this

is approximated using Patnaik's approximation. An algorithm is proposed in this thesis to examine the accuracy of Patnaik's approximation. The comparison of this algorithm's output with Patnaik's approximation shows that the latter can save massive computation time without losing much accuracy.

The proposed Bayesian multiple outlier identification procedure is applied to some simulated data sets. Various simulation and prior parameters are used to study the sensitivity of the posteriors to the priors. The area under the ROC curves (AUC) is calculated for each combination of parameters. A factorial design analysis on AUC is carried out by choosing various simulation and prior parameters as factors. The resulting AUC values are high for various selected parameters, indicating that the proposed method can identify the majority of outliers within tolerable errors. The results of the factorial design show that the priors do not have much effect on the marginal posterior probability as long as the sample size is not too small.

In this thesis, the proposed Bayesian procedure is also applied to a real data set obtained by Kanduc *et al.* in 2008. The proteomes of thirty viruses examined by Kanduc *et al.* are found to share a high number of pentapeptide overlaps to the human proteome. In a linear regression analysis of the level of viral overlaps to the human proteome and the length of viral proteome, it is reported by Kanduc *et al.* that among the thirty viruses, human T-lymphotropic virus 1, Rubella virus, and hepatitis C virus, present relatively higher levels of overlaps with the human proteome than the predicted level of overlaps. The results obtained using the proposed procedure indicate that the four viruses with extremely large sizes (Human herpesvirus 4, Human herpesvirus 6, Variola virus, and Human herpesvirus 5) are more likely to be the outliers than the three reported viruses. The results with the four extreme viruses deleted confirm the claim of Kanduc *et al.*

Acknowledgements

First of all, the author would like to express her sincere appreciation and gratitude to her co-supervisors, Dr. M. G. Bickis and Dr. C. E. Soteros for their invaluable guidance, patience, encouragement and support throughout the course of this research work and in the preparation of this thesis. The author has greatly benefited from Dr. M. Bickis and Dr. C. E. Soteros in-depth knowledge of statistics and programming.

The author's thanks are extended to the advisory committee members, Dr. A. Kusalik, Dr. W. H. Laverty, Dr. J. R. Martin, and Dr. R. Srinivasan for their advice and guidance.

The author would like to take this opportunity to acknowledge the constant encouragement, patience and support from her parents, Guoqiang Yin and Anna Cheng and husband, Po Hu throughout her Ph.D. program. The author presents this thesis as a gift to them.

Financial assistances from the College of Graduate Studies and Research in the form of a Ph. D. Scholarship, from the Department of Mathematics and Statistics at the University of Saskatchewan in the form of a Graduate Teaching Assistantship, and from her co-supervisors' NSERC (Natural Science and Engineering Research Council of Canada) Discovery Grants in the form of a Research Assistantship are gratefully acknowledged.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AFDR	Adaptive FDR controlling procedure
AUC	Area under ROC Curve
BH	Benjamini and Hochberg procedure
EL	extremely large viruses
FDP	False discovery proportion
FDR	False discovery rate
FEL	four extremely large (viruses)
FNP	False non-discovery proportion
FNR	False non-discovery rate
FPR	False positive rate
FSL	Fixed significance level procedure
FWER	Family-wise error rate
k-FWER	k-Family-wise error rate
KI	Kanduc *et al.* identified (viruses)
MCMC	Markov chain Monte Carlo
MMLE	Marginal maximum likelihood estimate
MPT	Most powerful test
MVN	multivariate normal (distribution)
PCER	Per comparison error rate
PFER	Per family error rate
pFDR	Positive false discovery rate
pFNR	Positive false non-discovery rate
PFP	Proportion of false discoveries (positives)
ROC	Receiver operating characteristic
TPR	True positive rate
UMPT	Uniformly most powerful test

List of Symbols

$\sim$	a random variable follows a certain distribution
ϕ	the density of the standard normal distribution
Φ	the CDF of the standard normal distribution
m	the total number of tests
n	the number of iterations in a simulation study or the number of random samples generated by using a Monte Carlo method
m_0	the number of true null hypotheses
m_1	the number of false null hypotheses
π_0	the proportion of true null hypotheses
π_1	the proportion of false null hypotheses
$\widehat{\pi}_0$	an estimator of π_0
h_{i0}	the null hypothesis for ith test
h_{i1}	the alternative hypothesis for ith test
H_i	the index of the alternative for ith test
$\boldsymbol{H}$	the vector of indices H_i
$\{0,1\}^m$	the m-dimensional vector with elements equal to 0 or 1
θ_i	the distribution parameter (or parameters) of ith test statistic
Θ	the parameter space of θ_i
Θ_{i0}	the parameter space of θ_i when h_{i0} is true
Θ_{i1}	the parameter space of θ_i when h_{i0} is false
d_i	an indicator of rejecting h_{i0}
P_i	the p-value of ith test
$P_{(i)}$	the ordered p-value of ith test
$h_{(i)}$	the null hypothesis corresponding to $P_{(i)}$
$\boldsymbol{\beta}$	the vector of unknown parameters in linear regression
$\hat{\boldsymbol{\beta}}$	the vector of least square estimators of $\boldsymbol{\beta}$
ε	the random errors of ith observation in linear regression
$\boldsymbol{\varepsilon}$	the vector of random errors
r_i	the residual for ith observation in linear regression
$\boldsymbol{r}$	the vector of residuals

$R(\hat{\boldsymbol{\beta}})$	the sum of residual squares
G	the hat matrix
g_i	the ith diagonal element of G
g_{ij}	the ijth diagonal element of G
s^2	the least square estimator of the variance of the random errors
r_i'	the standardized residual for ith observation in linear regression
$\boldsymbol{H}_{(i)}$	the vector of indices of outliers with ith observation deleted
$\hat{\boldsymbol{\beta}}_{(i)}$	the vector of least square estimates of $\boldsymbol{\beta}$ with ith observation deleted
$\boldsymbol{r}_{(i)}$	the vector of residuals with ith observation deleted
$G_{(i)}$	the hat matrix with ith observation deleted
$g_{(i)j}$	the jth diagonal element of $G_{(i)}$
$g_{(i)jk}$	the jkth diagonal element of $G_{(i)}$
$s^2_{(i)}$	the least square estimator of the variance of the random errors with ith observation deleted
r_i^*	the deletion residual for ith observation
$P(H_i = 1\|r_i^{*2})$	the marginal posterior probability that ith observation is an outlier given its deletion residual
$\chi^{2\prime}_\nu(\eta)$	the noncentral χ^2 random variable with degrees of freedom ν and the noncentrality parameter η
$t'_\nu(\xi)$	the singly noncentral t random variable with degrees of freedom ν and noncentrality parameter ξ
$t''_\nu(\xi,\eta)$	the doubly noncentral t random variable with degrees of freedom ν and noncentrality parameters ξ and η
$F'_{\nu_1,\nu_2}(\zeta)$	the singly noncentral F random variable with degrees of freedom ν_1, ν_2 and noncentrality parameter ζ
$F''_{\nu_1,\nu_2}(\zeta,\eta)$	the doubly noncentral F random variable with degrees of freedom ν_1, ν_2 and noncentrality parameters ζ and η

Chapter 1

Introduction

1.1 Motivation

Multiple hypothesis testing, the testing of more than one hypothesis simultaneously, has a broad range of applications such as factorial design [31, 51, 63, 87], regression diagnostics [1, 33], analyzing DNA microarray data [27, 28, 29, 30, 53, 57, 90, 92, 93, 97, 98], and classifying regions in image data [42]. Multiple hypothesis testing is a subfield of *multiple inference* (or simultaneous inference, *multiple comparison*), which includes both multiple estimation and multiple testing. Since the late 1940's, many statisticians have been interested in this field. For example, Shaffer [87] lists more than one hundred journal articles and books about simultaneous inference published before 1995.

Most recently renewed interest in multiple hypothesis testing stems from biological examples [27, 28, 29, 30, 53, 57, 90, 92, 93, 97, 98]. Modern biological technologies are growing quickly with the help of computers and result in many large, complex data sets. For instance, DNA microarrays, a novel biological technology, can be used to measure the expression levels for thousands to tens-of-thousands of genes simultaneously [29, 30, 90, 92, 93, 98]. A common goal of microarray experiments is to identify the genes that show changes in expression level across two or more different biological conditions, for example, the same cell type in a healthy and diseased state. Thus we can assign each gene a null hypothesis that there is no differential gene expression, versus an alternative hypothesis that there is a change in gene expression level. For each gene, its expression level data can be reduced to a test statistic for that gene. But with thousands of genes, we need to test thousands of hypotheses simultaneously. Although the number of genes is very large, the number of available arrays is small due to the cost of microarray experiments. A medium-size microarray study may obtain a hundred arrays with thousands of genes per sample, while a large clinical study, which is more traditional than the microarray

study, can collect 100 data items per unit for thousands of units [99]. Hence traditional multiple testing methods may not be appropriate for analyzing microarray data. Moreover, the high-dimensional multivariate distributions of associated test statistics involve many unknown parameters as well as complex and unknown dependence structures among the statistics. This motivates the rapid development of multiple hypothesis testing techniques. For many microarrays, the proportion of differentially expressed genes is expected to be small, yet identification of them is important [29, 30, 92, 93, 97, 98]. Therefore, in order to identify as many differentially expressed genes as possible, misidentification of a few identically expressed genes is tolerable [92, 93, 98].

Multiple hypothesis testing can also be applied to regression diagnostics. Regression is a statistical tool to analyze data generated in many fields of study. There exist many books introducing regression models and applications, for example, [26, 72, 73, 74]. The standard results of regression analysis are only valid for "clean" data, which satisfy certain assumptions. However, real data are usually contaminated and contain observations which violate the model assumptions. Such observations are called *outliers* in regression analysis. They may or may not be observable. Other authors define outliers as observations which are numerically distant from the rest of the data [7]. Such observations are referred to as *apparent outliers* in this thesis. In order to distinguish the outliers from the other observations, I call an observation *atypical* if it is an outlier and *typical* if not in this thesis. Sometimes, atypical observations are more interesting than typical ones. One motivating example is a dataset obtained from Kusalik [58] and this dataset was published in Kanduc *et al.* [56]. This dataset contains 30 viral proteomes, which are shown to present a high number of pentapeptide overlaps to the human proteome, and my goal is to identify viruses that share significantly higher or lower level of pentapeptide overlaps with human proteome than the predicted level of overlaps from the linear regression model. The proteome is the full complement of proteins produced by a particular genome. Such viruses are examples of outliers and are more interesting than the other viruses in genomic studies. A powerful method for identifying a single outlier is introduced in Cook and Weisberg [25] and Atkinson [1]. However, when there is more than one outlier, they may hide the effect of each other and lead to a "masking" problem. The problem of identifying multiple outliers has been studied in the past. Hadi and Simonoff [44] gave a review of early works of multi-step methods. Recent works on multiple diagnostics are by Atkinson [2],

Hadi [44], Hadi and Simonoff [45], and so on. In fact, the multiple deletion diagnostics problem in regression analysis can be viewed as a problem of multiple hypothesis testing. Each observation can be assigned a null hypothesis that this observation does follow the assumed distribution, and an alternative hypothesis that it is an outlier which follows a distribution different from the null distribution. By assuming appropriate distributions for typical and atypical observations, I construct a Bayesian multiple hypothesis testing method for regression diagnostics.

To solve the multiple testing problem, I need to state it mathematically. Hypothesis testing is naturally a frequentist concept. However, Bayesian methods can also be applied to solve the problem of hypothesis testing, especially when it is considered from a decision-theoretic viewpoint. In Section 1.2, I describe the problem of single hypothesis testing from both frequentist and Bayesian points of view. The statement of the multiple hypothesis testing problem and definitions of various error measures are given in Section 1.3.1. Multiple hypothesis testing methods based on sequential p-values are introduced in Section 1.3.3. In Section 1.3.4, some commonly used mixture models are introduced. In Section 1.4, I introduce the regression diagnostics problem and state it as a multiple hypothesis testing problem. A review of regression diagnostics methods is also given in this section. Finally, the scope of the thesis is presented in Section 1.5.

1.2 Single Hypothesis Testing Problem

For the problem of *single hypothesis testing,* only one test is considered, of which there is only one null hypothesis versus one alternative hypothesis, respectively denoted by h_0 and h_1. One wishes to decide whether h_0 or h_1 is true. The choice lies between only two decisions: accepting or rejecting h_0. A decision procedure for such a problem is called a *test* of the hypothesis h_0 (Lehmann [59]). The decision is based on a sequence of observations of a random variable X which has probability distribution P_θ. We assume P_θ belongs to a distribution class $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$, where Θ is the parameter space of θ. The non-empty subsets of Θ under h_0 and h_1 are denoted respectively by Θ_0 and Θ_1, where $\Theta_0 \cup \Theta_1 = \Theta$, $\Theta_0 \cap \Theta_1 = \emptyset$. Then P_θ either belongs to the class of nulls or the class of alternatives. If $\Theta_0 = \{\theta_0\}$, $\mathcal{P}$ is called simple, otherwise it is said to be composite. Thus the statistical

hypothesis testing problem can be written as

$$h_0 : \theta \in \Theta_0 \text{ vs. } h_1 : \theta \in \Theta_1. \tag{1.1}$$

In this thesis, we are usually interested in the case that h_0 is not true. For example, we are interested in the genes with different expression levels between control and treatment. Therefore, sometimes "rejecting h_0" is also called "discovering h_1". Let H be an indicator of the alternative hypothesis, that is $H = 0$ when the null hypothesis is true and $H = 1$ when it is false. H is considered as an unknown parameter in frequentist methods, but a random variable in Bayesian methods. Typically it is assumed that the distribution of X is known under both null and alternative hypotheses.

To attain a decision, a *test statistic* Y is defined, which is a function of observations of X, from which a realization of Y, denoted by y, is obtained. Let the distribution function of Y be $F_\theta(y)$ with associated probability density function $f_\theta(y)$.

For the special hypothesis testing problem, in which both null and alternative hypotheses are simple, that is,

$$h_0 : \theta \in \Theta_0 = \{\theta_0\} \text{ vs. } h_1 : \theta \in \Theta_1 = \{\theta_1\}, \tag{1.2}$$

it is assumed that

$$Y \mid H = 0 \sim F_0 \text{ and } Y \mid H = 1 \sim F_1, \tag{1.3}$$

where the distribution functions F_0 and F_1 respectively possess densities f_0 and f_1.

1.2.1 Frequentist Hypothesis Testing

In a traditional frequentist hypothesis testing procedure, we specify a *rejection region* Γ (or critical region). Rejecting or accepting the null hypothesis depends on whether the observed value y falls into the critical region or not. Since the decision from the hypothesis test is uncertain, one may attain the correct one, or may commit one of two errors: *Type I error*, that is, rejecting the null hypothesis when it is true, or *Type II error*, that is, accepting the null hypothesis when it is false. Type I error is also called *false rejection*, *false discovery* or *false positive*, and type II error is also called *false non-rejection*, or *false non-discovery*, or *false negative*. It is desirable to minimize the probabilities of the two types of errors at the same time. Unfortunately, the two probabilities cannot be controlled

simultaneously for a given y. As the type I error rate $\Pr_\theta(Y \in \Gamma)$, $\theta \in \Theta_0$ decreases, type II error rate $\Pr_\theta(Y \notin \Gamma)$, $\theta \in \Theta_1$ increases, and then power decreases, where *power* is defined to be the probability of correctly rejecting the false null, $\Pr_\theta(Y \in \Gamma)$, $\theta \in \Theta_1$, which also equals 1− type II error rate. (For convenience, in this thesis I use $\Pr_\theta(Y \in \Gamma \mid H = 0)$ and $\Pr_\theta(Y \notin \Gamma \mid H = 1)$ to denote respectively type I error rate and type II error rate, though H is fixed, i.e. $\Pr(H = 0)$ is an element of $\{0, 1\}$, for frequentists). A traditional way to control these errors is to assign an upper bound, called the significance level, to the type I error rate and to attempt to minimize the type II error rate, or equivalently to maximize power. Such an upper bound is called *the level of significance* of a test procedure, which is a number α between 0 and 1, and the actual probability of a type I error is called the *size* of the test procedure. Constraining the type I error rate below a given significance level is called "controlling" the type I error rate. Therefore, after observing a y, one could determine whether the hypothesis is accepted or rejected at a given significance level α. The rejection regions corresponding to the level α are denoted by Γ_α satisfying $\Pr_\theta(Y \in \Gamma_\alpha \mid H = 0) \leq \alpha$. A test is said to be *conservative* if it attains a type I error rate that is strictly smaller than α (often considerably), and it results in a loss of power.

Suppose that $\Theta_1 = \{\theta_1\}$. A level α test of $h_0 : \theta \in \Theta_0$ versus $h_1 : \theta = \theta_1$ is called a *most powerful test* (*MPT*) at level α if it has the greatest power among all level α tests. The Neyman-Pearson fundamental lemma, which can be found in [59], provides the most powerful test for the simple null hypothesis versus simple alternative hypothesis testing problem.

In the composite alternative hypothesis case, a procedure which maximizes the power for all $\theta \in \Theta_1$ while controlling the type I error rate at level α is called a *uniformly most powerful test (UMPT)*. However, UMP tests do not always exist except for a real-parameter family of densities possessing a monotone likelihood ratio [59]. Generally, UMPT's do not usually exist for multidimensional parameters.

In statistical hypothesis testing, a *p-value* is the probability of obtaining a value of the test statistic Y at least as extreme as its actual observation y given that the null hypothesis is true. The p-value is a measure of how strongly the data contradict the null hypothesis. Therefore, a p-value smaller than or equal to a given significance level indicates rejection of the null hypothesis. A more precise definition of p-value is given below. This definition is modified from Lehmann [59].

Definition 1.2.1 *Consider a family of tests of $h_0 : \theta \in \Theta_0$, with level α rejection regions Γ_α such that (a) $\Pr_\theta(Y \in \Gamma_\alpha | H = 0) \leq \alpha$ for any $\alpha \in (0,1)$ and for any $\theta \in \Theta_0$, and (b) $\Gamma_\alpha \subseteq \Gamma_{\alpha'}$ whenever $\alpha \leq \alpha'$. The smallest significance level at which the null hypothesis is rejected for the given y is called the p-value of a realization y of the statistic Y, that is*

$$p\text{-}value(y) \equiv \inf\{\alpha : y \in \Gamma_\alpha\} \tag{1.4}$$

When Y is continuous, the equality holds in the assumption (a) of definition 1.2.1, and thus we have:

Definition 1.2.2 *For a continuous statistic, the p-value of an observation y is*

$$p\text{-}value(y) = \min_{\{\alpha : y \in \Gamma_\alpha\}} \Pr_\theta(Y \in \Gamma_\alpha \mid H = 0)\,, \tag{1.5}$$

i.e. given the set of nested rejection regions, the p-value is the minimum type I error that can occur when rejecting the null hypothesis with y.

Since a p-value is a function of the statistic Y, it is a statistic as well. Let P denote p-value(Y) and p denote an observation of P. The Lemma given below provides the following important property of the null distribution of a p-value. I also give a proof of this Lemma which is a modification of the proof of Lemma 1.1 in [60] because this lemma and its proof are important for the rest of this chapter.

Lemma 1.2.1 *For a p-value as defined in 1.2.1, we have for any $\theta \in \Theta_0$*
(1).

$$\Pr_\theta(P \leq u \mid H = 0) \leq u, \tag{1.6}$$

(2).

$$\Pr_\theta(P \leq u \mid H = 0) \geq \Pr_\theta(Y \in \Gamma_u | H = 0) \tag{1.7}$$

Therefore, if the equality holds for all $\theta \in \Theta_0$ in the assumption (a) of definition 1.2.1, then P is uniformly distributed on $(0,1)$ when $H = 0$.

Proof. [Modified from the proof of Lemma 1.1 in [60].]
(1).
Since $p = p$-value$(y) = \inf\{\alpha : y \in \Gamma_\alpha\}$, then for any $\varepsilon > 0$, there exists an $\alpha < p + \varepsilon, \alpha \in$

$(0,1)$ such that $\Pr_\theta(Y = y \in \Gamma_\alpha \,|H = 0) \leq \alpha$, and $\Gamma_\alpha \subseteq \Gamma_{p+\varepsilon}$. Thus the event $\{P \leq u\}$ implies $\{Y \in \Gamma_{u+\varepsilon}\}$, and therefore

$$\Pr_\theta(P \leq u \,|H = 0) \leq \Pr_\theta(Y \in \Gamma_{u+\varepsilon} \,|H = 0) \leq u + \varepsilon. \tag{1.8}$$

The second inequality is obtained by the assumption (a) of definition 1.2.1. Then letting $\varepsilon \to 0$, we have 1.6.

(2).

Note that $\{Y \in \Gamma_u \,|\, H = 0\} \subseteq \{P \leq u \,|\, H = 0\}$, and hence 1.7 follows. ∎

Hence, when the simple null hypothesis is true, the p-value of a continuous statistic is uniformly distributed on $(0,1)$. The rejection region based on the p-value is simply $\{p \mid p \leq \alpha\}$ for a given significance level α.

A "good" frequentist method for single hypothesis testing obtains a conclusion satisfying the given significance level with high power.

Example 1.2.1 *Let $X_1, X_2, \cdots, X_m$ be an i.i.d. sample from a normal population $N(\mu, \sigma)$ with known variance σ^2. Suppose one is interested in testing the population mean for $h_0 : \mu = 0$ vs. $h_1 : \mu = \mu_1 > 0$ at the significance level α.*

The test statistic is $Y = \frac{\overline{X}}{\sigma/\sqrt{m}}$, which has the standard normal distribution given that h_0 is true. The level α rejection region is $\Gamma = \{y \mid y > \Phi^{-1}(1-\alpha)\}$, where Φ is the standard normal distribution function. Then if an observation $y \in \Gamma$, the null hypothesis is rejected. The power of this test is $\Phi(\Phi^{-1}(1-\alpha)+\mu_1)$. Since $\Pr_{\theta_0}\{Y = \Phi^{-1}(1-\alpha)\} = 0$, then by the Neyman-Pearson fundamental lemma this test is a UMPT for testing $h_0 : \mu = 0$ against $h_1 : \mu = \mu_1 > 0$ at level α.

The p-value of the observation y is calculated as $1 - \Phi(y)$. A calculated p-value smaller than or equal to the target α results in the rejection of the null hypothesis. The p-value has distribution $U(0,1)$ under the null hypothesis and distribution function $\Phi(\Phi^{-1}(1-p)+\mu_1)$ under the alternative hypothesis, $h_1 : \mu = \mu_1$.

When connected to decision theory, the problem of hypothesis testing is to find an optimal procedure that minimizes some risk function. Suppose that we want to test hypotheses (1.2) that satisfy the assumption (1.3). There are only two possible decisions, rejecting or accepting the null hypothesis, indicated by $d = 1$ and $d = 0$, respectively. We first assign a decision rule $\delta : Y \longmapsto \{0,1\}$, to each possible value of Y. Let $d = \delta(y)$, and $d \in \{0,1\}$.

In order to choose a δ, we must compare the consequences of using different rules. A *loss function*, $L(\theta, d)$, is employed to indicate the consequence of taking decision d when the conditional distribution of Y given $\theta = \theta_i$ is $F_i(y)$. Then the long-term average loss is the expectation $R(\theta, \delta) \equiv \mathrm{E}_\theta [L(\theta, \delta(Y))]$, which is called the *risk function*. A decision rule δ is *inadmissible* if there is another decision rule δ_1 such that $R(\theta, \delta_1) \leq R(\theta, \delta)$ for all θ with strict inequality for some θ. If there is no such δ_1, then we say δ is *admissible* (Lehmann [59]). Here $\mathrm{E}_\theta[\cdot]$ means $\mathrm{E}[\cdot \mid \theta]$. For simplicity, let $E_0[\cdot] = E_{\theta_0}[\cdot]$ when $\Theta_0 = \{\theta_0\}$ and $E_1[\cdot] = E_{\theta_1}[\cdot]$ when $\Theta_1 = \{\theta_1\}$.

1.2.2 Bayesian Hypothesis Testing

Unlike frequentists, Bayesians postulate prior probabilities, π_0 and $\pi_1 = 1 - \pi_0$, respectively, to the event that a null hypothesis is true and to one that it is false. Thus for simple h_0 vs. simple h_1 case, the marginal density of Y is

$$f(y) = \pi_0 f_0(y) + \pi_1 f_1(y). \tag{1.9}$$

The *Bayes risk* is then defined to be

$$r(\delta) = \pi_0 \mathrm{E}_0 [L(\theta_0, \delta(Y))] + \pi_1 \mathrm{E}_1 [L(\theta_1, \delta(Y))]. \tag{1.10}$$

Thus Bayesians are interested in finding an optimal procedure that minimizes the Bayes risk, and the optimal procedure is called the *Bayes rule* of the given decision problem. A simple example is given below. In this example, the Bayes rule is derived for testing $h_0 : \theta = \theta_0$ versus $h_1 : \theta = \theta_1$ a simple alternative hypothesis by using a simple loss function, 0-1 loss.

Example 1.2.2 *Consider a simple loss function, 0-1 loss,*

$$L(\theta, d) = \begin{cases} 1 & \theta = \theta_0,\ d = 1 \\ 0 & \theta = \theta_0,\ d = 0 \\ 1 & \theta = \theta_1,\ d = 0 \\ 0 & \theta = \theta_1,\ d = 1 \end{cases} \tag{1.11}$$

which makes the Bayes risk equal to an overall probability of an error,

$$\pi_0 \mathrm{Pr}(\delta(y) = 1 \,|\, H = 0) + \pi_1 \mathrm{Pr}(\delta(y) = 0 \,|\, H = 1) \tag{1.12}$$

resulting from the use of a decision rule. It can be shown that the Bayes rule minimizing the above probability is

$$\delta(y) = \begin{cases} 0 & f_1(y) < \frac{\pi_0}{\pi_1} f_0(y) \\ 1 & f_1(y) > \frac{\pi_0}{\pi_1} f_0(y) \end{cases}. \tag{1.13}$$

By Bayes' rule, the posterior probability of H, i.e. the conditional probability of H given $Y = y$, is

$$\Pr(H = i | Y = y) = \frac{\pi_i f_i(y)}{\pi_0 f_0(y) + \pi_1 f_1(y)} = \frac{\pi_i f_i(y)}{f(y)}, i = 0, 1 \tag{1.14}$$

and therefore Bayesians would like to accept the null hypothesis if $\Pr(H = 0 \mid Y = y) > \Pr(H = 1 \mid Y = y)$ *(Lehmann [59]).*

Consider next the composite alternative hypotheses situation, $h_0 : \theta = \theta_0$ versus $h : \theta \in \Theta_1 = \Theta - \{\theta_0\}$, where θ is a parameter (possibly a vector) of interest that belongs to the parameter space Θ. Suppose the conditional density of Y under the alternative is $f(y \mid \theta), \theta \in \Theta_1$ and the parameter θ under the alternative has a prior distribution $\pi(\theta)$. In this case, the marginal density of Y under the alternative is $f_1(y) = \int_{\Theta_1} f(y \mid \theta)\pi(\theta)d\theta$. If h_0 is also composite, then we can assume the parameter θ has a prior distribution $\pi(\theta)$ for all $\theta \in \Theta$. One can choose the proper model or proper prior distribution of θ to find the Bayes rule of the testing problem. The model selection usually depends on one's experience and is one of the most important steps in Bayesian testing procedures. The prior distribution of θ can be determined subjectively. One may attempt to use Bayesian methods even when the prior information about the parameter θ is unavailable. A prior that contains no (or minimal) information about θ is called a *noninformative prior* [12].

In Bayesian analysis, Bayes factors are widely used. For the Bayesian model (1.9) given above, the *Bayes factor* is defined to be

$$B = \frac{\int_{\Theta_1} f(y|\theta)\pi(\theta)d\theta}{\int_{\Theta_0} f(y|\theta)\pi(\theta)d\theta} = \frac{\pi_0 f_0(y)}{\pi_1 f_1(y)}. \tag{1.15}$$

Therefore the marginal posterior probability that the null hypothesis is false can be expressed in term of the Bayes factor $\Pr(H = 1 | Y = y) = 1/(1 + B)$.

Frequentist single hypothesis testing can also be considered as a special case of the statistical decision problem. The decision rule is obtained by minimizing the risk function $R(\theta, \delta)$ instead of the Bayes risk. Corresponding to the two types of errors, we can consider two types of loss functions,

$$L_1(\theta, d) = \begin{cases} 1 & \theta \in \Theta_0,\ d = 1 \\ 0 & \theta \in \Theta_1,\ d = 1, \text{ or } \theta \in \Theta,\ d = 0 \end{cases} \tag{1.16}$$

and

$$L_2(\theta, d) = \begin{cases} 1 & \theta \in \Theta_1,\ d = 0 \\ 0 & \theta \in \Theta_0,\ d = 0,\ \text{or } \theta \in \Theta,\ d = 1 \end{cases}. \tag{1.17}$$

Then minimizing $E_\theta\left[L_2(\theta, \delta(Y))\right]$ subject to the restriction $E_\theta\left[L_1(\theta, \delta(Y))\right] \leq \alpha$ is equivalent to maximizing the power while controlling the type I error at the level α (Lehmann [59]).

1.3 Multiple Hypothesis Testing Problem

1.3.1 Problem Statement and Definitions of Compound Error Measures

Multiple testing, in which the number of hypotheses plays an important role, is much more complex than single hypothesis testing. In this case each test has a type I error and a type II error, and it becomes another problem to measure the overall error rate when we have a large number of tests simultaneously. First, an appropriate compound error measure according to the false rejections for multiple testing should be defined. Then, *average power*, which is the proportion of the false hypotheses that are correctly rejected (Benjamini and Hochberg [9]), is commonly employed as a criterion to compare the performance of two multiple testing procedures.

Consider the problem of testing m null hypotheses $h_{1,0}, h_{2,0}, \cdots, h_{m,0}$ versus m alternative hypotheses $h_{1,1}, h_{2,1}, \cdots, h_{m,1}$ simultaneously, of which m_0 is the number of true nulls. For frequentists, m_0 is assumed to be fixed but unknown. Suppose that $\boldsymbol{Y} = (Y_1, Y_2, \cdots, Y_m)$ is the vector of test statistics for m tests, which have joint distribution indexed by the set of parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_m)$, where the θ_i's can be vectors. Let Θ_{i0} and Θ_{i1} be the non-empty subsets of the parameter spaces Θ_i for θ_i under the ith null and alternative hypotheses. Let $\Theta^m = \Theta_1 \times \Theta_2 \times \cdots \times \Theta_m$ be the sample space of $\boldsymbol{\theta}$, and the subset $\Theta_0^m = \Theta_{10} \times \Theta_{20} \times \cdots \times \Theta_{m0}$ be the sample space when all nulls are true. Let $\mathcal{I} = \{1, 2, \cdots, m\}$ be the set of indices of all nulls and $\mathcal{I}_0 = \{i_1, i_2, \cdots, i_{m_0}\}$ denote that of true nulls, and $\mathcal{I}_1 = \mathcal{I} - \mathcal{I}_0$. Thus $\{h_{i,0} : i \in \mathcal{I}_0\}$ and $\{h_{i,1} : i \in \mathcal{I}_1\}$ are the sets of true and false nulls. Let $H_i = 0$ when the ith null hypothesis is true and $H_i = 1$ when it is false. Let $\boldsymbol{H}$ denote the vector $(H_1, H_2, \cdots, H_m)^T$, and therefore $\boldsymbol{H} \in \{0, 1\}^m$. H_i is fixed for frequentists but random for Bayesians. Also let d_i be an indicator of rejecting $h_{i,0}$. Table 1.1 categorizes the m tests into all possible outcomes.

In Table 1.1, R and W denote respectively the total number of rejected hypotheses

	# of **Accepted nulls**	# of **Rejected nulls**	**Total**
# of True Nulls	$A = \sum_{i=1}^{m}(1-H_i)(1-d_i)$	$V = \sum_{i=1}^{m}(1-H_i)d_i$	$m_0 = \sum_{i=1}^{m}(1-H_i)$
# of False Nulls	$T = \sum_{i=1}^{m}H_i(1-d_i)$	$S = \sum_{i=1}^{m}H_i d_i$	$m_1 = \sum_{i=1}^{m}H_i$
Total	$W = \sum_{i=1}^{m}(1-d_i)$	$R = \sum_{i=1}^{m}d_i$	$m = m_0 + m_1$

Table 1.1: All possible outcomes of m hypothesis tests.

and the total number of accepted hypotheses, V and T are respectively the number of false discoveries and the number of false non-discoveries, and A and S are the number of correctly accepted true null hypotheses and the number of correctly rejected false null hypotheses. R is an observed random variable, whereas A, V, S and T are unobserved random variables.

In multiple hypothesis testing, a well-defined compound error rate plays the role of the single hypothesis type I error rate. In terms of the random variables in Table 1.1, we can give the following definitions of error measures according to the number of false rejections V. Use $\mathrm{E}[\cdot]$ to denote expectation taken on the distribution of $\boldsymbol{Y}$.

(a). *Per family error rate*:

$$\mathrm{PFER} \equiv \mathrm{E}[V] \tag{1.18}$$

(Shaffer [87]).

Note that

$$\begin{aligned}
\mathrm{E}[V] &= \mathrm{E}_\theta[\textstyle\sum_{i=1}^{m}(1-H_i)d_i] \\
&= \textstyle\sum_{i=1}^{m}\mathrm{E}_\theta[(1-H_i)d_i] \\
&= \textstyle\sum_{i=1}^{m}\mathrm{Pr}_\theta\left(\{d_i=1\}\cup\{H_i=0\}\right) \\
&= \textstyle\sum_{i\in I_0}\mathrm{Pr}_{\theta_i}(Y_i\in\Gamma^m \mid H_i=0)
\end{aligned} \tag{1.19}$$

where Γ^m is the joint critical region.

(b). *Per comparison error rate*:

$$\mathrm{PCER} \equiv \mathrm{E}\left[\frac{V}{m}\right] = \frac{\mathrm{E}[V]}{m} \tag{1.20}$$

(Shaffer [87]). Miller [63] calls this error rate the *expected family error rate* when $m = m_0$.

(c). *Family-wise error rate*:

$$\text{FWER} \equiv \Pr_\theta(V \geq 1) \tag{1.21}$$

(Shaffer [87]). Miller [63] calls this error rate the *probability of a nonzero family error rate* when $m = m_0$. Assuming $m = m_0$, FWER is also called the experimentwise error rate when it is used in factorial design [31, 51].

It can be shown that FWER $\leq$ PCER [43].

(d). *k-Family-wise error rate*:

$$k\text{-FWER} = \Pr_\theta(V \geq k) \tag{1.22}$$

for fixed $k > 1$ (Lehmann and Romano [60]).

(e). *False discovery rate* (or *expected false discovery rate*):

$$\text{FDR} \equiv \text{E}\left[\frac{V}{R \vee 1}\right] = \text{E}_\theta\left[\frac{V}{R} \mid R > 0\right] \Pr_\theta(R > 0), \tag{1.23}$$

where $R \vee 1 = \max(R, 1)$ (Storey [90]). FDR is loosely defined by Benjamini and Hochberg [9] to be $E\left[\frac{V}{R}\right]$ and equal to zero when there is no rejection. In fact, we are not interested in the case that there is no rejection and thus the definition above is more precise.

(f). *False discovery proportion* (or *realized false discovery rate*):

$$\text{FDP} \equiv \frac{V}{R} \tag{1.24}$$

(Lehmann and Romano [60]).

(g). *Positive false discovery rate*:

$$\text{pFDR} \equiv \text{E}\left[\frac{V}{R} \mid R > 0\right] \tag{1.25}$$

(Storey [90]).

(h). *Proportion of false discoveries (positives)*:

$$\text{PFP} \equiv \text{E}[V] \,/\, \text{E}[R] \tag{1.26}$$

(Bayarri and Berger [8]). Benjamini and Hochberg [9] briefly discussed pFDR and PFP as well.

When the problem of multiple hypothesis testing is viewed as a decision problem, in order to derive a risk function, a well-defined compound error measure analogous to the single hypothesis type II error rate is also needed. In this case, definitions of error measure according to the random number of false non-discoveries T, are given below.

(i). *False non-discovery rate* (or *expected false non-discovery rate*):

$$\mathrm{FNR} \equiv \mathrm{E}\left[\frac{T}{W \vee 1}\right] = \mathrm{E}_{\boldsymbol{\theta}}\left[\frac{T}{W} \,|\, W > 0\right] \Pr_{\boldsymbol{\theta}}(W > 0), \tag{1.27}$$

(Storey [90], Genovese and Wasserman [38]). Genovese and Wasserman [38] referred to FNR as the "dual error rate" to FDR.

(j). *False non-discovery proportion* (or *realized false non-discovery rate*):

$$\mathrm{FNP} \equiv \frac{T}{W} \tag{1.28}$$

(Genovese and Wasserman [39]). FNP is referred to as the "dual error rate" to FDP.

(k). *Positive false non-discovery rate*

$$\mathrm{pFNR} \equiv \mathrm{E}\left[\frac{T}{W} \,|\, W > 0\right] \tag{1.29}$$

(Storey [90]). pFNR is referred to as the "dual error rate" to pFDR.

Given a compound error measure according to the number of false rejections V and a significance level α, the traditional frequentist goal is to determine a multiple hypothesis procedure (*i.e.* a set of test statistics and a set of rejection regions) that maximizes the average power ($\frac{S}{m_1}$, using the notation of Table 1.1) subject to controlling the error rate at α. For example, given a significance level $\alpha > 0$, a procedure controlling FWER is one that yields a FWER less than or equal to α. Usually, one desires that a method controls a certain error rate for all possible combinations of true and false hypotheses ($\boldsymbol{H} \in \{0,1\}^m$). Such control is usually called *strong control*. Procedures that control a certain error rate only when all the null hypotheses are true are said to exhibit *weak control* (Hochberg and Tamhane [51]). For example, given $\alpha > 0$, if there is a procedure that yields $\Pr(V \geq 1 \mid \boldsymbol{H} = \{0\}^m) \leq \alpha$, then we say the FWER is weakly controlled by that procedure; if another procedure guarantees $\max \Pr(V \geq 1 \mid \boldsymbol{H} \in \{0,1\}^m) \leq \alpha$, then this procedure is said to strongly control the FWER. Since weak control is not applicable

for most real problems, the term "control" in this thesis will refer to strong control, unless otherwise noted.

On the other hand, from a Bayesian viewpoint, the number of null and alternative hypotheses are random and there exists prior distributions for the H_i's. In the microarray context, we usually have strong prior information about the probability that a null hypothesis is true. This information is that this probability is usually large [88], though the range of this probability may depend on different microarrays. A Bayesian firstly chooses an appropriate model and loss function to derive the Bayes risk and the Bayes rule for the risk function, or equivalently, finds the rejection rule in terms of the posterior probability of a null hypothesis being false given the observations.

I give a literature review of multiple hypothesis testing methods in the following sections, but the methods introduced are not all related to my work in the thesis. These methods can be mainly sorted into frequentist ones and Bayesian ones. However, in the multiple testing situation there exist procedures that combine both of them. The traditional multiple comparison procedures for comparing normally distributed means are reviewed in section 1.3.2. Those methods are famous and widely used in the analysis of variance. In section 1.3.3, multiple hypothesis testing procedures based on marginal p-values are introduced. They are easily applied and are distribution-free. A comparison of some procedures introduced in this section is given in Chapter 2. In section 1.3.4, some widely used Bayesian models and multiple hypothesis testing methods under these models are given. A Bayesian model, which is developed for regression diagnostics and is introduced in Chapter 3, is motivated by these models.

1.3.2 Multiple Comparison Methods

The traditional frequentist multiple comparison procedures are designed for comparing homogeneity of the means in the analysis of variance, and hence they are mostly based on the joint distributions of all normally distributed observations, such as Tukey's studentized range test, Scheffé's F projections and Duncan's multiple range test [51, 63, 87]. Both Tukey's and Scheffé's tests are *single-stage* procedures whose rejection thresholds do not depend on the data and which strongly control the FWER [31, 51]. The former is more powerful than the latter for pair-wise comparison, and has "generally slightly smaller power for overall tests" than the latter [31]. Both procedures can be modified to *multi-stage*

procedures, in which the rejection threshold is data-dependent, to gain more power [31, 51, 87]. The best known multi-stage multiple testing procedure is the one of Duncan. Duncan's method does not control the FWER. Einot and Gabriel [31] modified Duncan's method to ensure the control of the FWER. They also modified an approach of Ryan [79, 80], and showed that their modified Ryan's procedure is more powerful than the modified Duncan's method, when both methods control the FWER at the same level. More extensive reviews of the traditional frequentist methods for comparing normally distributed means than those presented here can be found in [51, 63, 87].

1.3.3 Multiple Hypothesis Testing Methods Based on Sequential p-values

There are other frequentist methods only based on the empirical distribution of marginal p-values (defined in equation 1.4), and hence they are distribution free. Let $P_1, P_2, \cdots, P_m$ be the p-values corresponding to testing the null hypotheses $h_{1,0}, h_{2,0}, \cdots, h_{m,0}$. Let $P_{(1)} \leq P_{(2)} \cdots \leq P_{(m)}$ be the ordered p-values, and let the null hypothesis corresponding to $P_{(i)}$ be denoted by $h_{(i),0}$. Sequential p-value based methods reject all hypotheses whose p-values are less than a threshold t, where t could be a constant or could be a function of the p-values.

The simplest method, referred to as *uncorrected testing*, rejects $h_{i,0}$ if $P_i \leq \alpha$ to guarantee that the per comparison error rate (PCER) $\leq \alpha$ (Genovese and Wasserman [38]). Obviously, this method ignores the multiplicity m. By the definition of PFER in 1.18, $\mathrm{E}[V] = m_0\alpha$, and hence we expect a large number of false rejections for a large number of hypotheses.

The most commonly controlled error rate when testing multiple hypotheses is the family-wise error rate (FWER), which is the probability of committing at least one Type I error out of all the hypotheses tested (as defined in (1.21)). A multiple testing method controlling FWER, which was proposed earlier than the term "multiple comparison" was introduced, is *Fisher's inverse* χ^2 method [35, p. 99]. This method is based on the fact that $-2\ln P_i$ has a χ^2 distribution with 2 degrees of freedom when $h_{i,0}$ is true. Thus when all p-values are independent, $-2\sum_{i=1}^{m} \ln P_i$ has a χ^2 distribution with $2m$ degrees of freedom when all $h_{i,0}$'s are true. Fisher's inverse χ^2 method can be used to test the null hypothesis $H_1 = H_2 = \cdots = H_m$, so it weakly controls the FWER. However, as mentioned in 1.3.1, weak control is not desirable for applications.

The most famous procedure controlling FWER in the strong sense is the *Bonferroni procedure* (Miller [63]). The Bonferroni procedure rejects h_{i0} if $P_i \leq \frac{\alpha}{m}$ guaranteeing that FWER$\leq \alpha$. The name "Bonferroni" is used in the famous textbook of Miller (1966) (the first edition of [63]). The procedure is given the name of an Italian mathematician, Carlo Emilio Bonferroni, because this method is based on his *First-Order Bonferroni inequality*, also known as Boole's inequality, which states that: given any set of events $\boldsymbol{A}= \{A_1, A_2, \cdots, A_m\}$,

$$\Pr(\bigcup_i A_i) \leq \sum_i \Pr(A_i). \tag{1.30}$$

Let $A_i = \{d_i = 1\}$ denote rejection of $h_{i,0}$, and let $\mathcal{I}_0 = \{i_1, i_2, \cdots, i_{m_0}\}$ denote the set of indices of true nulls. Then

$$\Pr(V \geq 1) = \Pr(\bigcup_{i \in I_0} \{d_i = 1\}) \tag{1.31}$$

$$\leq \sum_{i \in \mathcal{I}_0} \Pr(d_i = 1) \tag{1.32}$$

$$= \sum_{i \in \mathcal{I}_0} \Pr(P_i \leq \frac{\alpha}{m}) \tag{1.33}$$

$$= m_0 \frac{\alpha}{m} \leq \alpha, \tag{1.34}$$

since $\Pr(P_i \leq \frac{\alpha}{m}) = \frac{\alpha}{m}$ for $i \in \mathcal{I}_0$ [63, 87]. Obviously the Bonferroni procedure is usually conservative. If more information of the joint probabilities is given, there are higher order Bonferroni inequalities giving upper and lower bounds on the probability that k, $1 < k \leq m$, or more of the m events in $\boldsymbol{A}$ occur simultaneously (see Feller [32, p. 110]). In fact, the Bonferroni procedure also controls the per family error rate because

$$\mathrm{E}[V] = \sum_{i=1}^{m} \mathrm{E}[(1 - H_i)d_i] = \sum_{i \in \mathcal{I}_0} \Pr(d_i = 1 \mid H_i = 0) = m_0 \cdot \frac{\alpha}{m} \leq \alpha. \tag{1.35}$$

Holm [52] modified the Bonferroni procedure to a step-down procedure that also controls the FWER with increased average power. The *Holm procedure*, based on the ordered p-values, starts with the most significant p-value and continues rejecting hypotheses whose p-values are less than a stage-dependent threshold. First, if $P_{(1)} \geq \frac{\alpha}{m}$, accept $h_{(1),0}, \cdots, h_{(m),0}$ and stop. Otherwise, reject $h_{(1),0}$ and test the remaining $m-1$ hypotheses at level $\frac{\alpha}{m-1}$. At step i, if $P_{(i)} \geq \frac{\alpha}{m+1-i}$, accept $h_{(i),0}, \cdots, h_{(m),0}$ and stop. Otherwise, reject $h_{(i),0}$ as well and test the remaining $m-i$ hypotheses at level $\frac{\alpha}{m-i}$. In other words, reject $h_{(i),0}$ when

$$P_{(t)} \leq \frac{\alpha}{m+1-t} \tag{1.36}$$

for all $1 \leq i \leq t$. One advantage that should be mentioned here is that both the Bonferroni procedure and the Holm procedure make no assumptions concerning the dependence structure of the p-values of the individual tests.

Šidák [85] improved the significance level for each test of the Bonferroni procedure to $1-(1-\alpha)^{1/m}$ when p-values are independent, although the degree of improvement is slight for small values of α. Similarly, the Holm procedure can also be improved by replacing $\alpha/(m+1-t)$ in (1.36) with $1-(1-\alpha)^{1/(m+1-t)}$.

There are some stepwise multiple testing procedures that are based on the *Simes equality* and are proved to be more powerful than multi-step Bonferroni-type procedures. Simes [86] proved that if all null hypotheses are true and the associated test statistics are independent, then

$$\Pr(\textstyle\bigcap_{i=1}^{m}\{P_{(i)} > i\alpha\,/\,m\}) = 1-\alpha \tag{1.37}$$

for any $\alpha \in (0,1)$. Hence the Simes procedure that rejects any $h_{(i),0}$ when $P_{(i)} \leq i\alpha\,/\,m$ controls the PCER and only weakly controls the FWER. It has been proved that $\Pr(\bigcap_{i=1}^{m}\{P_{(i)} > i\alpha\,/\,m\}) > 1-\alpha$ for many types of dependence structure of p-values [87]. Hochberg [50] gave a step-up procedure utilizing the Simes result. The *Hochberg procedure* finds

$$t = \max\left\{1 \leq i \leq m : P_{(i)} \leq \alpha\,/\,(m-i+1)\right\}, \tag{1.38}$$

and rejects all $h_{(i),0}$'s with $i \leq t$. It can also be described as following: If $P_{(m)} \leq \alpha$, reject all $h_{(i),0}$'s; otherwise, $P_{(m)}$ cannot be rejected, and if $P_{(m-1)} \leq \alpha\,/\,2$, reject $h_{(1),0}, \cdots, h_{(m-1),0}$, etc. This procedure strongly controls the FWER and is more powerful than the Holm procedure. Other methods controlling FWER are presented in Hochberg and Tamhane [51], Miller [63] and Shaffer [87].

Although the multi-stage methods can improve the average power, the resulting number of rejections are still quite small, especially when m is very large. Benjamini and Hochberg [9] argued that the FWER controlling methods provide a demanding control for large m. Given $\alpha > 0$, to ensure that $P(V \geq 1) \leq \alpha$, those methods must test each true null hypothesis at a very small level assuming that m_0 is large, and hence result in a small average power. In DNA microarray experiments, we usually test thousands of genes at the same time, where the number of genes having the same expression levels in treatment as in control is often large [29, 30, 88, 90, 92, 93, 98]. In such a situation, one may prefer to pay more attention to comparing the number of false discoveries with the total number of

discoveries rather than paying attention to whether there is one or more type I errors. To address this, Benjamini and Hochberg [9] introduced the concept of the false discovery rate (FDR), the expected proportion of false rejections out of the total rejections (as defined in (1.23)). They also proposed a multi-stage procedure (here referred to as the BH) that strongly controls FDR. The procedure calculates the data-dependent threshold

$$t = \max\left\{1 \leq i \leq m : P_{(i)} \leq \frac{i}{m}\alpha\right\} \tag{1.39}$$

first and then rejects all the $h_{(i),0}$'s with $i \leq t$. Benjamini and Hochberg proved that if the p-values corresponding to the true null hypotheses are uniformly distributed and independent of each other and of the p-values corresponding to the alternative hypotheses, then BH guarantees that FDR $\leq \pi_0\alpha$, where $\pi_0 = m_0/m$ is the proportion of true null hypotheses in the collection. They also use a simulated example to provide evidence that the average power of this method is uniformly larger than the FWER controlling procedures of Bonferroni and of Holm in the different configurations of the hypotheses they considered.

This method is based on the assumption of independent p-values. However, when multiple testing is applied to DNA microarrays, the independence assumption does not usually apply. It is well known that in a microarray experiment most genes are relevant to each other, so the independence assumption is not appropriate [92]. Therefore, an extension to the dependent case is needed. Benjamini and Yekutieli [11] extended the BH procedure to the case that the test statistics have positive regression dependence on each of the test statistics corresponding to the true null hypotheses.

Finner and Roters [34] proved under the same assumptions in Benjamini and Hochberg [9] that indeed FDR $= \pi_0\alpha$. (For related results, see also Benjamini and Yekutieli [11]; Genovese and Wasserman [38]; and Sarkar [81].) Therefore, the BH procedure is conservative when $m_0 < m$. This suggests that the BH procedure can be more efficient by incorporating an estimate of π_0. Benjamini and Hochberg [10] proposed a data-adaptive procedure to estimate this proportion and modify the BH procedure, but they didn't show that this procedure provides strong control of the FDR.

Storey [89] proposed an alternative procedure incorporating an estimate of π_0 to overcome the conservativity of the BH method. He considered the FDR arising from multiple testing with a fixed significance level and proposed an estimator of the FDR which he showed to have positive bias. Choosing this significance level to control the estimated

FDR will also control the FDR. This fixed significance level method (hereafter called FSL) is based on a Bayesian model which is introduced in section 1.3.4, though it is mainly a frequentist method. Storey [89] defined the FDR as in (1.23), where R is the number of rejections and V is the number of false discoveries. Whereas the rejection threshold t calculated by BH is random, Storey suggested to reject hypotheses $h_{i,0}$, for which $P_i \leq \gamma$ for a fixed pre-selected rejection threshold γ and then estimating the FDR. Since $\mathrm{E}(V(\gamma)) = \pi_0\gamma$ and $V(\gamma) = \#\{P_i \leq \gamma \mid H_i = 0\}$, he estimated the FDR by

$$\widehat{\mathrm{FDR}}_\lambda(\gamma) = \frac{\hat{\pi}_0\gamma}{R(\gamma) \vee 1} \tag{1.40}$$

where

$$\hat{\pi}_0(\lambda) = \frac{1 - R(\lambda)}{1 - \lambda},\ R(\lambda) = \#\{P_i > \lambda\} \tag{1.41}$$

for some suitably chosen λ. Since FDR = pFDR $\times$ Pr$\{R(\gamma) > 0\}$, Storey also gave a slightly different conservative estimate of pFDR

$$\widehat{\mathrm{pFDR}}_\lambda(\gamma) = \frac{\hat{\pi}_0\gamma}{\{R(\gamma) \vee 1\}\{1 - (1 - \gamma)^m\}}, \tag{1.42}$$

where $1 - (1 - \gamma)^m$ is a lower bound for Pr$\{R(\gamma) > 0\}$. Both estimators are based on the empirical distribution of marginal p-values. The conservativity property of the estimators are proved under the assumption of independent p-values in [89], and Storey and Tibshirani [92] modified the estimators and proved that the conservativity is valid under several dependent cases for the modified estimators. Recently, Sarkar and Liu [84] proposed a modification of $\widehat{\mathrm{FDR}}$, replacing the denominator in (1.40) by $\left(R(\gamma) \vee \frac{1}{m} + 1\right)/(m + 1)$. This estimator has better small-sample properties. However, their estimator of FDR is close to (1.40) when m is large.

Storey [89] also proposed a bootstrap approach to estimate the optimal λ which minimizes the mean square error of the estimate. However, Black [18] argued that the bootstrap method may underestimate π_0 and thus results in an underestimate of FDR. My simulation results in Chapter 2 confirm his argument and also show that the bootstrap method may produce a larger bias than using a fixed λ. Storey and Tibshirani [93] used a spline method to estimate π_0. My simulation results show that this method needs less computing time and produces a smaller bias compared to the bootstrap one. The simulation estimates of π_0 by using a fixed $\lambda = 0.9$ and by using the spline method are close. The other methods estimating π_0 include Bickis *et al.* [17], Bickis [15], Efron *et al.* [30], Ferreira and Zwinderman [33], and Tusher *et al.* [98].

Storey [89] showed numerically that FSL can gain more power than BH while controlling the same FDR. In Chapter 2 it is proved that when the number of alternatives (m_1) is small compared to the total number of hypotheses, FSL can be less powerful than BH. Black [18] proposed an adaptive FDR controlling procedure (hereafter called AFDR) that adjusts BH for this conservatism using Storey's estimator of π_0. Given a fixed γ, the AFDR implements BH with the target FDR set to the data-dependent value

$$\hat{\alpha} = \widehat{\mathrm{FDR}}_\lambda(\gamma)/\widehat{\pi}_0(\lambda). \tag{1.43}$$

His simulation results showed that the AFDR has power comparable to the fixed rejection region method of Storey [89]. It is shown in Chapter 2 that AFDR has superior power to FSL.

Storey *et al.* [94] investigated picking a data dependent significance level to control the estimated FDR of Storey [89]. For a given target α, Storey *et al.* [94] proposed, instead of FSL, choosing the random significance level Γ by

$$\Gamma = \sup\left\{0 \le p \le 1 \ : \ \widehat{\mathrm{FDR}}_\lambda(p) \le \alpha\right\}, \tag{1.44}$$

which they showed to be equivalent to BH with target level α in the case that $\hat{\pi}_0(\lambda)$ is fixed at 1 rather than being an estimator. It is shown in Chapter 2 that this procedure is equivalent to a procedure that is connected to AFDR. In Chapter 2, a clarification of the relationships of BH, FSL and AFDR is also given.

Because the rejection threshold of the BH method is data-dependent, it is difficult to work out an explicit expression for its power function. However, there is some work on the asymptotic behaviour of the BH method. Genovese and Wasserman [38] gave the asymptotic rejection threshold assuming that the p-values are independent and the distribution function of p-value under alternative is strictly concave. Ferreira and Zwinderman [33] generalized the results of [38] to the case that the independence and concavity assumptions are no longer needed. They also proved that asymptotic FDR of the BH method as $m \to \infty$ still equal to $\pi_0\alpha$, even under a dependent structure. However, my simulation results in chapter 2 show that the empirical power of the BH method is quite different from the asymptotic power even for m as large as 5000. Chi [22] studied the FDR, pFDR, and power of the BH procedure asymptotically under a random effect model which is introduced in chapter 1.3.4. Wu [100] generalized Chi's results to a conditional dependence model.

Since the FWER is criticized to be too strict for a multiple testing problem where one is willing to tolerate a few false discoveries, Lehmann and Romano [60] suggested to control the k-FWER, the probability of committing at least k false rejections out of all the hypotheses tested (as defined in (1.22)). Obviously, when $k > 1$, the control of k-FWER is less conservative than the control of FWER. It is straightforward to extend the existing methods controlling FWER to ones controlling k-FWER. Lehmann and Romano [60] modified the Bonferroni and Holm procedures to obtain both single stage and multi-stage procedures controlling k-FWER in the strong sense. These two methods do not make any assumption concerning the dependence structure of the p-values of the individual tests. Romano and Wolf [77] proposed a more powerful multi-stage procedure by taking account of the dependence structure of the p-values of the individual tests. Sarkar [83] generalized Hochberg's step-up procedure which controls the FWER to strongly control the k-FWER. The generalized Hochberg's procedure is more powerful than generalized the Holm procedure of Lehmann and Romano. It is noted that the choice of k value should be seriously considered.

Similar to generalizing the control of FWER, Gordon *et al.* [43] suggest to assign an upper bound for the per family error rate (PFER), the expected number of false rejections (as defined in (1.18)). The upper bound α can be greater than one when more than one false rejection is allowed. The Bonferroni multiple testing procedure controls not only the FWER but also the PFER. Now if the significance level α is interpreted as the upper bound for PFER, then the power of the Bonferroni procedure can be improved. Their simulation results show the Bonferroni procedure has smaller variances of both the number of true rejections and the total number of rejections than the BH. However, experience and caution is really needed for choosing the upper bound α.

1.3.4 Mixture models for multiple hypothesis testing

In this section, some simple and broadly used mixture models are given, and the multiple hypothesis testing methods based on these model are discussed.

As mentioned before, a natural Bayesian approach for multiple hypothesis testing assumes the common prior probability $\Pr(H_i = 0) = \pi_0$ and then $\Pr(H_i = 1) = 1 - \pi_0 = \pi_1$, *i.e.* H_i's are i.i.d. random variables with distribution Bernoulli(π_0). Let $Y_1, \cdots Y_m$ be the *i.i.d.* test statistics of m tests. Suppose Y_i are generated from a mixture model: the

common distribution of $Y_1, \cdots Y_m$ under the null is $F_0(y)$ with density $f_0(y)$; $Y_1, \cdots Y_m$ under the alternative also have a common distribution $F_1(y)$ with density $f_1(y)$, that is

$$Y_i \mid H_i = 0 \sim F_0,\ Y_i \mid H_i = 1 \sim F_1, \text{ and } H_i \sim \text{Bernoulli } (\pi_0)\ . \tag{1.45}$$

Thus $Y_1, \cdots Y_m$ marginally follow the mixture distribution

$$F(y) = \pi_0 F_0(y) + \pi_1 F_1(y) \tag{1.46}$$

with mixture density

$$f(y) = \pi_0 f_0(y) + \pi_1 f_1(y). \tag{1.47}$$

Under the above model assumptions, Efron *et al.* [30] introduced a "nonparametric empirical Bayesian" approach to estimate the marginal posterior probabilities

$$p_0(y) = \Pr(H_i = 0 \mid Y_i = y) = \frac{\pi_0 f_0(y)}{f(y)} \tag{1.48}$$

$$p_1(y) = \Pr(H_i = 1 \mid Y_i = y) = 1 - \frac{\pi_0 f_0(y)}{f(y)} \tag{1.49}$$

and used an upper bound estimate of π_0. This method only assumes that the parameters arise from some common population with unknown distribution and thus is nonparametric. In order to improve the estimation of f_0, they employed a permutation method to generate the values of Y under the null hypotheses. Thus the ratio $f_0(y)/f(y)$ can be estimated directly from the empirical distributions of Y and the null version of Y. A logistic regression approach was used to estimate this ratio. The proportion of the true null hypotheses is estimated by an upper bound of it, $\min_y \{f(y)/f_0(y)\}$. Both density estimates and the estimate of π_0 are based on the empirical distribution of the test statistics and a permutation method which does not depend on the distribution of the statistics or any dependence structure of the test statistics. A small value of the estimate results in a rejection of the associated null hypothesis. The empirical Bayesian method was applied to an oligonucleotide microarray experiment in [30] as well as a spotted cDNA microarray experiment in Efron and Tibshirani [29]. Efron and Tibshirani also presented the connection between the proposed empirical Bayes approach of Efron *et al.* and the FDR control of Benjamini and Hochberg [9]. For the rejection region of type $\Gamma = \{y_i \mid y_i \leq y\}$, they defined the *Bayesian FDR* as

$$\text{Fdr } (y) \equiv \pi_0 F_0(y) \ / \ F(y) = \Pr(H_i = 0 \mid Y_i \leq y)\ . \tag{1.50}$$

They also proved that $\mathrm{Fdr}(y)$ approximates FDR for the rejection region $\Gamma = \{y_i \mid y_i \leq y\}$. Efron *et al.* defined *local false discovery rate* as

$$\text{fdr } (y) \equiv \pi_0 f_0(y) \ / \ f(y) \ . \tag{1.51}$$

Both Fdr and fdr are based on partial data and known distributions. The Bayesian FDR can be defined for general Γ as well. They also showed that there is a relationship between the local FDR and the Bayesian FDR. The Averaging Theorem in [29] states that "$\mathrm{Fdr}(\Gamma)$ is the conditional f-average of $\mathrm{fdr}(y)$ for $y \in \Gamma$", *i.e.* $\mathrm{Fdr}(\Gamma) = \mathrm{E}_f \{\mathrm{fdr}(y) \mid y \in \Gamma\}$. Moreover, neither Efron *et al.* nor Efron and Tibshirani took into account any loss function to derive the optimal rule. They made a decision based only on the marginal posterior probability p_1.

Scott and Berger [88] proposed a hierarchical Bayesian approach to multiple testing motivated by the need for significance analysis of microarrays. In their model, observations arise independently from normal densities with different means μ_i and a common but unknown variance σ^2 under the null and alternative hypotheses, where $\mu_i = 0$ if $H_i = 0$ but $\mu_i \neq 0$ if $H_i = 0$. The goal is to identify genes which have over or under expression in treatment compared to control, *i.e.* determine which of the μ_i's are nonzero. Then the nonzero μ_i's are modelled as having a common normal distribution with zero mean and an unknown variance V. Since there is no strong prior information about V and σ^2, an noninformative prior is used. As mentioned before, there is strong prior information about π_0, which is that it is usually expected to be large in the microarray context because most genes are expected to have the same expression in treatment and control. Therefore, they chose a Beta distribution with the parameters a and b, and b is fixed to be one. When $a = 1$, the hyper-prior of π_0 is the uniform prior. Scott and Berger suggested to specify α by making a "best guess" $\hat{\pi}_0^*$ for π_0. If $\hat{\pi}_0^*$ is interpreted as the median of the prior distribution, then $a = \log(0.5)/\log(\hat{\pi}_0^*)$. The hierarchical Bayesian model given by Scott and Berger [88] can be written as

$$\begin{aligned} &F_0 \sim N(0, \sigma^2), \ F_1 \sim N(\mu, \sigma^2) \\ &\mu \sim N(0, V) \\ &\pi(V, \sigma^2) = (V + \sigma^2)^{-2}, \ \pi_0 \sim \mathrm{Beta}(a, 1) \end{aligned} \quad . \tag{1.52}$$

The Scott and Berger method is a parametric method involving the calculation of high dimensional integrals, which can only be solved by simulation methods. In this paper,

the importance sampling method was used to calculate the posterior probabilities of the parameters. They discussed the appropriate choices of prior distributions and how the choices affect posterior probabilities. The approach also computes the marginal posterior probability $p_i = \Pr(H_i = 0 \mid D)$, where D denotes the observed dataset, and rejects $h_{i,0}$ with p_i less than or equal to some threshold, which is the optimal decision rule using a loss function proportional to the distance between null and alternative distributions, *i.e.*

$$L(\mu_i, d) = \begin{cases} 1 & \mu_i = 0,\ d = 1 \\ 0 & \mu_i = 0,\ d = 0 \\ c\,|\mu_i| & \mu_i \neq 0,\ d = 0 \\ 0 & \mu_i \neq 0,\ d = 1 \end{cases}, \tag{1.53}$$

where c is a constant. The hyper-prior distribution of π_0 used in [88] has a large mass at 1, which means that with high probability there is no gene showing different expression levels in treatment and control. However, although the distributions of a large proportion of the genes expression levels are unaltered in treatment and control, there are some genes showing change in their expression.

Muller *et al.* [64] introduced a more useful Bayesian version of FDR,

$$\overline{\text{FDR}} = E(FDP \mid D) = E(\frac{\sum(1 - H_i)d_i}{\sum d_i} \mid D) = \frac{\sum(1 - v_i)d_i}{\sum d_i} \tag{1.54}$$

where $v_i = \Pr(H_i = 1 \mid D)$ is the marginal posterior probability of the ith null being a true null. The similar definitions of Bayesian FNR, Bayesian false discoveries, and Bayesian false non-discoveries can also be derived as following

$$\overline{\text{FD}} = \sum(1 - v_i)d_i \tag{1.55}$$

$$\overline{\text{FN}} = \sum v_i(1 - d_i) \tag{1.56}$$

$$\overline{\text{FNR}} = \frac{\overline{\text{FN}}}{m - \sum d_i} = \frac{\sum v_i(1 - d_i)}{m - \sum d_i}. \tag{1.57}$$

They derived the optimal decision rules for several loss functions based on those quantities, such as $L_N(d, y) = c\overline{\text{FD}} + \overline{\text{FN}}$, $L_R(d, y) = c\overline{\text{FDR}} + \overline{\text{FNR}}$.

One may wonder whether the rule defined in Benjamini and Hochberg [9] is optimal under a certain loss function. Cohen and Sackrowitz [23, 24] showed that the step-up procedures including BH are inadmissible under a loss function including the combination of $\overline{\text{FD}}$ and $\overline{\text{FN}}$.

The traditional Bayes methods need to specify proper prior distributions for model parameters under null and alternative hypotheses. Sometimes subjective priors may not be available for some parameters, and the calculation usually involves computing high dimensional integrals. Johnson [54] proposed an approach for calculating Bayes factors from modelling the sample distributions of test statistics directly rather than those of raw data. Usually the null distribution of a test statistic does not depend on any unknown parameters. When the alternative hypothesis is the negation of the null hypothesis, Johnson suggested to model the distribution of the test statistic with a minimum number of parameters under "a reasonable broad class of alternative model" [54]. The examples in [54] shows that the Bayes factor associated with standard test statistics, such as χ^2, F, t and z, have relatively simple expressions, and the parameters that are implicit to the alternative hypothesis can be estimated by the marginal maximum likelihood estimate (MMLE). Ji *et al.* [53] used the results of Johnson [54] and Muller *et al.* [64] to proposed a Bayesian approach for the problem of multiple hypothesis testing arising from the microarray analysis. In their model, both null and alternative distributions of the test statistics are modelled as scaled F distributions with different scale parameters. An inverse gamma and a gamma distribution are used as prior distributions of the scale parameters under alternative and null hypotheses, and π_0 is assumed having the same Beta prior distribution as in Scott and Berger [88]. The joint posterior probability of all parameters is simpler than that computed from the Bayesian model in [88], but it still needs to be solved by simulation methods. In Ji *et al.* [53], a Markov chain Monte Carlo (MCMC) algorithm is employed to estimate the joint posterior probability. Then they calculated the optimal rejection threshold that minimizes $\overline{\text{FNR}}$ subject to fixing $\overline{\text{FDR}}$ at a given level.

In Chapter 3, I propose a Bayesian model for the multiple outlier identification problem that is often encountered in regression analysis. This model is motivated by the mixture models introduced in this section. In the next section, an introduction to the problem of multiple deletion diagnostics is presented.

1.4 Regression diagnostics

1.4.1 Linear regression model and Least Squares

I start with standard results from least squares for the linear regression model. All the results in this section can be found in Atkinson [1], Rao [74] or Ravishanker and Dey [73]. In the linear regression model

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.58}$$

where $\boldsymbol{y}$ is the $m \times 1$ vector of *responses* (also called dependent variable), X is the $m \times k$ full-rank design matrix of $k-1$ known vectors of *explanatory* variables (or independent variable), with all elements of the first column equal to 1 and ith row $\boldsymbol{x}_i^T$; $\boldsymbol{\beta}$ is a vector of k unknown parameters; and $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_m)$ is a vector of m unknown random errors. Note that in Section 1.2 and 1.3, y denotes a realization of a test statistic Y, and X denotes data, but in Section 1.4 and Chapter 3, $\boldsymbol{y} = (y_1, \cdots, y_m)^T$ and X denote respectively the vector of responses and the design matrix of explanatory variables. The only test statistic introduced in this section is called the deletion residual, and is denoted by r_i^* for the ith observation.

If the linear regression model is true, then the following assumption is satisfied:

Assumption 1.4.1 *(1) Model:* $\mathrm{E}[y_i \mid \boldsymbol{x}_i] = \boldsymbol{x}_i^T\boldsymbol{\beta}$.

(2) Independence: the y_i's are conditionally independent, given the $\boldsymbol{x}_i$'s.

(3) Homoscedasticity: $\mathrm{Var}[y_i \mid \boldsymbol{x}_i] = \sigma^2$, *or equivalently* $\mathrm{Var}[\varepsilon_i] = \sigma^2$, *where σ^2 is unknown.*

(4) Normality: given $\boldsymbol{x}_i$, y_i has a normal distribution, or equivalently the ε_i has a normal distribution.

Assumptions 1-4 can be equivalently written as: the vector of the observations $\boldsymbol{y}$ has a multivariate normal distribution (MVN)

$$\boldsymbol{y} \sim \mathrm{MVN}(X\boldsymbol{\beta},\, \sigma^2 I) \tag{1.59}$$

The vector of *least square* estimates $\hat{\boldsymbol{\beta}}$ of the vector of parameters $\boldsymbol{\beta}$ minimizes the sum of squares

$$R(\boldsymbol{\beta}) = (\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta}), \tag{1.60}$$

and thus satisfies the *normal equation*

$$X^T X\hat{\boldsymbol{\beta}} = X^T \boldsymbol{y}. \tag{1.61}$$

Therefore, the vector of least square estimates is

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}. \tag{1.62}$$

With this vector of estimates, the vector of least square residuals is

$$\boldsymbol{r} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - X\hat{\boldsymbol{\beta}} = (I - G)\boldsymbol{y} \tag{1.63}$$

and the minimized sum of squares is

$$R(\hat{\boldsymbol{\beta}}) = \boldsymbol{r}^T \boldsymbol{r} = \boldsymbol{y}^T (I - G)\boldsymbol{y} \tag{1.64}$$

where $G = X(X^T X)^{-1} X^T$ is the so-called "*hat*" matrix. The matrix G has diagonal elements $g_i = \boldsymbol{x}_i^T (X^T X)^{-1} \boldsymbol{x}_i$ and off-diagonal elements g_{ij}. The ith diagonal element of G, g_{ii}, is abbreviated to g_i. In the literature, the hat matrix and its diagonal and off-diagonal elements are usually respectively denoted by H, h_i and h_{ij}, but in this thesis H is used as the indicator for the alternative and h_{ij}, $j = 0, 1$ denotes the ith null and alternative hypotheses, respectively, among m tests in this thesis. Therefore in this thesis, G, g_i and g_{ij} are used instead. The hat matrix has the properties shown in the following result. The results in the remaining part of this subsection can be found in Rao [74] or Ravishanker and Dey [73].

Result 1.4.1 *(1) G is symmetric and idempotent, that is, $G^T = G$ and $G^2 = G$.*

(2) $\sum_{i=1}^{m} g_i = \text{trace}(G) = \text{trace}\left\{X(X^T X)^{-1} X^T\right\} = \text{trace}\left\{(X^T X)^{-1} X^T X\right\} = \text{trace}(I_k) = k = \text{rank}(G)$.

It is easy to show that

$$\text{E}\left[R(\hat{\boldsymbol{\beta}})\right] = \text{E}\left[\boldsymbol{y}^T (I - G)\boldsymbol{y}\right] = \text{trace}(I - G)\text{Var}(y) = \sigma^2 (m - k). \tag{1.65}$$

Hence an unbiased estimator of σ^2 is

$$s^2 = R(\hat{\boldsymbol{\beta}}) \;/\; (m - k). \tag{1.66}$$

By (1.59), it is easy to show that $\hat{\boldsymbol{\beta}}$ is an unbiased vector of estimates for $\boldsymbol{\beta}$, and also has a multivariate normal distribution. Moreover, $\boldsymbol{\beta}$ and s^2 are independent. Hence, we have the following result.

Result 1.4.2 *Since*

$$\begin{bmatrix}\hat{\boldsymbol{\beta}}\\ \boldsymbol{r}\end{bmatrix} = \begin{bmatrix}(X^TX)^{-1}X^T\\ (I-G)\end{bmatrix}\boldsymbol{y}, \tag{1.67}$$

then under the Assumption 1.4.1, we have the following results:

(1) The vector of estimates $\hat{\boldsymbol{\beta}}$ follows a multivariate normal distribution, i.e.,

$$\hat{\boldsymbol{\beta}} \sim \text{MVN}(\boldsymbol{\beta},\sigma^2(X^TX)^{-1}). \tag{1.68}$$

(2) The vector of residuals $\boldsymbol{r}$ has a multivariate normal distribution, i.e.,

$$\boldsymbol{r} \sim \text{MVN}(\boldsymbol{0}, \sigma^2(I-G)). \tag{1.69}$$

(3) $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{r}$ are jointly MVN and $\text{Cov}(\hat{\boldsymbol{\beta}}, \boldsymbol{r}) = 0$, so $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{r}$ are independent. Since $s^2 = \frac{1}{m-k}\boldsymbol{r}^T\boldsymbol{r}$, then $\hat{\boldsymbol{\beta}}$ and $s^2$are also independent.

Note that the residuals r_i's defined in (1.63) do not all have the same variance, and particularly, $Var(r_i) = \sigma^2(1-g_i)$. The standardized residual r_i' is then defined as

$$r_i' = \frac{r_i}{s\sqrt{1-g_i}}. \tag{1.70}$$

As shown in the First Fundamental Theorem in [74, p. 189], the distribution of $R(\hat{\boldsymbol{\beta}})$ is a central chi-square distribution with $m-k$.

Theorem 1.4.1 (Rao [74]) *Under Assumption 1.4.1, $(m-k)s^2/\sigma^2$ has a chi-square distribution with $m-k$ degrees of freedom.*

In order to prove this theorem, we need a lemma which is given on page 186, part (ii) in [74]. I also use this lemma to prove the new theorem I propose in Chapter 3.

Lemma 1.4.1 (Rao [74]) *Let $y_1, \cdots, y_n$ be i.i.d. random variables from $N(\mu_i, 1)$, and let $\boldsymbol{y}^T = (y_1, \cdots, y_n)$ and $\boldsymbol{\mu}^T = (\mu_1, \cdots, \mu_n)$. A necessary and sufficient condition that $\boldsymbol{y}^TA\boldsymbol{y}$ has a chi-square distribution is that A is idempotent, in which case, the degrees of freedom of χ^2 is $\text{rank}(A) = \text{trace}(A)$, and if μ_i's are not all zero, then χ^2 is non-central with centrality parameter $\boldsymbol{\mu}^TA\boldsymbol{\mu}$.*

It is easy to prove Theorem 1.4.1 by using this lemma.

Proof. [of Theorem 1.4.1]
Since y_i/σ^2 are i.i.d. $N(\boldsymbol{x}_i^T\boldsymbol{\beta},1)$ and $I-G$ is idempotent, $\boldsymbol{y}^T(I-G)\boldsymbol{y}\ /\ \sigma^2$ has a chi-square distribution with the degrees of freedom trace$(I-G)=m-k$ and non-central parameter $\frac{1}{\sigma^2}(X\boldsymbol{\beta})^T(I-G)X\boldsymbol{\beta}=\frac{\mathbf{1}}{\boldsymbol{\sigma^2}}\boldsymbol{\beta}^TX^T\left\{I-X(X^TX)^{-1}X^T\right\}X\boldsymbol{\beta}=0$. ∎

The theorem 1.4.1 is true when Assumption 1.4.1 is valid. In Chapter 3 I derive the distribution of s^2 when Assumption 1.4.1 is violated.

1.4.2 Identification of one outlier

As mention in Section 1.1, a powerful method of identifying a single outlier is introduced in Cook and Weisberg [25] and Atkinson [1]. This method is based on the deletion of single observation [6]. The test statistic calculated from the deletion of a single observation is the *deletion residual* (also called *studentized residual* in some literature). The deletion residual for the ith observation is the standardized residual for the ith observation using a model fitted with other observations

In order to identify outliers which are numerically distant from the rest of the data, we need to compare the regression models with and without some observations. The algebra of deletion discussed below is taken from Atkinson [1]. Atkinson [1] gave results for deleting a set of observations. I only list results for deleting a single observation, which is a special case of Atkinson's results, and I also present calculations of some results. Let $X_{(i)}$ denote the design matrix with the ith row $\boldsymbol{x}_i^T$ deleted from X, and $\boldsymbol{y}_{(i)}$ denote the vector of responses with the ith observation y_i deleted. Similarly, a notation with a subscript i in parentheses denotes the quantity obtained with the ith observation y_i deleted. The vector of least square estimates $\hat{\boldsymbol{\beta}}_{(i)}$ with the ith observation being deleted is therefore

$$\hat{\boldsymbol{\beta}}_{(i)}=(X_{(i)}^TX_{(i)})^{-1}X_{(i)}^T\boldsymbol{y}_{(i)}, \tag{1.71}$$

and the corresponding residuals are

$$\boldsymbol{r}_{(i)}=\boldsymbol{y}_{(i)}-\hat{\boldsymbol{y}}_{(i)}=\boldsymbol{y}_{(i)}-\boldsymbol{x}_{(i)}\hat{\boldsymbol{\beta}}_{(i)}=(I-G_{(i)})\boldsymbol{y}_{(i)}, \tag{1.72}$$

where

$$G_{(i)}=X_{(i)}(X_{(i)}^TX_{(i)})^{-1}X_{(i)}^T. \tag{1.73}$$

The estimate of σ^2 is then

$$s_{(i)}^2=R(\hat{\boldsymbol{\beta}}_{(i)})\ /\ (m-k-1)=\boldsymbol{r}_{(i)}^T\boldsymbol{r}_{(i)}\ /\ (m-k-1)=\boldsymbol{y}_{(i)}^T(I_{m-1}-G_{(i)})\boldsymbol{y}_{(i)}, \tag{1.74}$$

where I_{m-1} denotes the $(m-1)\times(m-1)$ identity matrix. The results for $\hat{\boldsymbol{\beta}}$, $\boldsymbol{r}$, and s in Section 1.4.1 are still true for $\hat{\boldsymbol{\beta}}_{(i)}$, $\boldsymbol{r}_{(i)}$ and $s_{(i)}$ provided the model assumptions are valid for $\boldsymbol{y}_{(i)}$.

In fact, $(X_{(i)}^T X_{(i)})^{-1}$ can be calculated from $(X^T X)^{-1}$. To see this, first note that

$$\begin{aligned} X^T X &= [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m][\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_m^T]^T \\ &= \boldsymbol{x}_1\boldsymbol{x}_1^T + \cdots + \boldsymbol{x}_i\boldsymbol{x}_i^T + \boldsymbol{x}_m\boldsymbol{x}_m^T \\ &= X_{(i)}^T X_{(i)} + \boldsymbol{x}_i\boldsymbol{x}_i^T. \end{aligned} \tag{1.75}$$

Let A be a square matrix and let B and C^T be matrices with the same dimension and with the number of rows of B equal to that of A. It is easy to verify that

$$(A - BC)^{-1} = A^{-1} + A^{-1}B(I - CA^{-1}B)^{-1}CA^{-1}. \tag{1.76}$$

Let $A = X^T X$, and $B = C^T = \boldsymbol{x}_i$, then we have

$$(X_{(i)}^T X_{(i)})^{-1} = (X^T X - \boldsymbol{x}_i\boldsymbol{x}_i^T)^{-1} \tag{1.77}$$

$$= (X^T X)^{-1} + (X^T X)^{-1}\boldsymbol{x}_i(I - \boldsymbol{x}_i(X^T X)^{-1}\boldsymbol{x}_i^T)^{-1}\boldsymbol{x}_i^T(X^T X)^{-1} \tag{1.78}$$

$$= (X^T X)^{-1} + (X^T X)^{-1}\boldsymbol{x}_i(1 - g_i)^{-1}\boldsymbol{x}_i^T(X^T X)^{-1} \tag{1.79}$$

$$= (X^T X)^{-1}\left\{I + \frac{1}{(1-g_i)}\boldsymbol{x}_i\boldsymbol{x}_i^T(X^T X)^{-1}\right\}. \tag{1.80}$$

Hence, the jkth element of $G_{(i)}$ has the form

$$g_{(i)jk} = \mathbf{x}_j^T(X_{(i)}^T X_{(i)})^{-1}\mathbf{x}_k \tag{1.81}$$

$$= \mathbf{x}_j^T(X^T X)^{-1}\mathbf{x}_k + \frac{\mathbf{x}_j^T(X^T X)^{-1}\mathbf{x}_i\mathbf{x}_i^T(X^T X)^{-1}\mathbf{x}_k}{1 - g_i} \tag{1.82}$$

$$= g_{jk} + \frac{g_{ij}g_{ik}}{1-g_i},\ j,k = 1,\ldots,i-1,i+1,\ldots m,\ j \neq k, \tag{1.83}$$

and the jth diagonal element of $G_{(i)}$ can be written as

$$g_{(i)jj} \triangleq g_{(i)j} = g_j + \frac{g_j^2}{1-g_i},\ j = 1,\ldots,i-1,i+1,\ldots m, \tag{1.84}$$

where g_j is the jth diagonal element of G and g_{jk} is the jkth off-diagonal element of G.

Also $\hat{\boldsymbol{\beta}}_{(i)}$ and $s_{(i)}$ respectively have the following relationships with $\hat{\boldsymbol{\beta}}$ and s.

Result 1.4.3

$$\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}} = -(X^T X)^{-1}\boldsymbol{x}_i r_i \ /\ (1 - g_i), \tag{1.85}$$

and

$$(m-k-1)s^2_{(i)} = (m-k)s^2 - r_i^2 \ / \ (1-g_i). \tag{1.86}$$

Result 1.4.3 can be shown from the following calculations:

$$\begin{aligned}
&\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}} \\
&= (X^T_{(i)}X_{(i)})^{-1}X^T_{(i)}\boldsymbol{y}_{(i)} - (X^TX)^{-1}X^T\boldsymbol{y} && (1.87)\\
&= (X^T_{(i)}X_{(i)})^{-1}(X^T\boldsymbol{y} - \boldsymbol{x}_i y_i) - (X^TX)^{-1}X^T\boldsymbol{y} && (1.88)\\
&= (X^TX)^{-1}(X^T\boldsymbol{y} - \boldsymbol{x}_i y_i) + \frac{1}{(1-g_i)}(X^TX)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^T(X^TX)^{-1}(X^T\boldsymbol{y} - \boldsymbol{x}_i y_i) && (1.89)\\
&- (X^TX)^{-1}X^T\boldsymbol{y} && (1.90)\\
&= -(X^TX)^{-1}\boldsymbol{x}_i y_i + \frac{1}{(1-g_i)}(X^TX)^{-1}\boldsymbol{x}_i\left\{\boldsymbol{x}_i^T(X^TX)^{-1}X^T\boldsymbol{y} - \boldsymbol{x}_i^T(X^TX)^{-1}\boldsymbol{x}_i y_i\right\} && (1.91)\\
&= -(X^TX)^{-1}\boldsymbol{x}_i y_i + \frac{1}{(1-g_i)}(X^TX)^{-1}\boldsymbol{x}_i\left\{\hat{y}_i - g_i y_i\right\} && (1.92)\\
&= (X^TX)^{-1}\boldsymbol{x}_i\left\{-y_i + \frac{\hat{y}_i - g_i y_i}{(1-g_i)}\right\} && (1.93)\\
&= (X^TX)^{-1}\boldsymbol{x}_i\frac{\hat{y}_i - y_i}{(1-g_i)} && (1.94)\\
&= -(X^TX)^{-1}\boldsymbol{x}_i r_i \ / \ (1-g_i), && (1.95)
\end{aligned}$$

and then

$$\begin{aligned}
&(m-k-1)s^2_{(i)} \\
&= \left(\boldsymbol{y}_{(i)} - X_{(i)}\hat{\boldsymbol{\beta}}_{(i)}\right)^T\left(\boldsymbol{y}_{(i)} - X_{(i)}\hat{\boldsymbol{\beta}}_{(i)}\right) && (1.96)\\
&= \boldsymbol{y}^T_{(i)}\boldsymbol{y}_{(i)} - \hat{\boldsymbol{\beta}}^T_{(i)}X^T_{(i)}\boldsymbol{y}_{(i)} && (1.97)\\
&= \boldsymbol{y}^T\boldsymbol{y} - y_i^2 - \hat{\boldsymbol{\beta}}^T_{(i)}\left(X^T\boldsymbol{y} - \boldsymbol{x}_i y_i\right) && (1.98)\\
&= \boldsymbol{y}^T\boldsymbol{y} - y_i^2 - \hat{\boldsymbol{\beta}}^T\left(X^T\boldsymbol{y} - \boldsymbol{x}_i y_i\right) + \frac{1}{(1-g_i)}\left((X^TX)^{-1}\boldsymbol{x}_i r_i\right)^T\left(X^T\boldsymbol{y} - \boldsymbol{x}_i y_i\right) && (1.99)\\
&= \boldsymbol{y}^T\boldsymbol{y} - \hat{\boldsymbol{\beta}}^T X^T\boldsymbol{y} - y_i^2 + \hat{\boldsymbol{\beta}}^T\boldsymbol{x}_i y_i + \frac{r_i}{(1-g_i)} && (1.100)\\
&\left(\boldsymbol{x}_i^T(X^TX)^{-1}X^T\boldsymbol{y} - \boldsymbol{x}_i^T(X^TX)^{-1}\boldsymbol{x}_i y_i\right) \\
&= (m-k)s^2 - y_i\left(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}\right) + \frac{r_i}{(1-g_i)}\left(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} - g_i y_i\right) && (1.101)\\
&= (m-k)s^2 - y_i r_i + \frac{r_i}{(1-g_i)}(\hat{y}_i - g_i y_i) && (1.102)\\
&= (m-k)s^2 + \frac{r_i}{(1-g_i)}(\hat{y}_i - g_i y_i - y_i + g_i y_i) && (1.103)
\end{aligned}$$

$$= (m-k)s^2 - \frac{r_i^2}{(1-g_i)}. \tag{1.104}$$

The difference between the ith observation and the ith prediction obtained when the ith observation is excluded in the estimation is used to examine whether the deletion of the ith observation has a remarkable effect on prediction. This difference $y_i - \hat{y}_{(i)} = y_i - x_i^T\hat{\boldsymbol{\beta}}_{(i)}$ has variance $\sigma^2\left\{1 + \boldsymbol{x}_i^T(X_{(i)}^TX_{(i)})^{-1}\boldsymbol{x}_i\right\}$, which can be estimated by $s_{(i)}^2$. $\hat{y}_{(i)}$ and $s_{(i)}^2$ are both independent of y_i. The prediction of the ith observation that is obtained with the ith observation deleted is $\hat{y}_{(i)i}$ and it is abbreviated to $\hat{y}_{(i)}$. The test statistic for testing whether the ith observation is an outlier is the *deletion residual*, and it is defined as

$$r_i^* = \frac{y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{(i)}}{s_{(i)}\sqrt{1 + \boldsymbol{x}_i^T(X_{(i)}^TX_{(i)})^{-1}\boldsymbol{x}_i}} = \frac{r_i}{s_{(i)}\sqrt{(1-g_i)}}, \tag{1.105}$$

where the last equality is obtained by observing that

$$\boldsymbol{x}_i^T(X_{(i)}^TX_{(i)})^{-1}\boldsymbol{x}_i = \boldsymbol{x}_i^T(X^TX)^{-1}\boldsymbol{x}_i + \frac{\boldsymbol{x}_i^T(X^TX)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^T(X^TX)^{-1}\boldsymbol{x}_i}{1-g_i} \tag{1.106}$$

$$= g_i + \frac{g_i^2}{1-g_i} \tag{1.107}$$

$$= \frac{g_i}{1-g_i}, \tag{1.108}$$

and

$$\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{(i)} = \boldsymbol{x}_i^T\,\hat{\boldsymbol{\beta}} - \frac{\boldsymbol{x}_i^T(X^TX)^{-1}\boldsymbol{x}_ir_i}{1-g_i} \tag{1.109}$$

$$= \hat{y}_i - \frac{g_ir_i}{1-g_i}. \tag{1.110}$$

It follows that

$$r_{(i)} = y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{(i)} \tag{1.111}$$

$$= r_i + \frac{g_ir_i}{1-g_i} \tag{1.112}$$

$$= \frac{r_i}{1-g_i}. \tag{1.113}$$

Thus it is easy to find the distribution of the deletion residual if there is no outlier, which is given in the following result. This result is given in Atkinson [1], but the proof of the result cannot be found in [1]. I give a proof after presenting the result.

Result 1.4.4 *When Assumption 1.4.1 is satisfied for all observations, the deletion residual r_i^* has a central t distribution on $m-k-1$ degrees of freedom.*

Proof.

Under Assumption 1.4.1, the vector of observations $\boldsymbol{y}$ has a multivariate normal distribution

$$\boldsymbol{y} \sim \mathrm{MVN}(X\boldsymbol{\beta}, \sigma^2 I),$$

and the vector of observations after deleting y_i still has a multivariate normal distribution, *i.e.*

$$\boldsymbol{y}_{(i)} \sim \mathrm{MVN}(X_{(i)}\boldsymbol{\beta}_{(i)}, \sigma^2 I). \tag{1.114}$$

Because all y_i's are independent (Assumption 1.4.1 (2)), $y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{(i)}$ still follows a normal distribution with a mean

$$\begin{aligned}
\mathrm{E}\left[y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{(i)}\right] &= \mathrm{E}(y_i) - \boldsymbol{x}_i^T(X_{(i)}^TX_{(i)})^{-1}X_{(i)}^T \cdot \mathrm{E}(\boldsymbol{y}_{(i)}) && (1.115)\\
&= \boldsymbol{x}_i^T\boldsymbol{\beta} - \boldsymbol{x}_i^T(X_{(i)}^TX_{(i)})^{-1}X_{(i)}^TX_{(i)}\boldsymbol{\beta}_{(i)} && (1.116)\\
&= \boldsymbol{x}_i^T\boldsymbol{\beta} - \boldsymbol{x}_i^T\boldsymbol{\beta}_{(i)} = 0, && (1.117)
\end{aligned}$$

and a variance

$$\mathrm{Var}\left[y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{(i)}\right] = \mathrm{Var}(y_i) + \boldsymbol{x}_i^T(X_{(i)}^TX_{(i)})^{-1}X_{(i)}^T\left(\boldsymbol{x}_i^T(X_{(i)}^TX_{(i)})^{-1}X_{(i)}^T\right)^T\mathrm{Var}(\boldsymbol{y}_{(i)}) \tag{1.118}$$

$$\begin{aligned}
&= \sigma^2 + \boldsymbol{x}_i^T(X_{(i)}^TX_{(i)})^{-1}X_{(i)}^TX_{(i)}(X_{(i)}^TX_{(i)})^{-1}\boldsymbol{x}_i\sigma^2 I_{m-1} && (1.119)\\
&= \sigma^2 + \boldsymbol{x}_i^T(X_{(i)}^TX_{(i)})^{-1}\boldsymbol{x}_i\sigma^2 && (1.120)\\
&= (1 + \boldsymbol{x}_i^T(X_{(i)}^TX_{(i)})^{-1}\boldsymbol{x}_i)\sigma^2. && (1.121)
\end{aligned}$$

Thus $z_i = \frac{y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{(i)}}{\sigma\sqrt{1+\boldsymbol{x}_i^T(X^TX)^{-1}\boldsymbol{x}_i}} = \frac{1}{\sigma}r_{(i)}\sqrt{1-g_i}$ has a standard normal distribution. Then by the Theorem 1.4.1, $(m-p-1)s_{(i)}^2/\sigma^2$ has a central chi-square distribution with degrees of freedom $m-k-1$. Since $(y_i, \hat{\boldsymbol{\beta}}_{(i)})$ and $s_{(i)}^2$ are independent (Result 1.4.2 (3)), so are z_i and $s_{(i)}^2$. Thus, $r_i^* = \frac{z_i}{(m-p-1)s_{(i)}^2/\sigma^2}$ has a central t distribution on $m-k-1$ degrees of freedom. ∎

The central t distribution of r_i^* when there is no outlier is a standard result presented in many literature (for example [1]). However, if there is more than one outlier, the marginal distribution of r_i^* is not a central t distribution even if the ith observation is not an outlier. I show this result by giving a proper distribution for outliers in Chapter 3.

The problem of identifying single outlier can be solved as a single hypothesis testing problem by using the deletion residual as the test statistics.

1.4.3 Review of methods for identifying multiple outliers in regression

In this subsection, I clarify the methods for identifying multiple outliers in regression analysis that appears in the literature as "single-step", "backward-search" and "forward-search". A single-step method identifies outliers in one step; a backward search starts from the set of all the observations, and in each step examines the "outlyingness" of all observations included and excludes the most extreme one; a forward search starts from a subset of observations that excludes all outliers and in each step an observation is reincluded by a certain rule. The word "*outlyingness*" means the plausibility that an observation is an outlier and has been used in Atkinson, Riani and Cerioli [6] and Hadi and Simonoff [45].

The method of identifying a single outlier, which is based on the deletion of a single observation, is introduced in the previous section. The algebra of the deletion of a single observation can be generalized to the deletion of a set of observations, and hence the deletion residual calculated from multiple deletions can be used to test multiple outliers. The multiple deletions is a single-step method which is also introduced in Cook and Weisberg [25] and Atkinson [1]. When the number of outliers is small, for example two or three, the number of deletion residuals that need to be calculated is moderate, $\binom{m}{2}$ or $\binom{m}{3}$. However, the number of outliers is usually unknown and small in real data sets. When m observations are included in the dataset, then $\sum_{i=1}^{m'}\binom{m}{i}$ hypotheses are need to be tested, where m' is a moderate number smaller than m. If the number of atypical observations are more than that of typical observations, then typical observations become atypical. Therefore, the number of outliers is usually smaller than or equal to $m/2$. Data coming from two clusters of equal size is an example for $m_0 = m_1$, where m_0 and m_1 denote respectively the number of typical observations and that of atypical observations [6]. When m and m_1 are both large, it requires massive computation time to test all $\sum_{i=1}^{m'}\binom{m}{i}$ hypotheses.

Since the single step multiple deletion of Cook and Weisberg [25] and Atkinson [1] is time consuming, one may consider to use multi-step methods instead. It is straightforward to use single-case diagnostics backward, that is, include all observations for estimation and identify potential outliers from most to least extreme (Prescott [70], Tietjen, Moore, and Beckman [95]). Such methods are backward search methods. Hadi and Simonoff [44] gave a review of early works of multi-step methods, but they did not distinguish forward

methods from backwards methods. The methods using single-case diagnostics backward are criticized for having the "masking" problem [48], that is, multiple outliers may hide the effect of each other because including other outliers in parameter estimation may lead to a small residual for one outlier [4, 5, 45]. Hawkins [49] suggested the forward search that excludes all possible outliers and then tests excluded observations sequentially for reinclusion. However, the starting data set and the test statistics for reinclusion were not specified in [49]. On the other hand, the classical least square estimator is criticized for its lack of robustness since a single outlier can have a large effect on the estimates. Rousseeuw [78] proposed the *least median of squares* estimator which minimizes the median of the squared residual, whereas the least square estimator minimizes the sum of the residual squares.

Recently Atkinson [2] combined the least median of squares with forward search and this method is introduced in detail in the books [3] and [6]. This forward search starts from a small, robustly selected subset of observations which excludes all outliers. Then the size of the subset used for estimating increases sequentially, and in each step, the least extreme observation among those excluded is tested for outlyingness before being reincluded. The search stops if the least extreme observation among those excluded is declared significant at that step. A drawback of Atkinson's forward search is the difficulty in finding the distribution of the test statistic, which is used for testing outlyingness in each step. Also an adjustment to multiplicity is needed for this step-wise procedure if all outliers are tested simultaneously. Atkinson and Riani [5] introduced some methods to approximate this distribution. They also addressed the simultaneous inference of their methods in this article. There are also other forward search methods such as those of Hadi [44] and Hadi and Simonoff [45]. All the methods mentioned so far are frequentist methods. Many graphical methods and influence statistics such as plots of leverage and Cook's D are not mentioned in this thesis. These methods are introduced in many standard books, for example, [1, 25, 26, 72].

Outliers usually have strong effects on parameter estimation and lead to wrong models. If one is interested in minimizing the effects of outliers, then robustness of estimates need to be studied. The primary purpose of the thesis is to construct an efficient method to identify outliers rather than to study the robustness of estimates. The choice of explanatory variable may also influence the results of deletion diagnostics. Since the main objective of

the thesis is the identification of outliers, the effects of variable selection are not discussed in this thesis.

1.4.4 ROC curves and AUC

In Chapter 3, *Receiver Operating Characteristic* (ROC) curves, which is also called (relative operating characteristic curves), are used to present the simulation results. I give a brief introduction of ROC analysis in this section. ROC analysis, which was original developed in the field of radar signal detection theory in the 1950's, has recently been used in medicine, radiology, psychology [62, 69]. It is related to diagnostic decision making and provides tools to compare the performance of different testing methods. The ROC curve is obtained by plotting the *true positive rate* (TPR) versus *false positive rate* (FPR) for different thresholds. Multiple testing can be treated as a decision making problem, where each test is decided to be null or alternative. In the context of multiple testing, TPR is the proportion of identified false nulls out of all false nulls, *i.e.* TPR $= \frac{S}{m_1}$ (using symbols in Table 1.1 in Section 1.3.1), which is equal to the average power. TPR is also called *sensitivity* or *hit rate*. In multiple testing, FPR is the proportion of true nulls that is claimed to be alternative to the total number of true nulls, *i.e.* FPR $= \frac{V}{m_0}$ (using symbols in Table 1.1 in Section 1.3.1). FPR is also called *false-alarm rate* and 1-FPR is also called *specificity*. A ROC curve does not depend on a specific selection of a decision threshold, and can describe the performance of a method. The *area under ROC curve* (AUC) can be interpreted as the probability that the posterior probability of a false null being assigned false is higher than that of a true null being assigned false, where a high value of the posterior probability of a null being assigned false favors the decision to call the observation atypical [46, 62, 69]. Clearly if test A is uniformly better than test B in the sense that the ROC curve of A is above that of B, then AUC of A is higher than that of B [69]. Although the converse is not necessarily true, AUC can still be used to assess the overall performance of a testing method.

1.5 Scope of the thesis

So far I have introduced the motivating examples, and stated the problems of multiple hypothesis testing and the problem of multiple outlier identification. I have also reviewed

the methods for solving these problems.

In Chapter 2, I discuss the FDR-controlling method of Benjamini and Hochberg [9], the fixed rejection region method of Storey [89] and the adaptive FDR-controlling method of Black [18], and carry out a clarifying comparison of the three methods. Simulation studies involving a wider range of simulation parameters than those used in Storey [89] and Black [18] are also presented in this chapter. All the computations in this thesis are done by using the "R" language which can be downloaded from the open source "http://www.r-project.org/" [71]. I find that the method of Storey [89] is not necessarily more powerful than the original FDR-controlling method of Benjamini and Hochberg [9], contrary to Storey's claim, and the adaptive procedure of Black [18], which modifies the Benjamini and Hochberg method to incorporate an estimate of the proportion of true nulls, is more powerful than the method of Storey [89]. I also present a novel simulation study for a fair comparison of the method of Benjamini and Hochberg [9] and the method of Storey [89] by setting their parameters such that they both have the same true FDR, and then comparing the respective powers. The results show that the former has superior power for the majority of parameter values of my simulation.

In Chapter 3, I propose a Bayesian multiple testing approach, which is based on the deletion residuals, to identify multiple outliers. I also prove a new result that the marginal distribution of the deletion residual of an observation is a doubly noncentral t distribution when there is more than one outlier by assuming a mean shift for outliers. By assuming prior distributions for the proportion and the mean shift of outliers, I use an importance sampling method to compute the marginal posterior probability that an observation is an outlier given its deletion residual. This posterior probability can be used to measure the outlyingness of an observation. The calculation of the posterior probabilities involves the computation of the density of the doubly noncentral F distribution, which is approximated by using the method of Patnaik [68]. In order to examine the accuracy of Patnaik's approximation, I also propose an algorithm to compute the density of the doubly noncentral t distribution and compare the results obtained by using both methods. At the end of this chapter, the proposed Bayesian method is applied to some simulated datasets. The simulation parameters vary over a set of values, and various priors are employed to study the sensitivity of the posteriors to the priors. For each combination of simulation and prior parameter levels, the area under ROC curves is calculated. Then a factorial design

study is carried out to compare the AUC for different levels of the simulation parameters and prior parameters. The resulting AUC values are high for various choices of priors, indicating that the proposed method can identify the majority of outliers with tolerable error. The results of the factorial design analysis show that the choices of the priors do not appreciably affect the marginal posterior probability that the ith observation is an outlier given the ith deletion residual, as long as the sample size is not too small. Both Patnaik's approximation and the proposed algorithm are used to calculate the densities of the doubly noncentral F distribution for one simulated dataset. The results show that the densities calculated by the former are not far from those computed by the latter, but the former is much faster than the latter.

In Chapter 4, one application of the Bayesian multiple outlier identification method proposed in Chapter 3 is presented. The dataset given in Kanduc *et al.* [56] is analyzed by using the proposed Bayesian method. Kanduc *et al.* [56] examined thirty proteomes for amino acid sequence similarity to the human proteome. They also carried out a linear regression analysis to the level of overlaps of the viral proteomes to the human proteome and the size of viral proteome, and they concluded that three viruses, human T-lymphotropic virus 1, Rubella virus, and hepatitis C virus, present the relatively highest number of viral overlaps to the human proteome. Hereafter these three viruses will be referred to as Kanduc *et al.* identified (KI) viruses. The goal is to identify the outliers in the levels of viral overlaps to the human proteome. In order to study how sensitive the posterior distribution is to the prior distributions, the marginal posterior probabilities that an observation is an outlier given its deletion residual are calculated for the thirty viruses with several choices of prior distributions. The results show that the four viruses with extremely large size (Human herpesvirus 4, Human herpesvirus 6, Variola virus, and Human herpesvirus 5) are more likely to be the outliers than the KI viruses. Hereafter these four viruses will be referred to as the four extremely large (FEL) viruses. Among the other 26 viruses, the KI viruses still cannot be rejected without other viruses being rejected. Then I remove the FEL viruses and apply the proposed Bayesian method to the reduced dataset. The results for the reduced dataset confirm the claim of Kanduc *et al.* [56]. Among the 26 viruses in the reduced dataset, the FEL viruses and Lake victoria marburgvirus have the four largest posterior probabilities of being outliers.

The conclusions and the ideas for future work are presented in the last chapter.

Chapter 2

A clarifying comparison of methods for controlling the false discovery rate

2.1 Introduction

The goal of the traditional hypothesis test is to control the type I error rate below a fixed predetermined level while maintaining high power. In multiple hypothesis testing, a well-defined compound error rate plays the role of the single hypothesis type I error rate, and average power is commonly employed as a criterion to compare the performance of two procedures. The most commonly controlled error rate when testing multiple hypotheses is the family-wise error rate (FWER), which is the probability of committing at least one Type I error out of all the hypotheses tested. In some applications, for example, DNA microarray experiments, one tests a large number of hypotheses simultaneously, while the proportion of true null hypotheses is often high. In this situation one may tolerate more than one false rejection, and thus FWER is considered to provide too strict a control.

As mentioned in Section 1.3.3, Benjamini and Hochberg [9] introduced the concept of the false discovery rate (FDR), which offers a less strict multiple testing criterion than FWER. They also proposed a step-up procedure (**BH**) that controls FDR, determining the rejection region as a function of the FDR. Storey [89] considered the FDR arising from multiple testing with a fixed significance level and proposed an estimator of the FDR which he showed to have positive bias. Choosing this significance level to control the estimated FDR also controls the FDR. By simulation, Storey demonstrated that a multiple testing procedure (**FSL**) with a fixed significance level could have higher power than **BH** while controlling the FDR at the same level. It was pointed out by Black [18], however, that the apparent advantage of Storey's procedure was due to his incorporating an estimate of π_0, the proportion of true nulls. **BH** with target FDR equal to α actually

controls the FDR at level $\pi_0\alpha$, so that it can be extremely conservative if π_0 is small. Black demonstrated that an adaptive procedure (**AFDR**) that adjusts **BH** for this conservatism using Storey's estimator of π_0, $\widehat{\pi}_0$, has power comparable to **FSL**. Storey *et al.* [94] investigated picking a data dependent significance level to control Storey's [89] estimated FDR, which is equivalent to **BH** when the target FDR is inversely proportional to $\widehat{\pi}_0$. More recently, Storey [91] has proposed an alternative method of dealing with multiplicity which is based on the joint likelihood.

Readers of Storey [89] and Black [18] could be left with the impression that **FSL** is always more powerful than **BH** and comparable in power to **AFDR**. In this chapter, I clarify the relationship between **BH** and **FSL** and show analytically and graphically that when the number of true alternatives is relatively small, then **BH** can reject *more* false null hypotheses than **FSL** at the same estimated FDR. This observation seems to contradict a claim made by Storey [89]. Moreover, the adaptive method of Black [18], **AFDR**, in all cases rejects at least as many hypotheses as **FSL** and tends to be more powerful. I also present simulation results comparing the power of **BH** and **FSL** when the *actual* FDR's are the same. Such a "fair" comparison of the procedures has already been published by us in Yin *et al.* [101]. The results presented are for finite sample sizes m (not more than 5000) since it is known in general (although under some mild assumptions) that the random rejection threshold, Γ, for the **BH** procedure approaches a fixed rejection threshold γ (dependent on α and π_0) as $m \to \infty$ (Genovese and Wasserman [38]; Storey *et al.* [94]; and most generally, Ferreira and Zwinderman [33]), that is **BH** is asymptotically equivalent to **FSL**.

In Section 2.2 I present the basic concepts and notation for discussing the methods, using the empirical distribution function of the p-values. In Section 2.3, I compare **BH** and **FSL** and present simulation studies involving 100 and 1000 hypotheses. Section 2.4 proves that modifying **BH** to incorporate an estimate of π_0 as in **AFDR** gives a more powerful procedure than **FSL**. In Section 2.5, I carry out a simulation study for a fair comparison of **BH** and **FSL** by setting their parameters such that they both have the same true FDR, and then comparing the respective powers. A summary is presented in Section 2.6.

2.2 Concepts and Notation

Consider the problem of testing m null hypotheses $h_{1,0}, \ldots, h_{m,0}$ simultaneously, of which m_0 are true nulls. The proportion of true null hypotheses is denoted by $\pi_0 = m_0/m$. Benjamini and Hochberg [9] used R and V to denote respectively the total number of rejections and the number of false rejections, and this notation has persisted in the literature. The notations for all possible outcomes of the m test were given in Table 1.1.

Multiple testing procedures can be described in terms of the empirical distribution functions of the p-values corresponding to the hypotheses $h_{1,0}, \ldots, h_{m,0}$. As defined in Section 1.3.1, let H_i be the indicator for the alternative hypothesis, *i.e.*

$$\begin{aligned} H_i = 0 \quad &\text{if } h_i \text{ is true} \\ \text{and} \quad H_i = 1 \quad &\text{if } h_i \text{ is false,} \end{aligned} \tag{2.1}$$

and define

$$\hat{F}_0(p) = m^{-1} \sum_{i=1}^{m} (1 - H_i) \mathbf{1}_{[0,p]}(P_i) \tag{2.2}$$

$$\hat{F}_1(p) = m^{-1} \sum_{i=1}^{m} H_i \mathbf{1}_{[0,p]}(P_i), \tag{2.3}$$

where $\mathbf{1}_{[0,p]}$ denotes the indicator function of the interval $[0, p]$. Define

$$\hat{F} = \hat{F}_0 + \hat{F}_1. \tag{2.4}$$

Then the number of rejections at a significance level of γ is $R = m\hat{F}(\gamma)$, and the number of false discoveries is $V = m\hat{F}_0(\gamma)$. As mentioned in Section 1.3.1, Storey [89] defined:

$$\text{FDR} = E(V/(R \vee 1)). \tag{2.5}$$

Given a fixed target α for an upper bound on the false discovery rate, **BH** can be described as rejecting all hypotheses $h_{(1),0}, \ldots, h_{(T),0}$ where

$$T = \max\{1 \leq t \leq m : \hat{F}(P_{(t)}) \geq \alpha^{-1} P_{(t)}\}. \tag{2.6}$$

Since $\hat{F}$ is non-decreasing and right-continuous this is equivalent to rejecting all hypotheses h_i for which $p_i \leq \Gamma$, where the random threshold Γ is defined by

$$\Gamma = \sup\{p \in [0, 1] : \hat{F}(p) \geq \alpha^{-1} p\}. \tag{2.7}$$

The right-continuity of $\hat{F}$ also guarantees that $\hat{F}(p) = \alpha^{-1}p$, and it follows that

$$\text{FDR} = \alpha E\left(\frac{\hat{F}_0(\Gamma)}{\Gamma}\right). \tag{2.8}$$

Benjamini and Hochberg [9] proved that if the p-values corresponding to the true null hypotheses are uniformly distributed and independent of each other and of the p-values corresponding to the alternative hypotheses, then **BH** guarantees that $\text{FDR} \leq \pi_0\alpha$ where, as defined previously, $\pi_0 = E(\hat{F}_0(1)) = \hat{F}_0(1)$ is the proportion of true null hypotheses in the collection. Finner and Roters [34] proved under the same assumptions that indeed $\text{FDR} = \pi_0\alpha$. (For related results, see also Benjamini and Yekutieli [11]; Genovese and Wasserman [38]; and Sarkar [81].)

Whereas under **BH** the rejection threshold Γ is a random variable, Storey [89] proposed rejecting hypotheses for which $P_i \leq \gamma$ for a fixed pre-selected rejection threshold γ and then estimating the FDR. Since

$$\text{FDR} = E\left(\frac{\hat{F}_0(\gamma)}{\hat{F}(\gamma)}\right) \quad \text{and} \quad E\left(\hat{F}_0(\gamma)\right) = \pi_0\gamma \tag{2.9}$$

for fixed γ, Storey proposed estimating the FDR by

$$\widehat{\text{FDR}}_\lambda(\gamma) = \frac{\hat{\pi}_0\gamma}{\hat{F}(\gamma)} \tag{2.10}$$

where

$$\hat{\pi}_0(\lambda) = \frac{1 - \hat{F}(\lambda)}{1 - \lambda} \tag{2.11}$$

for some suitably chosen λ, provided that $\hat{F}(\gamma) > 0$. For the case of no rejections ($\hat{F}(\gamma) = 0$), Storey defined $\widehat{\text{FDR}} = m\hat{\pi}_0\gamma$. Since $0 \leq \pi_0 \leq 1$, and $0 \leq \text{FDR} \leq 1$, I truncate their estimates, $\widehat{\pi}_0(\lambda)$ and $\widehat{\text{FDR}}$ to one, should the formulas (2.10) and (2.11) result in values exceeding 1.

The estimator (2.11), which has a positive bias, was proposed as a simple practical solution by Bickis et al. [17], even though it was shown to have a slightly larger mean squared error than another estimator used earlier by Bickis and Krewski [16]. Storey [89] also introduced a bootstrap procedure to facilitate choosing λ. Such a bootstrapped estimator of π_0, however, need not have less bias than (2.11) in which λ is a predetermined constant, as can be seen in Tables 2 and 3 of Black [18]. My own simulations shown in Table 2.1 confirm that the bias of $\hat{\pi}_0$ is relatively insensitive to the choice of λ, and is not reduced by bootstrapping.

Given a fixed γ, Black [18] proposed, instead of **FSL**, implementing **BH** with the target FDR set to the data-dependent value

$$\hat{\alpha} = \widehat{\mathrm{FDR}}_\lambda(\gamma)/\widehat{\pi}_0(\lambda) \tag{2.12}$$

to obtain the adaptive FDR controlling procedure **AFDR**. This guarantees that

$$\mathrm{FDR} = \hat{\alpha}\pi_0 = \widehat{\mathrm{FDR}}_\lambda(\gamma) \cdot \pi_0/\widehat{\pi}_0(\lambda), \tag{2.13}$$

where I simply reject *all* the hypotheses should $\widehat{\pi}_0 = 0$.

For a given target α, Storey *et al.* [94] proposed, instead of **FSL**, choosing the random significance level Γ by

$$\Gamma = \sup\left\{0 \le p \le 1 \ : \ \widehat{\mathrm{FDR}}_\lambda(p) \le \alpha\right\}, \tag{2.14}$$

which they show to be equivalent to **BH** with target level α in the case that $\hat{\pi}_0(\lambda)$ is fixed at 1 rather than being an estimator. They also prove (Storey *et al.* [94], Lemma 2) that for

$$T = \max\{1 \le t \le m : \widehat{\mathrm{FDR}}_\lambda(P_{(t)}) \le \alpha), \tag{2.15}$$

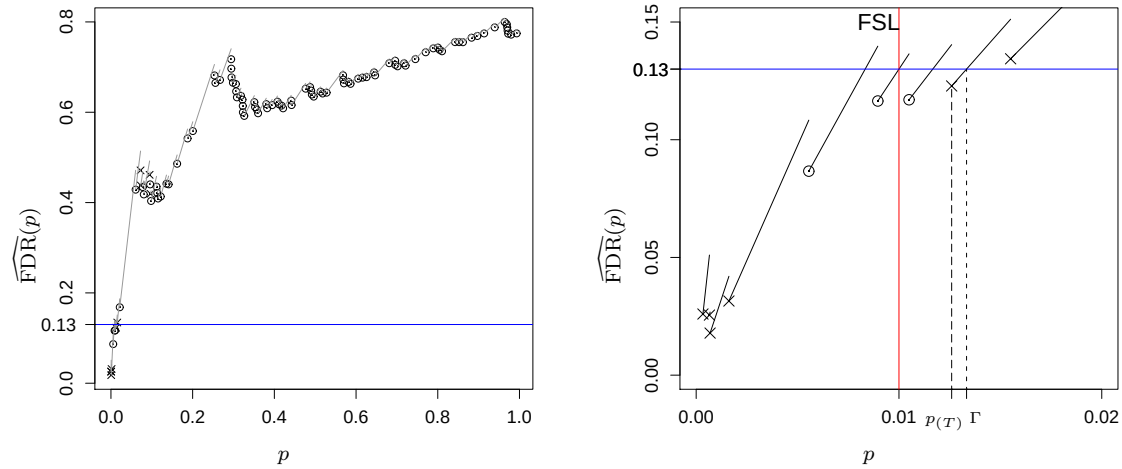
Γ as in (2.14) satisfies:

$$P_{(T)} \le \Gamma < P_{(T+1)}. \tag{2.16}$$

(One can see this graphically as in Figure 2.1.) Thus a procedure which rejects all null hypotheses, $h_{(i)}$, corresponding to the $P_{(i)}$ for $i \le T$ is equivalent, in terms of the rejection region, to one which rejects all null hypotheses corresponding to the $P_{(i)} \le \Gamma$, $i = 1, ..., m$. Hence the Storey *et al.* [94] procedure defined by (2.14) is equivalent to one, referred to herein as ***α*AFDR**, in which the target FDR in **BH** is set to the data-dependent value

$$\hat{\alpha} = \alpha/\widehat{\pi}_0(\lambda). \tag{2.17}$$

For a fixed target α, ***α*AFDR** (and hence the Storey *et al.* [94] procedure) can be shown to be equivalent to implementing **AFDR** with a data-dependent value of $\gamma = \Gamma^*$. To see this, first note that $\widehat{\mathrm{FDR}}_\lambda(\gamma)$ is a discontinuous function of γ, where the jump discontinuities occur at the ordered p-values. Although the function is continuously increasing between the jumps, it decreases at the jumps. Hence, if $\widehat{\pi}_0(\lambda) \ge \alpha$ then there exists Γ^* such that $\widehat{\mathrm{FDR}}_\lambda(\Gamma^*) = \alpha$. For any such Γ^*, if $\gamma = \Gamma^*$ in **AFDR**, then clearly $\widehat{\alpha}$ in (2.12) and (2.17) are identical. Hence this gives a procedure which is equivalent to ***α*AFDR**



(a) (b)

Figure 2.1: Figure (a) shows a simulated example of Storey's $\widehat{\mathrm{FDR}}$ (with $\lambda = 0.5$) as a function of the significance level p. For this dataset $\widehat{\pi}_0 = 0.78$. The p-values are shown as o's for the null case and x's for the alternative. Figure (b) shows the magnification of (a) for small p-values to show the relationship between the Storey *et al.* [94] procedure and **αAFDR**. A target FDR of $\alpha = 0.13$ corresponds to a significance level $p = \Gamma$ for the Storey *et al.* [94] procedure and a p-value, $p = p_{(T)}$, for **αAFDR**. Both procedures reject the same set of p-values. The vertical line at $p = 0.01$ corresponds to a fixed significance level of $\gamma = 0.01$ in **FSL** which gives $\widehat{\mathrm{FDR}} = 0.13$.

and that of Storey *et al.* [94]. If $\widehat{\pi}_0(\lambda) < \alpha$ then no such Γ^* may exist. However, in this case both **AFDR** and the Storey *et al.* [94] procedure would reject all hypotheses and in that sense are equivalent. Similarly, given a fixed γ, **AFDR** is equivalent to implementing $\boldsymbol{\alpha}$**AFDR** or Storey *et al.* [94] with a data-dependent $\alpha = \widehat{\mathrm{FDR}}_\lambda(\gamma)$.

Thus, there are a variety of proposed procedures and some relationships have been established between them. However, further clarification of these relationships is possible. Indeed, in the next section I clarify the relationship between **BH** and **FSL**.

2.3 Clarification of the relationship between BH and FSL

Storey explored the connection between the fixed error rate procedure **BH** and his proposed fixed rejection region method **FSL** and concluded that "using the Benjamini and Hochberg [9] method to control FDR at level α/π_0 is equivalent to using the proposed method to control FDR at level α" (Storey, [89, p. 485]). I find this statement to be inaccurate. The method proposed in Storey [89] involves a fixed significance level γ followed by an *estimation* of the FDR, $\widehat{\mathrm{FDR}}_\lambda(\gamma)$, so it is unclear how the "proposed method" *controls* the FDR. Storey's simulations (which I have reproduced in Tables 2.2 and 2.3) compare his proposed method, **FSL**, to **BH** with target level $\widehat{\mathrm{FDR}}_\lambda(\gamma)$. However, these two methods are not equivalent, even in the case $\widehat{\pi}_0 = 1$. In this section, I prove that the two methods do not agree with each other. Indeed, using the **BH** to control FDR at target level α/π_0 can result in more rejections than using **FSL**, leading to a greater power.

A consequence of the results of Storey *et al.* [94] and my discussion of them in Section 2.2 is that **BH** with target level α is equivalent to $\boldsymbol{\alpha}$**AFDR** in the case that $\widehat{\pi}_0(\lambda)$ is fixed at 1 (or equivalently λ is set to 0). Hence T from (2.6) satisfies

$$T = \max\left\{1 \le t \le m : \widehat{\mathrm{FDR}}_0\left(P_{(t)}\right) \le \alpha\right\}, \tag{2.18}$$

where

$$\widehat{\mathrm{FDR}}_0(P_{(t)}) = \frac{\widehat{\pi}_0(0) \cdot P_{(t)}}{t/m} = \frac{P_{(t)}}{t/m}, \tag{2.19}$$

and I have used (2.10) and the fact that $\hat{F}(P_{(t)}) = t/m$.

However, this procedure is not equivalent to **FSL** because $\widehat{\mathrm{FDR}}_0(P_{(t)})$ is not a monotone function of $P_{(t)}$. Suppose that, as prescribed by **FSL**, one rejects all p-values less than or equal to a fixed rejection threshold γ, and obtains an estimate $\widehat{\mathrm{FDR}}_\lambda(\gamma)$. Then choose

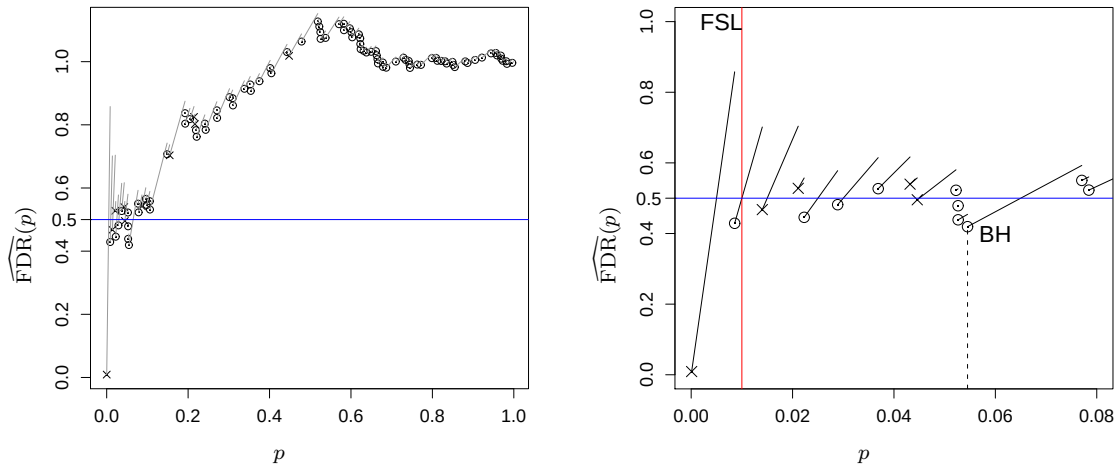
$\alpha = \widehat{\text{FDR}}_0(\gamma)$ in (2.18) and use (2.19) to obtain the number of rejections for **BH** as

$$T = \max\left\{1 \le t \le m : \frac{P_{(t)}}{t} \le \frac{\gamma}{r}\right\}, \tag{2.20}$$

where

$$r = \max\left\{1 \le t \le m : P_{(t)} \le \gamma\right\} \tag{2.21}$$

is the number of rejections for **FSL**. Since, by the definition of r, $P_{(r)}/r \le \gamma/r$, hence $T \ge r$. Indeed it is possible for T to be strictly greater than r and this is illustrated in Figure 2.2 for a data set that yields, after truncation, $\widehat{\pi}_0(0.5) = 1$ (this data set is the same as that depicted in Figure (a)). For this example with $\gamma = 0.01$, **FSL** gives $r = 2$ rejections and one false discovery whereas **BH** has $T = 13$ rejections and 8 false discoveries.



(a) (b)

Figure 2.2: Figure (a) shows a simulated example of Storey's $\widehat{\text{FDR}}$ (with $\lambda = 0.5$) as a function of the significance level p. This example is the same as that of Figure 2.4(a) and gives $\widehat{\pi}_0 = 1.1$ which is truncated to 1. The p-values are shown as o's for the null case and x's for the alternative. Figure (b) shows the magnification of (a) for small p-values to show the relationship between **FSL** and **BH**. The vertical line at $p = 0.01$ corresponds to the fixed significance level γ for **FSL** which leads to $\widehat{\text{FDR}} = 0.5$. **BH** at this target level will reject more hypotheses than **FSL**. The dashed vertical line labelled **BH** denotes the line $p = p_{(T)}$ where T defines the rejection region for **BH**.

I have thus proved:

Result 2.3.1 *When $\widehat{\pi}_0 \ge 1$, **BH** rejects at least as many hypotheses as **FSL**, and is therefore at least as powerful while controlling the FDR at the same level.*

Storey [89] performed a numerical study to compare **FSL** with **BH**. In this simulation, $m = 1000$ data were generated independently from a simple null distribution $N(0,1)$ and a simple alternative distribution $N(\mu,1), \mu = 2$ with π_0 varying over the range $0.1, 0.2, \cdots, 0.9$. For each observation z_i, the p-value was defined as $P_i = 1 - \Phi(z_i)$, where Φ denotes the standard normal distribution. Two different rejection thresholds $\gamma = 0.01$ and $\gamma = 0.001$ for **FSL** were used in this simulation. The procedure was performed for each value of π_0 and γ over 1000 replications, and the empirical power and $\widehat{\text{FDR}}_\lambda(\gamma)$ were calculated. Throughout this simulation λ was set to a constant 0.5. To make the performance of the two procedures comparable, **BH** was implemented at target level $\widehat{\text{FDR}}_\lambda(\gamma)$, which is a conservative estimate of the true FDR. In this set-up, the two procedures control the FDR at the same level. The average powers of the two methods over 1000 iterations were reported, suggesting that **FSL** may be more powerful than **BH** over all combinations of γ and π_0.

However, for a better comparison of the two procedures, one should consider the effect of m, the number of hypotheses. Moreover, the largest π_0 considered by Storey is 0.9, which may be too low for some situations such as those involving microarray data. In such situations thousands of hypotheses need to be tested, but most of them are expected to be true nulls. Thus I implemented a simulation study with the same set-up as Storey for two different sample sizes, $m = 100$ and $m = 1000$, but for $m = 1000$, I also considered more extreme values of π_0, *i.e.*, 0.95 and 0.99. I also performed simulations with $\mu = 1$, and also with $\gamma = 0.0005$ and $\gamma = 0.0001$ for both values of μ. In this study I truncated $\widehat{\pi}_0$ at 1 in the calculation of $\widehat{\text{FDR}}_\lambda(\gamma)$.

Storey [89] also proposed a bootstrap approach to estimate the optimal λ which minimizes the mean square error of the estimate. Black [18] argued that the bootstrap method may underestimate π_0 and thus results in an underestimate of FDR. Storey and Tibshirani [93] used a spline method to estimate π_0. I firstly compare the estimates of π_0 by using different methods. Table 2.1 presents the simulation estimates of π_0 for $\mu = 2$, $m = 1000, \gamma = 0.01$ and selected values of π_0 by using different methods to calculate λ. In the second and third columns, λ is fixed, while in the last two columns, λ is respectively calculated by using the bootstrap method of Storey [89] and the spline method of Storey and Tibshirani [93]. The results show that the bootstrap method underestimates π_0 and produces a larger bias than using a fixed λ or the spline method, whereas the estimates of

π_0 by using a fixed λ and by using the spline method are close and are not very different from the true values. I do not therefore examine the optimal λ further in this thesis, but rather set $\lambda = 0.5$, the value used in Section 5 of Storey [89].

π_0	$\lambda = 0.5$	$\lambda = 0.9$	bootstrap	spline
0.99	0.990	0.992	0.916	0.993
0.95	0.952	0.951	0.883	0.951
0.9	0.904	0.901	0.839	0.900
0.8	0.809	0.801	0.750	0.798

Table 2.1: Simulation estimates of π_0 by using different estimates of λ for $m = 1000$, $\mu = 2$, $\gamma = 0.01$ and selected values of π_0. The first column list the selected value of π_0; the second and third columns respectively are $\widehat{\pi}_0$ when λ is fixed to be 0.5 and 0.9; the last two columns present $\widehat{\pi}_0$ when λ is calculated by using the bootstrap method of Storey [89] and the spline method of Storey and Tibshirani [93].

Tables 2.2 and 2.3 show the simulation results for $m = 1000$, and $m = 100$, respectively, for $\mu = 2$, and $\gamma = 0.01$ and $\gamma = 0.001$; Tables 2.4 and 2.5 show the simulation results for $m = 1000$, and $m = 100$, respectively, for $\mu = 2$, and $\gamma = 0.0005$ and $\gamma = 0.0001$; Tables 2.6 and 2.7 show the simulation results for $m = 1000$, and $m = 100$, respectively, for $\mu = 1$, and $\gamma = 0.01$ and $\gamma = 0.001$; Tables 2.8 and 2.9 show the simulation results for $m = 1000$, and $m = 100$, respectively, for $\mu = 1$, and $\gamma = 0.0005$ and $\gamma = 0.0001$.

For π_0 varying from 0.1 to 0.9 in Tables 2.2 and 2.3, the results are similar to those reported by Storey [89], with the average power of **FSL** being greater than that of **BH**. For $\pi_0 = 0.95$ and 0.99 in Table 2.2 and $\pi_0 = 0.9$ in Table 2.3, an opposite result appears, that is, the average power of **BH** over the 1000 iterations is greater than that of **FSL**. This result can also be observed for $\pi_0 = 0.9, 0.95, 0.99$, $\mu = 2$, $m = 1000$ and $\gamma = 0.0005$ and 0.0001 in Table 2.4, $\pi_0 = 0.7, 0.8, 0.9$, $\mu = 2$, $m = 100$ and $\gamma = 0.0005$ and 0.0001 in Table 2.5, $\pi_0 = 0.9$, $\gamma = 0.01$, $\mu = 1$, $m = 1000$ and $\pi_0 = 0.6, 0.7, 0.8, 0.9$, $\gamma = 0.001$, $\mu = 1$, $m = 1000$ in Table 2.6, $\pi_0 = 0.5, 0.6, 0.7, 0.8, 0.9$, $\mu = 1$, $m = 100$ and $\gamma = 0.01$ and 0.001 in Table 2.7, $\pi_0 = 0.5, 0.6, 0.7, 0.8, 0.9$, $\mu = 1$, $m = 1000$ and $\gamma = 0.0005$ and 0.0001 in Table 2.8, and $\pi_0 = 0.9$, $\mu = 1$, $m = 100$ and $\gamma = 0.0005$ and 0.0001 in Table 2.8. As the rejection threshold γ becomes smaller or the distance between null and alternative distributions becomes smaller, the result that the average power of **BH** over the 1000 iterations is greater than that of **FSL** occurs for a larger range of π_0, except for extreme small γ and $\mu = 1$ and $m = 100$. When $\mu = 1$ and $m = 100$ and $\gamma = 0.0005$ and 0.0001,

	FDR			Power			
	BH	**FSL**	**AFDR**	**BH**	**FSL**	**AFDR**	$\#\{R_{\mathbf{FSL}} < R_{\mathbf{BH}}\}$
π_0				$\gamma = 0.01$			
0.1	0.0004	0.003	0.003	0.073	0.372	0.373	0
0.2	0.001	0.007	0.007	0.121	0.372	0.373	0
0.3	0.004	0.011	0.011	0.164	0.372	0.373	0
0.4	0.008	0.018	0.018	0.202	0.372	0.374	0
0.5	0.013	0.026	0.026	0.236	0.372	0.374	0
0.6	0.024	0.039	0.039	0.267	0.372	0.375	0
0.7	0.041	0.059	0.060	0.295	0.372	0.375	0
0.8	0.077	0.097	0.099	0.321	0.373	0.379	0
0.9	0.175	0.194	0.202	0.347	0.372	0.385	11
0.95	0.334	0.335	0.365	0.376	0.372	0.403	241
0.99	0.827	0.719	0.837	0.599	0.376	0.618	743
π_0				$\gamma = 0.001$			
0.1	0.00005	0.0007	0.0008	0.017	0.138	0.140	0
0.2	0.0004	0.002	0.002	0.031	0.138	0.140	0
0.3	0.001	0.003	0.003	0.045	0.138	0.140	0
0.4	0.002	0.005	0.005	0.059	0.138	0.141	0
0.5	0.004	0.007	0.007	0.073	0.137	0.141	0
0.6	0.007	0.010	0.011	0.086	0.137	0.142	0
0.7	0.011	0.015	0.016	0.099	0.137	0.143	0
0.8	0.022	0.027	0.030	0.112	0.136	0.146	3
0.9	0.058	0.060	0.069	0.135	0.137	0.157	140
0.95	0.142	0.119	0.154	0.171	0.136	0.183	422
0.99	0.64	0.37	0.65	0.48	0.137	0.49	688

Table 2.2: Simulation estimates of FDR and power for **BH**, **FSL**, and **AFDR** for $m = 1000$, $\lambda = 0.5$, $\mu = 2$ and $\gamma = 0.01, 0.001$. The final column gives the number of simulations (out of 1000) in which **FSL** rejected fewer hypotheses than **BH**. Standard errors are no larger in order of magnitude than the last significant figure reported.

	FDR			Power			
	BH	**FSL**	**AFDR**	**BH**	**FSL**	**AFDR**	$\#\{R_{\mathbf{FSL}} < R_{\mathbf{BH}}\}$
π_0				$\gamma = 0.01$			
0.1	0.0007	0.003	0.003	0.080	0.372	0.383	0
0.2	0.002	0.007	0.007	0.127	0.372	0.384	0
0.3	0.004	0.012	0.013	0.169	0.372	0.386	0
0.4	0.008	0.018	0.019	0.206	0.372	0.389	0
0.5	0.013	0.026	0.028	0.241	0.373	0.394	0
0.6	0.023	0.039	0.043	0.270	0.373	0.398	0
0.7	0.039	0.058	0.067	0.300	0.374	0.409	6
0.8	0.081	0.095	0.117	0.338	0.372	0.425	90
0.9	0.235	0.189	0.276	0.441	0.372	0.496	350
π_0				$\gamma = 0.001$			
0.1	0	0.001	0.001	0.025	0.138	0.154	0
0.2	0.0006	0.002	0.003	0.040	0.138	0.157	0
0.3	0.0009	0.004	0.005	0.054	0.139	0.160	0
0.4	0.001	0.005	0.007	0.068	0.139	0.164	1
0.5	0.002	0.007	0.010	0.084	0.140	0.170	11
0.6	0.006	0.010	0.015	0.101	0.139	0.178	36
0.7	0.013	0.014	0.024	0.127	0.138	0.193	132
0.8	0.029	0.024	0.040	0.160	0.139	0.209	267
0.9	0.064	0.045	0.073	0.181	0.137	0.203	290

Table 2.3: Simulation estimates of FDR and power for **BH**, **FSL**, and **AFDR** for $m = 100$, $\lambda = 0.5$, $\mu = 2$ and $\gamma = 0.01, 0.001$. The final column gives the number of simulations (out of 1000) in which **FSL** rejected fewer hypotheses than **BH**. Standard errors are no larger in order of magnitude than the last significant figure reported, the largest SE being 7×10^{-3}.

	FDR			Power			
	BH	**FSL**	**AFDR**	**BH**	**FSL**	**AFDR**	$\#\{R_{\mathbf{FSL}} < R_{\mathbf{BH}}\}$
π_0				$\gamma = 0.0005$			
0.1	0.0002	0.0005	0.0005	0.011	0.098	0.101	0
0.2	0.0003	0.001	0.001	0.020	0.098	0.101	0
0.3	0.0006	0.002	0.002	0.031	0.098	0.101	0
0.4	0.001	0.003	0.003	0.040	0.098	0.101	0
0.5	0.002	0.005	0.005	0.051	0.098	0.102	0
0.6	0.004	0.007	0.008	0.060	0.097	0.103	0
0.7	0.008	0.011	0.011	0.070	0.097	0.104	0
0.8	0.015	0.018	0.021	0.079	0.097	0.107	12
0.9	0.045	0.042	0.052	0.101	0.097	0.120	208
0.95	0.116	0.085	0.123	0.137	0.097	0.146	472
0.99	0.41	0.250	0.42	0.187	0.096	0.193	516
π_0				$\gamma = 0.0001$			
0.1	0.0002	0.0002	0.0001	0.004	0.043	0.046	0
0.2	0.0004	0.0005	0.0005	0.008	0.043	0.046	0
0.3	0.0002	0.0009	0.0009	0.012	0.043	0.046	0
0.4	0.0007	0.001	0.002	0.016	0.043	0.047	0
0.5	0.001	0.002	0.002	0.021	0.043	0.047	0
0.6	0.002	0.003	0.003	0.026	0.042	0.048	0
0.7	0.003	0.005	0.005	0.030	0.042	0.050	7
0.8	0.008	0.009	0.010	0.038	0.042	0.054	100
0.9	0.028	0.022	0.032	0.058	0.042	0.067	414
0.95	0.056	0.040	0.062	0.067	0.041	0.072	474
0.99	0.092	0.078	0.094	0.053	0.041	0.054	117

Table 2.4: Simulation estimates of FDR and power for **BH**, **FSL**, and **AFDR** for $m = 1000$, $\lambda = 0.5$, $\mu = 2$ and $\gamma = 0.0005, 0.0001$. The final column gives the number of simulations (out of 1000) in which **FSL** rejected fewer hypotheses than **BH**. Standard errors are no larger in order of magnitude than the last significant figure reported.

	FDR			Power			
	BH	**FSL**	**AFDR**	**BH**	**FSL**	**AFDR**	$\#\{R_{\mathbf{FSL}} < R_{\mathbf{BH}}\}$
π_0				$\gamma = 0.0005$			
0.1	0	0.0007	0.001	0.018	0.097	0.119	0
0.2	0	0.001	0.002	0.029	0.098	0.122	0
0.3	0.0007	0.002	0.003	0.040	0.098	0.127	2
0.4	0.001	0.003	0.005	0.051	0.098	0.131	10
0.5	0.002	0.003	0.007	0.064	0.099	0.139	28
0.6	0.005	0.005	0.010	0.082	0.098	0.145	94
0.7	0.008	0.007	0.016	0.102	0.097	0.153	182
0.8	0.016	0.012	0.027	0.121	0.097	0.153	280
0.9	0.034	0.025	0.039	0.121	0.095	0.134	208
π_0				$\gamma = 0.0001$			
0.1	0	0.0003	0.0007	0.008	0.043	0.069	0
0.2	0	0.0005	0.0008	0.014	0.043	0.070	0
0.3	0.0001	0.0007	0.001	0.019	0.043	0.071	8
0.4	0.0001	0.001	0.002	0.025	0.043	0.072	33
0.5	0.0001	0.001	0.002	0.032	0.043	0.073	64
0.6	0.0004	0.001	0.002	0.039	0.043	0.070	122
0.7	0.0008	0.002	0.004	0.045	0.042	0.066	151
0.8	0.002	0.003	0.004	0.050	0.043	0.062	139
0.9	0.004	0.006	0.006	0.047	0.042	0.051	59

Table 2.5: Simulation estimates of FDR and power for **BH**, **FSL**, and **AFDR** for $m = 100$, $\lambda = 0.5$, $\mu = 2$ and $\gamma = 0.0005, 0.0001$. The final column gives the number of simulations (out of 1000) in which **FSL** rejected fewer hypotheses than **BH**. Standard errors are no larger in order of magnitude than the last significant figure reported, the largest SE being 4×10^{-3}.

	FDR			Power			
	BH	**FSL**	**AFDR**	**BH**	**FSL**	**AFDR**	$\#\{R_{\mathbf{FSL}} < R_{\mathbf{BH}}\}$
π_0				$\gamma = 0.01$			
0.1	0.004	0.012	0.013	0.013	0.092	0.099	0
0.2	0.010	0.026	0.028	0.018	0.092	0.100	0
0.3	0.022	0.044	0.046	0.024	0.092	0.101	0
0.4	0.040	0.068	0.072	0.031	0.092	0.103	0
0.5	0.064	0.099	0.104	0.038	0.092	0.105	0
0.6	0.100	0.141	0.152	0.046	0.091	0.109	1
0.7	0.161	0.204	0.223	0.056	0.091	0.117	13
0.8	0.276	0.306	0.340	0.075	0.091	0.133	98
0.9	0.522	0.496	0.572	0.151	0.091	0.206	484
π_0				$\gamma = 0.001$			
0.1	0.001	0.006	0.007	0.003	0.018	0.028	0
0.2	0.005	0.012	0.014	0.004	0.018	0.029	0
0.3	0.011	0.021	0.026	0.006	0.018	0.031	2
0.4	0.024	0.033	0.040	0.008	0.018	0.034	13
0.5	0.036	0.050	0.062	0.012	0.018	0.037	39
0.6	0.059	0.072	0.094	0.019	0.018	0.045	112
0.7	0.114	0.107	0.155	0.036	0.018	0.064	285
0.8	0.237	0.174	0.277	0.093	0.018	0.126	516
0.9	0.485	0.31	0.518	0.26	0.018	0.30	693

Table 2.6: Simulation estimates of FDR and power for **BH**, **FSL**, and **AFDR** for $m = 1000$, $\lambda = 0.5$, $\mu = 1$ and $\gamma = 0.01, 0.001$. The final column gives the number of simulations (out of 1000) in which **FSL** rejected fewer hypotheses than **BH**. Standard errors are no larger in order of magnitude than the last significant figure reported.

	FDR			Power			
	BH	**FSL**	**AFDR**	**BH**	**FSL**	**AFDR**	$\#\{R_{\mathbf{FSL}} < R_{\mathbf{BH}}\}$
π_0				$\gamma = 0.01$			
0.1	0.005	0.012	0.015	0.026	0.092	0.151	6
0.2	0.014	0.026	0.036	0.036	0.092	0.163	18
0.3	0.027	0.047	0.061	0.046	0.093	0.169	31
0.4	0.044	0.068	0.094	0.066	0.093	0.186	73
0.5	0.071	0.097	0.138	0.094	0.093	0.207	151
0.6	0.131	0.142	0.207	0.135	0.092	0.239	265
0.7	0.227	0.201	0.311	0.207	0.092	0.301	397
0.8	0.398	0.29	0.467	0.32	0.092	0.40	533
0.9	0.671	0.43	0.707	0.51	0.090	0.57	694
π_0				$\gamma = 0.001$			
0.1	0.003	0.007	0.010	0.008	0.019	0.052	27
0.2	0.005	0.013	0.020	0.011	0.019	0.049	47
0.3	0.009	0.019	0.030	0.013	0.019	0.046	83
0.4	0.014	0.025	0.037	0.017	0.019	0.043	111
0.5	0.018	0.029	0.046	0.020	0.019	0.039	134
0.6	0.030	0.038	0.060	0.022	0.019	0.035	139
0.7	0.041	0.046	0.064	0.022	0.019	0.031	116
0.8	0.056	0.059	0.075	0.024	0.019	0.028	88
0.9	0.077	0.076	0.085	0.022	0.018	0.023	48

Table 2.7: Simulation estimates of FDR and power for **BH**, **FSL**, and **AFDR** for $m = 100$, $\lambda = 0.5$, $\mu = 1$ and $\gamma = 0.01, 0.001$. The final column gives the number of simulations (out of 1000) in which **FSL** rejected fewer hypotheses than **BH**. Standard errors are no larger in order of magnitude than the last significant figure reported.

	FDR			Power			
	BH	**FSL**	**AFDR**	**BH**	**FSL**	**AFDR**	$\#\{R_{\mathbf{FSL}} < R_{\mathbf{BH}}\}$
π_0				$\gamma = 0.01$			
0.1	0.001	0.005	0.006	0.002	0.011	0.024	4
0.2	0.005	0.011	0.014	0.004	0.011	0.026	11
0.3	0.010	0.019	0.024	0.005	0.011	0.029	29
0.4	0.022	0.030	0.039	0.009	0.011	0.035	68
0.5	0.038	0.044	0.060	0.015	0.011	0.041	142
0.6	0.064	0.060	0.093	0.022	0.011	0.049	250
0.7	0.115	0.084	0.152	0.039	0.011	0.067	398
0.8	0.208	0.134	0.246	0.050	0.011	0.077	577
0.9	0.350	0.21	0.381	0.048	0.011	0.059	630
π_0				$\gamma = 0.001$			
0.1	0.001	0.001	0.005	0.002	0.003	0.020	51
0.2	0.004	0.005	0.012	0.002	0.003	0.019	90
0.3	0.008	0.009	0.019	0.003	0.003	0.018	155
0.4	0.016	0.017	0.029	0.004	0.003	0.016	221
0.5	0.023	0.023	0.039	0.005	0.003	0.014	266
0.6	0.032	0.030	0.051	0.006	0.003	0.012	297
0.7	0.045	0.038	0.059	0.005	0.003	0.009	268
0.8	0.063	0.056	0.074	0.005	0.003	0.007	209
0.9	0.079	0.073	0.086	0.004	0.003	0.005	101

Table 2.8: Simulation estimates of FDR and power for **BH**, **FSL**, and **AFDR** for $m = 1000$, $\lambda = 0.5$, $\mu = 1$ and $\gamma = 0.0005, 0.0001$. The final column gives the number of simulations (out of 1000) in which **FSL** rejected fewer hypotheses than **BH**. Standard errors are no larger in order of magnitude than the last significant figure reported.

	FDR			Power			
	BH	**FSL**	**AFDR**	**BH**	**FSL**	**AFDR**	$\#\{R_{\mathbf{FSL}} < R_{\mathbf{BH}}\}$
π_0				$\gamma = 0.01$			
0.1	0.001	0.004	0.005	0.005	0.011	0.026	12
0.2	0.002	0.006	0.009	0.006	0.011	0.024	23
0.3	0.003	0.009	0.014	0.007	0.011	0.023	36
0.4	0.007	0.011	0.016	0.009	0.011	0.021	46
0.5	0.009	0.011	0.017	0.010	0.012	0.019	62
0.6	0.015	0.016	0.019	0.011	0.011	0.017	65
0.7	0.020	0.019	0.023	0.012	0.012	0.016	48
0.8	0.022	0.022	0.024	0.013	0.012	0.015	40
0.9	0.037	0.033	0.038	0.012	0.011	0.013	19
π_0				$\gamma = 0.001$			
0.1	0	0.002	0.001	0.001	0.003	0.005	0
0.2	0	0.003	0.001	0.002	0.003	0.004	1
0.3	0	0.003	0.002	0.002	0.003	0.004	3
0.4	0.001	0.004	0.003	0.002	0.003	0.004	6
0.5	0.001	0.004	0.003	0.002	0.003	0.004	8
0.6	0.002	0.004	0.003	0.003	0.003	0.004	7
0.7	0.003	0.005	0.004	0.003	0.003	0.004	2
0.8	0.003	0.005	0.004	0.003	0.003	0.004	3
0.9	0.005	0.007	0.006	0.003	0.002	0.003	3

Table 2.9: Simulation estimates of FDR and power for **BH**, **FSL**, and **AFDR** for $m = 100$, $\lambda = 0.5$, $\mu = 1$ and $\gamma = 0.0005, 0.0001$. The final column gives the number of simulations (out of 1000) in which **FSL** rejected fewer hypotheses than **BH**. Standard errors are no larger in order of magnitude than the last significant figure reported.

both power and FDR become very small and are similar for **BH** and **FSL.**

The last columns of Tables 2.2 – 2.9 list the numbers of iterations where **FSL** rejects fewer hypotheses than **BH**. Notably, when $m = 100$, $\mu = 1$, $\pi_0 = 0.9$, and $\gamma = 0.01$, there are 694 out of 1000 data sets in which there are more rejections by **BH** than **FSL**. Although one might wonder about the practical use of FDR's as high as the ones found for $\pi_0 = 0.99$ in Table 2.2, it should be noted that such large FDR's are what is required to achieve appreciable power. In this case especially, the power of **BH** substantially exceeds that of **FSL**. These results refute Storey's statement "using my approach, I reject a greater number of hypotheses while controlling the same error rate, which leads to greater power" (Storey, [89, p. 485]).

Figure 2.3 shows how the significance level γ affects the realized FDR and power for an extreme value of $\pi_0 = 0.99$ and various values of m and μ. The sample size and the mean of the alternative distribution vary in Figure 2.3 as: (a) $m = 1000$ and $\mu = 2$; (b) $m = 100$ and $\mu = 2$; (c) $m = 1000$ and $\mu = 1$; (d) $m = 100$ and $\mu = 1$. In all plots of Figure 2.3 I see that appreciable power can only be achieved at the cost of a high FDR, and that for a given γ, **BH** has both higher FDR and higher power than **FSL**. For a given FDR, it can also be seen that **BH** has higher power than **FSL**, whereas the behavior of **BH** and **AFDR** are very similar. I explore the relationship between **BH** and **FSL** further in Section 2.5.

Since **BH** tends to reject more hypotheses than **FSL** when the relative number of alternatives is small, I look at two specific data sets in which $m = 100$ and $\pi_0 = 0.9$. Figure 2.4 (a) and (b) show the p-value empirical distribution plots for two simulated data sets in which $\hat{\pi}_0 = 1.1$ and $\hat{\pi}_0 = 0.76$, respectively. I truncated $\hat{\pi}_0$ if it was greater than one, and obtained $\widehat{\mathrm{FDR}}_{0.5}(0.01)$ equal to 0.5 and 0.38, respectively, for the two data sets. The lower ends of the two plots are magnified in Figures 2.4 (c) and (d), which each shows the critical value for **FSL** of $\gamma = 0.01$ as well as the **BH** critical line of slope α^{-1}, where following Storey [89], α is set to $\widehat{\mathrm{FDR}}_{0.5}(0.01)$. I see that **BH** rejects 13 and 16 p-values for the two data sets, of which 5 and 6 are true alternatives, respectively. However, **FSL** rejects only two hypotheses in each data set, of which one and two are true alternatives, respectively. This demonstrates:

Result 2.3.2 *There exist data sets for which **BH** rejects more hypotheses than **FSL**, for both* $\hat{\pi}_0 \geq 1$ *and* $\hat{\pi}_0 < 1$.

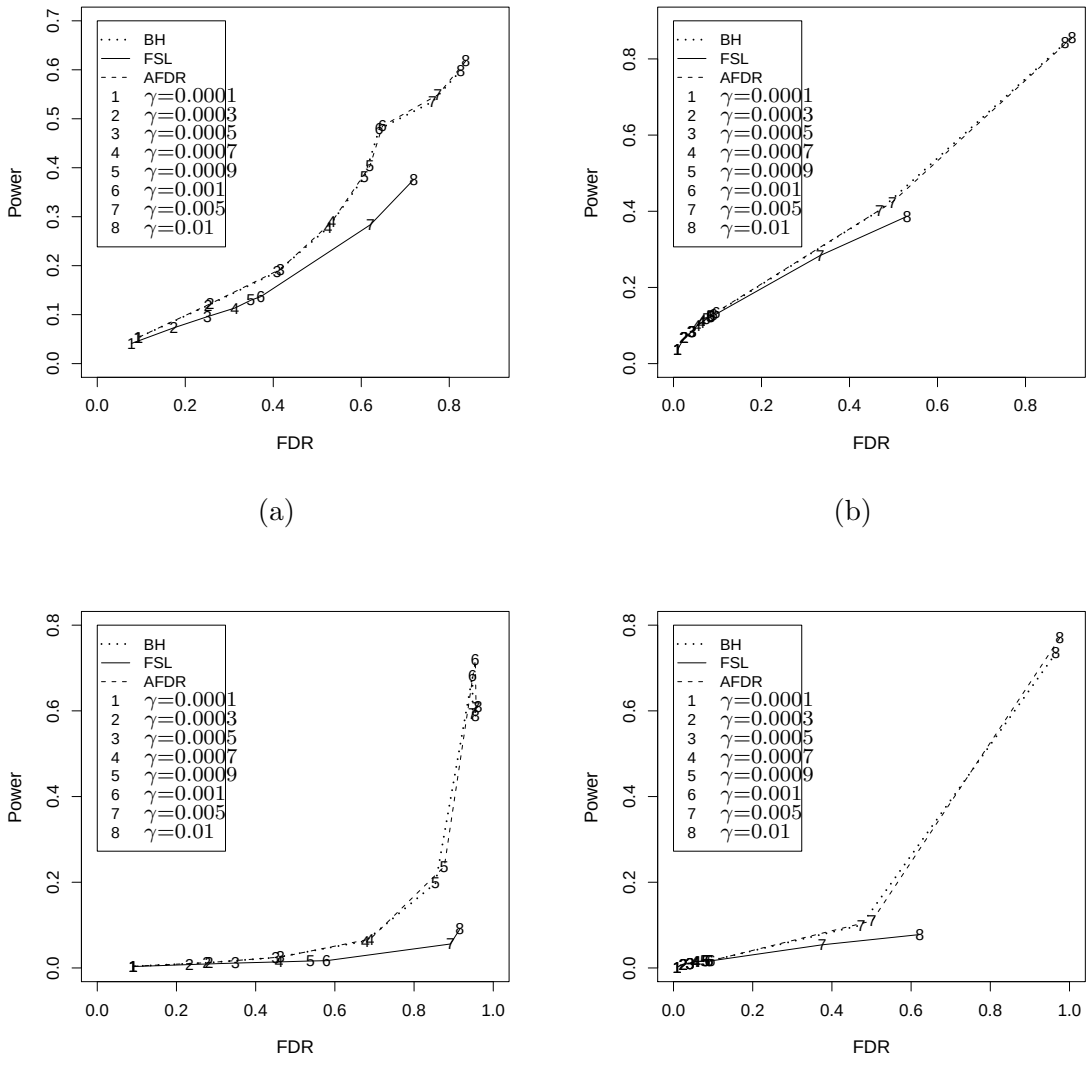


Figure 2.3: Relationship between power and false discovery rate for three methods. (a) $\mu = 2, m = 1000, \pi_0 = 0.99$; (b) $\mu = 2, m = 100, \pi_0 = 0.99$; (c) $\mu = 1, m = 1000, \pi_0 = 0.99$; (d) $\mu = 1, m = 100, \pi_0 = 0.99$. The plotted points for each γ are based on 1000 simulated data sets, each consisting of 1000 p-values.

Storey [89] stated that the two procedures are equivalent if the most conservative estimate $\widehat{\pi}_0 = 1$ is taken. If this statement were true, then the two procedures should have rejected the same number of hypotheses for the two cases, which contradicts my results. Therefore, the two methods are not equivalent.

Since the advantage of **BH** over **FSL** is most notable for large π_0, it might be expected that the data sets in which **BH** rejects more hypotheses would tend to be those for which $\widehat{\pi}_0 \geq 1$. A boxplot of $\hat{\pi}_0$ (before truncation) of the data sets for which **BH** rejects more hypotheses than **FSL** is shown in Figure 2.5. It is interesting that a large percentage of values of $\hat{\pi}_0$ are strictly less than 1.

It is not hard to see that if $\Pr\{P \leq \lambda\} > \lambda$ under the alternative, and the p-values are independent, then $\Pr\{\widehat{\pi}_0 > 1\}$ is an increasing function of π_0. Indeed the event

$$\{\widehat{\pi}_0 > 1\} = \left\{ \sum_{i=0}^{m} X_i < m\lambda \right\} \tag{2.22}$$

where X_i is the indicator of $\{P_i \leq \lambda\}$. Let us order the X_i's such that the first m_0 are nulls and the remainder are alternatives, and write

$$\Pr\{\widehat{\pi}_0 > 1\} = \Pr\left\{ X_{m_0+1} + \sum_{i \neq m_0+1} X_i < m\lambda \right\}. \tag{2.23}$$

Changing one of the alternatives to a null amounts to changing the distribution of X_{m_0+1} to the null distribution, which is stochastically smaller than the alternative. It thus follows that the sum $X_{m_0+1} + \sum_{i \neq m_0+1} X_i$ also is stochastically smaller and thus $\Pr\{\widehat{\pi}_0 > 1\}$ is increased.

Thus I have shown that:

Result 2.3.3 *The higher the true proportion of nulls, the greater the probability that* $\widehat{\pi}_0 > 1$.

2.4 The adaptive FDR controlling procedure

In this section I prove that incorporating the estimate $\widehat{\pi}_0$ in **BH** leads to rejecting at least as many hypotheses as with the fixed rejection region method **FSL** while controlling FDR at the same level.

Given a fixed γ, suppose that r is defined as in (2.21). **AFDR**, as defined in Section

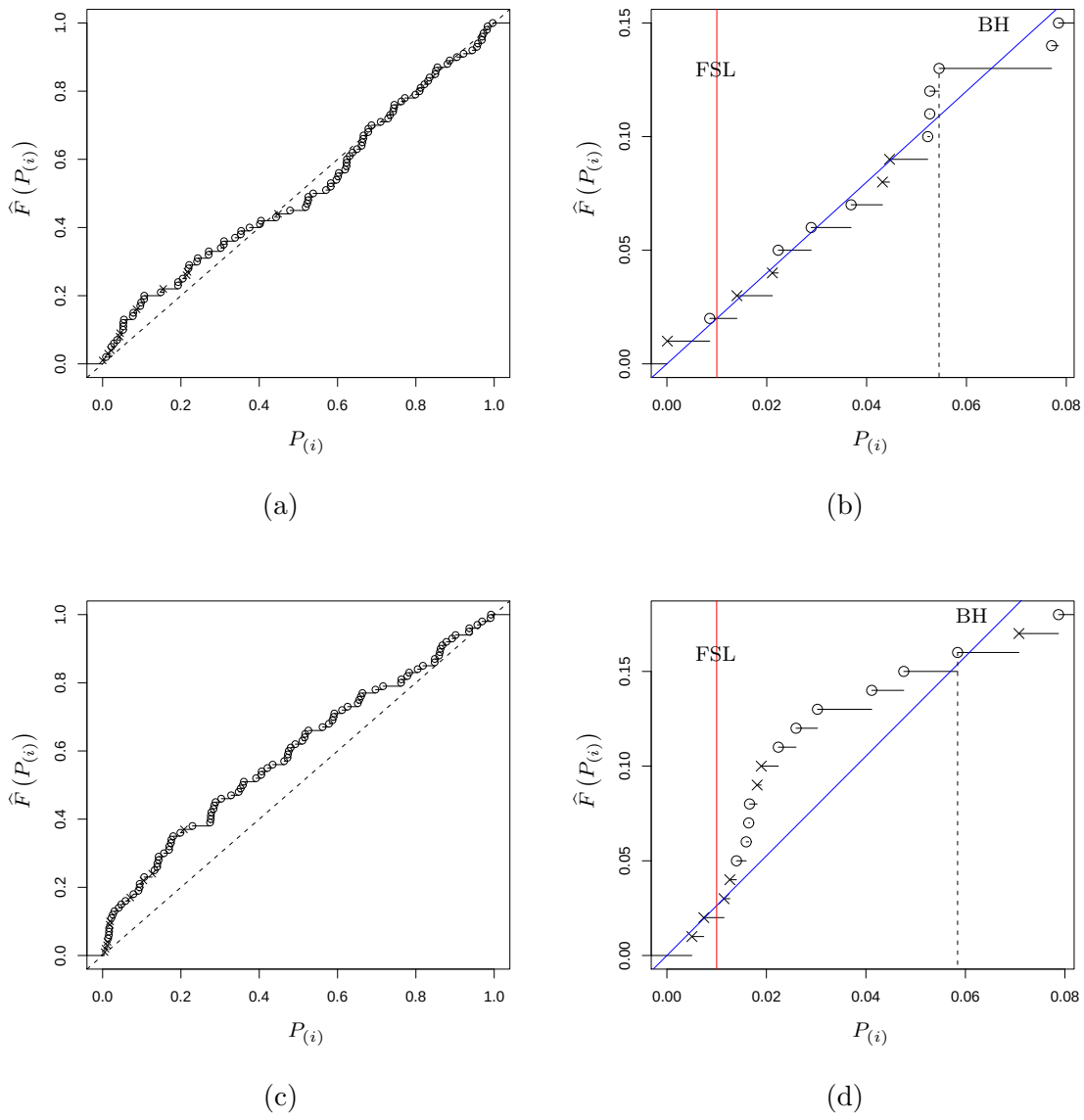


Figure 2.4: Empirical distribution plots of p-values calculated from two datasets with $m = 100$, of which 90 are simulated from $N(0, 1)$ and 10 are simulated from $N(2, 1)$. The upper two graphs (a) and (b) show the diagonal lines corresponding to the null distribution. The lower two graphs (c) and (d) magnify respectively the lower ends of plots (a) and (b) to show the rejection regions of **FSL** (solid vertical line) and **BH** (diagonal line). The p-values are shown as `o`'s for the null case and `x`'s for the alternative. The dashed vertical line indicates the value of the largest rejected p-value.

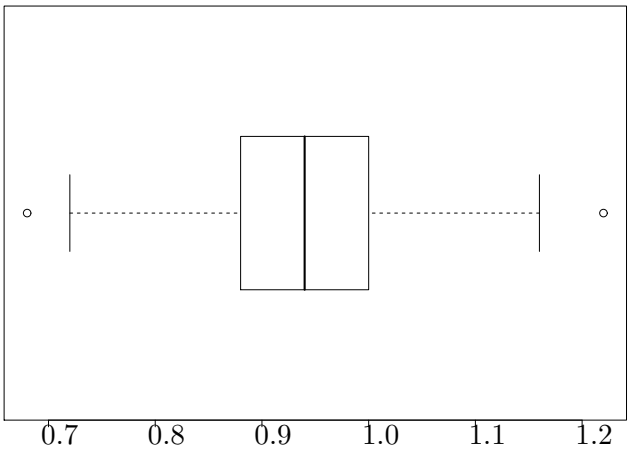


Figure 2.5: Boxplot of untruncated $\hat{\pi}_0$ for the cases in which BH rejects more hypotheses than FSL.

2.2 (cf. equation (2.12)), rejects $h_{(1)}, \cdots, h_{(T)}$ such that

$$T = \max\left\{1 \leq t \leq m : P_{(t)} \leq \frac{\widehat{\mathrm{FDR}}_\lambda(\gamma) \cdot t}{m\widehat{\pi}_0}\right\}. \tag{2.24}$$

Since if $r = 0$, neither **AFDR** nor **FSL** rejects any hypotheses, I only consider the cases where $r > 0$. Since

$$\widehat{\mathrm{FDR}}_\lambda(\gamma) = \frac{\widehat{\pi}_0\gamma}{r/m}, \tag{2.25}$$

therefore

$$T = \max\left\{1 \leq t \leq m : P_{(t)} \leq \gamma\frac{t}{r}\right\}. \tag{2.26}$$

Because $P_{(r)} \leq \gamma = \gamma(\frac{r}{r})$, then $T \geq r$, proving:

Result 2.4.1 *At least as many hypotheses are rejected by* ***AFDR*** *as by* ***FSL****.*

Black [18] reported a discrepancy between his simulations and Storey's claims, but attributed these to "various inaccuracies in the estimates of π_0 and FDR(γ)" (Black [18, p. 300]). However, these differences can in fact be explained by Result 2.4.1. Indeed it is possible for T to be strictly greater than r and this is illustrated in Figure 2.1 (b) for one data set. For this example $m = 100$, $\pi_0 = 0.9$, $\mu = 2$. Figure 2.1 (b) shows that for $\gamma = 0.01$, $\widehat{\mathrm{FDR}}_{0.5}(0.01) = 0.13$ and **FSL** gives $r = 6$ rejections and 2 false discoveries whereas **AFDR** has $T = 8$ rejections and 3 false discoveries.

To illustrate this phenomenon further, I also present simulation results to compare **AFDR** with **FSL** by using the same set-up described in Section 2.4. The results are shown in Table 2.2 - Table 2.9 for two different sample sizes and two different means of alternative distribution and four different rejection thresholds. Overall, I observed the adaptive method to be more powerful than **FSL** for the combinations of simulation parameters used, with the advantage becoming appreciable for large π_0. Closer alternatives ($\mu = 1$) displayed a greater advantage for **AFDR** over **FSL**. The same pattern was seen for $\gamma = 0.0005$ and $\gamma = 0.0001$ although *all* the powers were less than 0.1 with such small values of γ.

2.5 A fair comparison of BH and FSL

In Section 2.3, I performed a simulation study to compare **BH** with **FSL**. I first used **FSL** to calculate $\widehat{\text{FDR}}_\lambda(\gamma)$, and then performed **BH** with the target FDR at $\widehat{\text{FDR}}_\lambda(\gamma)$. As shown in Tables 2.2 - 2.9, the actual FDR's of the two methods differ, making the comparison unfair. Although π_0 and the true FDR are unknown when the procedures are applied to a real data set, I can calculate them in the case of a simulation. Thus a fair comparison, in which the two methods produce exactly the same FDR, is possible.

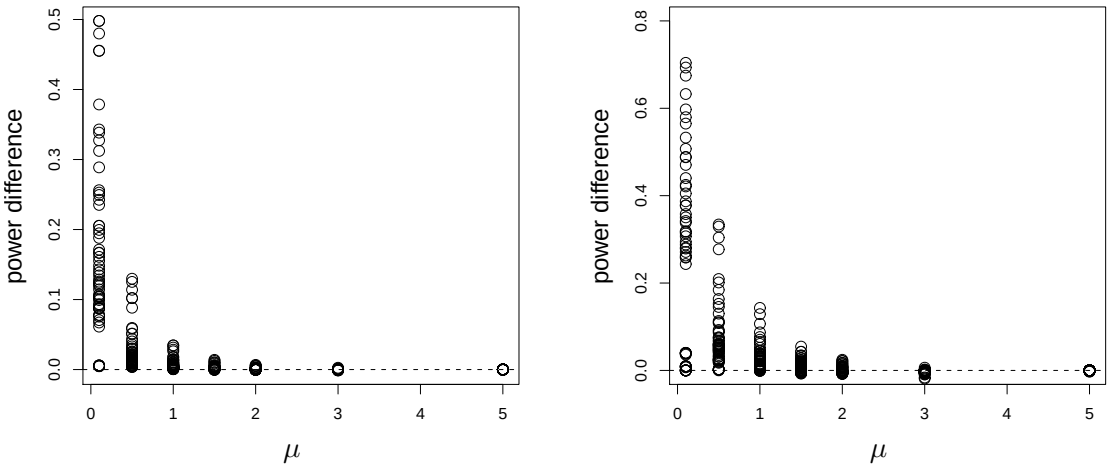
I generated data from a simple null distribution $N(0,1)$ and a simple alternative distribution $N(\mu,1)$, with $\mu \in \{0.1, 0.5, 1, 1.5, 2, 3, 5\}$. The proportion of true nulls π_0 also varied over the range $0.1, 0.2, \cdots, 0.9$. For each case, the sample size m was set to be 100, 1000, or 5000 and different fixed rejection thresholds $\gamma = 0.2, 0.1, 0.05, 0.01, 0.005, 0.001$ were used. For each combination of the parameters, I computed the power of **FSL** as $\Phi\left(\Phi^{-1}(\gamma) + \mu\right)$ as well as the *true* FDR (called SFDR) by

$$\text{SFDR} = \sum_{v=1}^{m_0} \sum_{s=0}^{m_1} \binom{m_0}{v} \binom{m_1}{s} \frac{v}{v+s} \gamma^v (1-\gamma)^{m_0 - v} \delta^s (1-\delta)^{m_1 - s} \tag{2.27}$$

where $\delta = \Phi\left(\Phi^{-1}(\gamma) + \mu\right)$. Then **BH** was performed at target level SFDR/π_0 on 1000 replications of m simulated p-values, giving the same true FDR as that of **FSL** at significance level γ. For each simulation, the power was estimated as the proportion of false hypotheses that were rejected, and this power was averaged over the 1000 replications to give the empirical power of **BH**. The advantage of **BH** over **FSL** was then measured by the difference

$$\text{empirical power of } \mathbf{BH} - \text{power of } \mathbf{FSL}.$$

Figure 2.6 (a) and (b) shows the power difference versus μ when $m = 1000$ and $m = 100$, respectively. In both graphs, the powers of the two procedures are very close when the distance between the alternative and the null distributions is large. However, when the distance is small, **BH** in most cases outperforms **FSL**, although there are some cases for which **FSL** had slightly larger power than **BH**. The numerical results of $m = 100$ are similar to those of $m = 1000$, but with an even greater power difference in favour of **BH**. In Figure 2.7 and 2.8 the power difference and μ are plotted respectively for $m = 1000$ and $m = 100$, and various γ and π_0. When γ is small, the performance of the two methods is almost identical for all μ and π_0. However, **BH** yields greater power than **FSL** in cases with large π_0, which is, for example, typical of microarray experiments.



(a) $m = 1000$ (b) $m = 100$

Figure 2.6: The power advantage of **BH** over **FSL** when the true FDR is fixed at the same level, showing how it relates to the separation μ between the null and alternative distributions. The points represent various choices of γ and π_0 as described in the text, with m fixed at 1000 in (a) and 100 in (b).

To examine the effects of increasing numbers of hypotheses I compared my finite-sample simulations with the asymptotic results of Ferreira and Zwinderman [33]. Because more hypotheses require a stronger adjustment for multiplicity, power will decrease with increasing m. In Figure 2.9 – 2.11 I have plotted the power of **BH** against the true FDR for several values of m, μ, and π_0. The values of m are indicated as numbers in each plot of Figure 2.9 – 2.11, the values of μ and π_0 are shown in the subtitle of every graph in

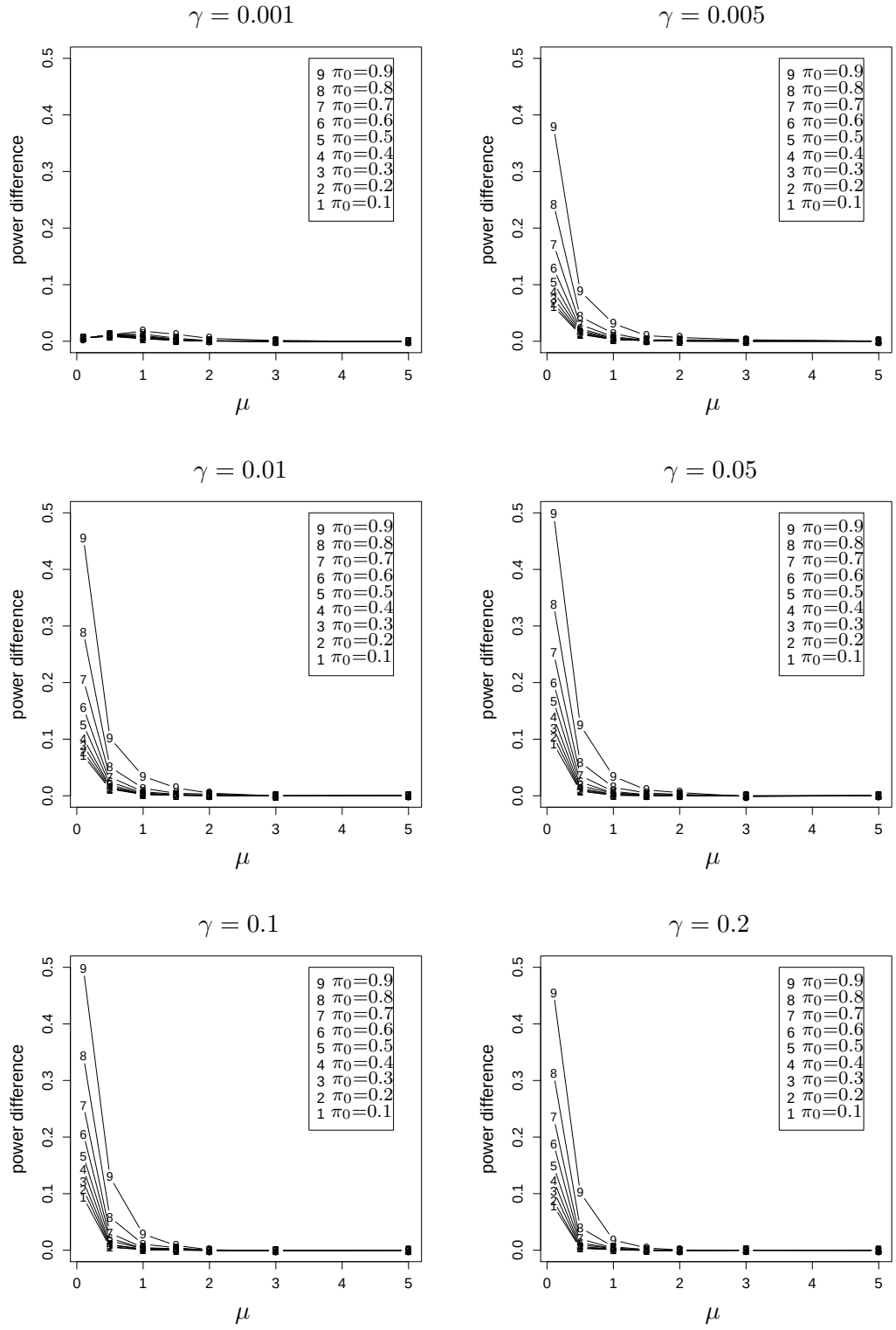


Figure 2.7: The difference between the powers of the two methods versus μ for $m = 1000$ and a range γ's. In each graph, the numbers $1, 2, \cdots, 9$ represent the powers for $\pi_0 = 0.1, 0.2, \cdots, 0.9$, respectively.

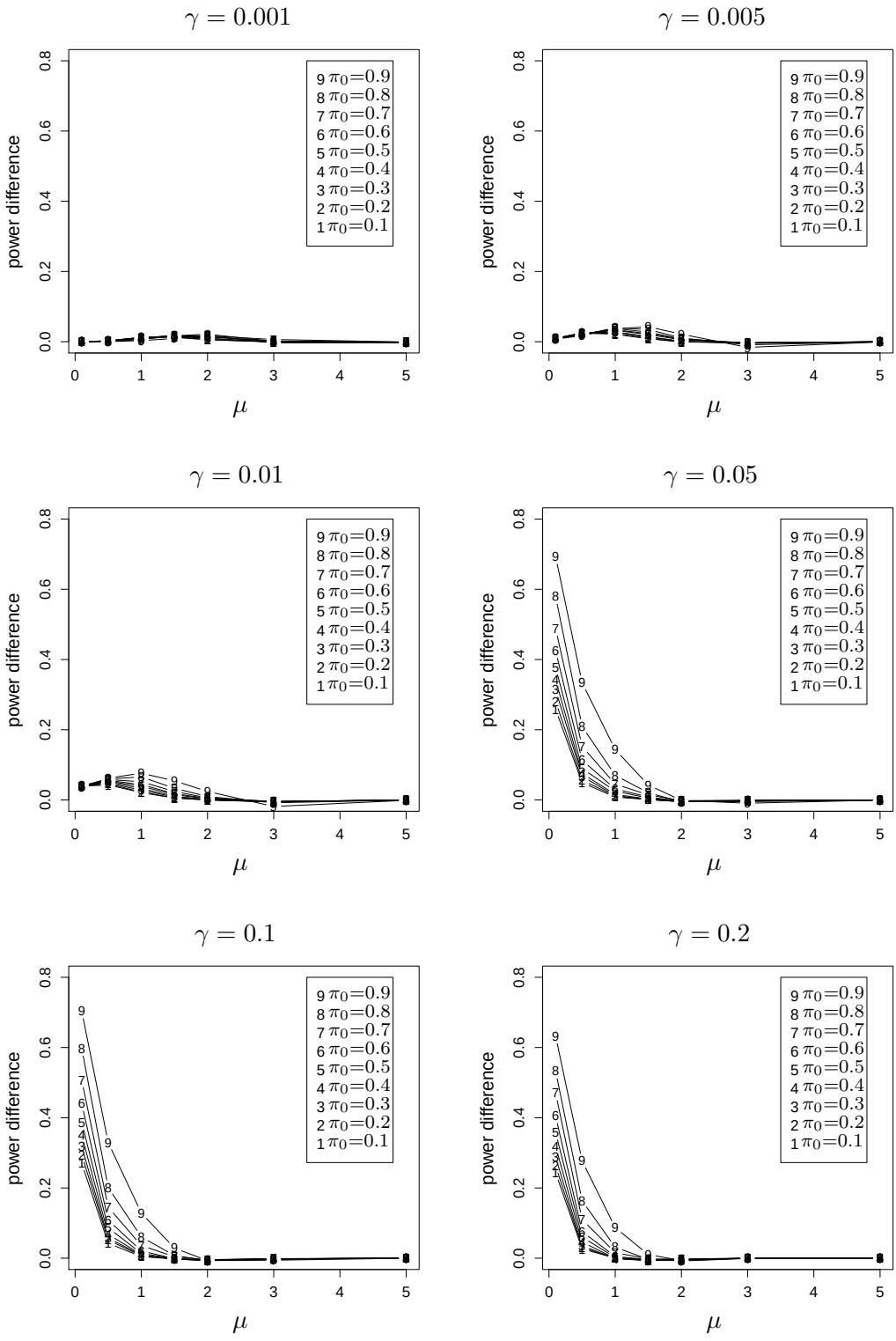


Figure 2.8: The difference between the powers of the two methods versus μ for $m = 100$ and a range γ's. In each graph, the numbers $1, 2, \cdots, 9$ represent the powers for $\pi_0 = 0.1, 0.2, \cdots, 0.9$, respectively.

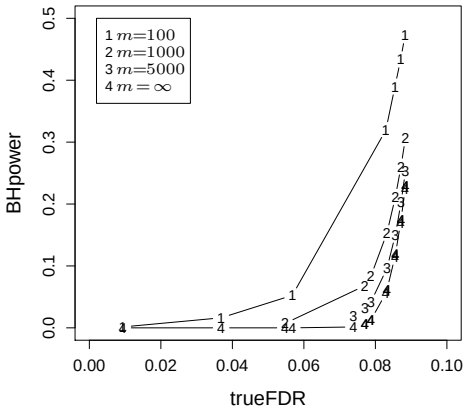
Figure 2.9 – 2.11. Note that for close alternatives and large π_0, the asymptotic power is very low, but for a small number of hypotheses ($m = 100$) the power can be improved if one is willing to accept a high FDR. At $\mu = 1.5$ and $\pi_0 = 0.9$ I see that power is roughly equal to FDR for all values of m. One can also note that not only is the power substantially higher when π_0 is small, but it also depends less on m. As would be expected, the power increases substantially for large μ, and graphs in Figure 2.11 indicate that in such cases the power is not much affected by changes in m.

2.6 Discussion

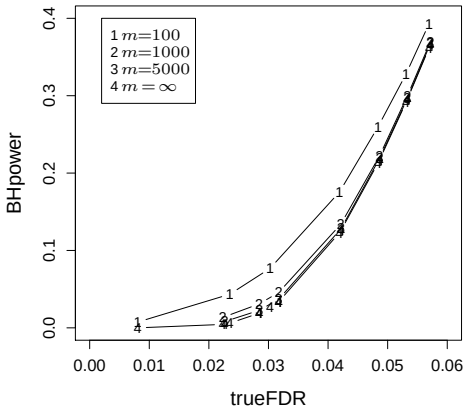
In this chapter I compared the original FDR-controlling method of Benjamini and Hochberg with the fixed rejection region method of Storey and found that the latter is not necessarily more powerful than the former, contrary to Storey's claim. In my simulation study, **BH** performed better than **FSL** when the relative number of alternatives was small (e.g. $m_1 = 10$, $m = 100$ or $m_1 = 50$, $m = 1000$). In the case where the estimate of the proportion of null hypotheses equals one, I have proved that **BH** rejects at least the same number of hypotheses as **FSL**, and therefore has at least the same power. The simulation results showed that, **BH** can reject more hypotheses than **FSL**, and therefore may have greater power. Indeed, in the fair comparison, my simulations for the majority of parameter values showed that **BH** had superior power.

Setting the target FDR level to α in **BH** yields an actual FDR equal to $\pi_0\alpha$. Therefore, this method is quite conservative when the proportion of true null hypotheses is small. However it could become less conservative by incorporating the estimate $\widehat{\pi}_0$ that is used by Storey. By setting the target level to $\widehat{\mathrm{FDR}}_\lambda(\gamma)/\widehat{\pi}_0(\lambda)$, **BH** becomes the adaptive FDR controlling procedure (**AFDR**), as introduced by Black [18]. I have proved that **AFDR** is at least as powerful as **BH** and **FSL**, and have shown that it can be substantially more powerful when π_0 is large.

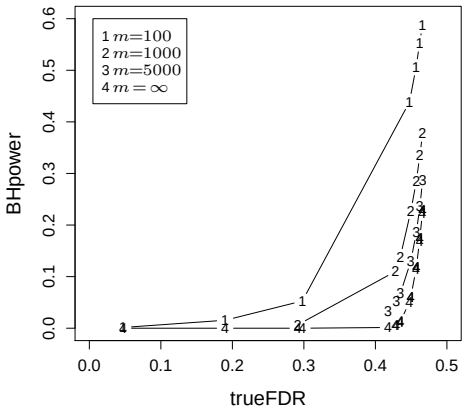
One of the main motivations for recent rapid developments in multiple hypothesis testing procedures is the need to analyze DNA microarray data. DNA microarray data can be used to measure the expression levels for thousands of genes simultaneously. It is interesting to identify the genes that show changes in expression level across two or more different biological conditions. There may be genes that are differentially expressed, but



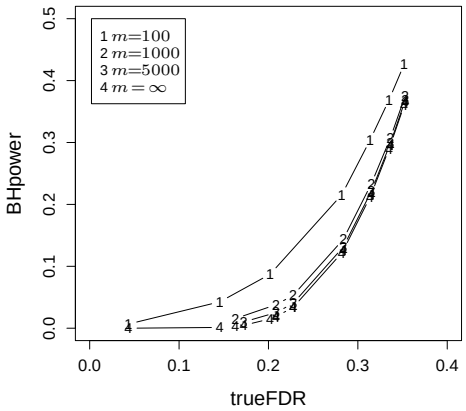
(a) $\mu = 0.1, \pi_0 = 0.1$



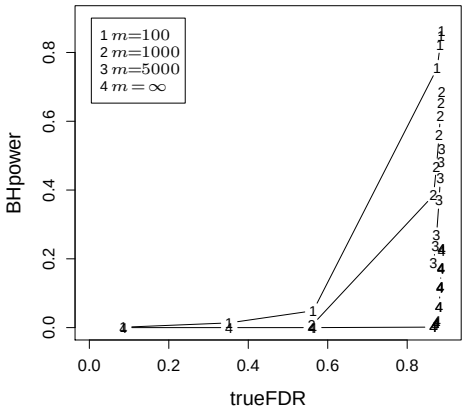
(d) $\mu = 0.5, \pi_0 = 0.1$



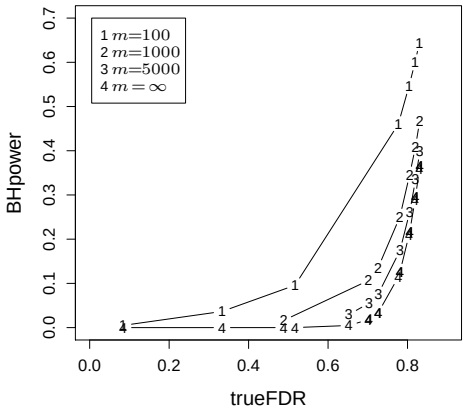
(b) $\mu = 0.1, \pi_0 = 0.5$



(e) $\mu = 0.5, \pi_0 = 0.5$

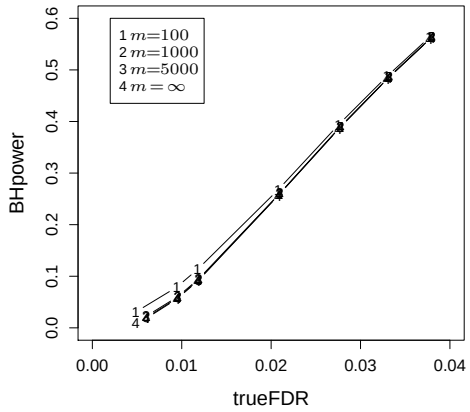


(c) $\mu = 0.1, \pi_0 = 0.9$

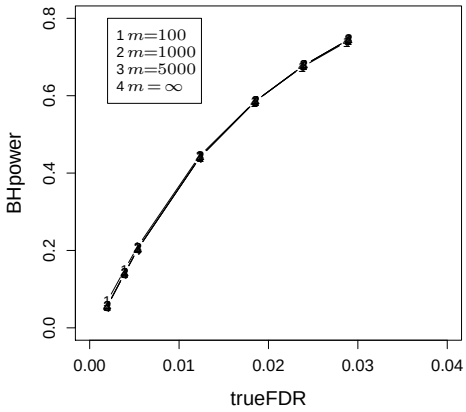


(f) $\mu = 0.5, \pi_0 = 0.9$

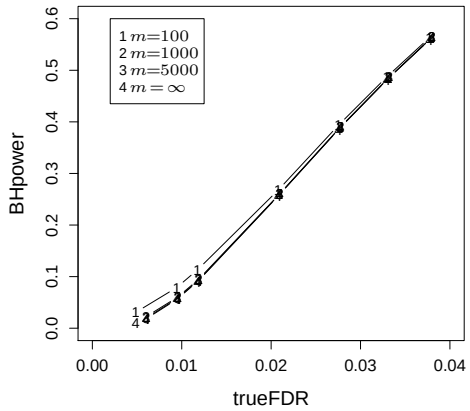
Figure 2.9: Relationship between power and true FDR of BH for different m.



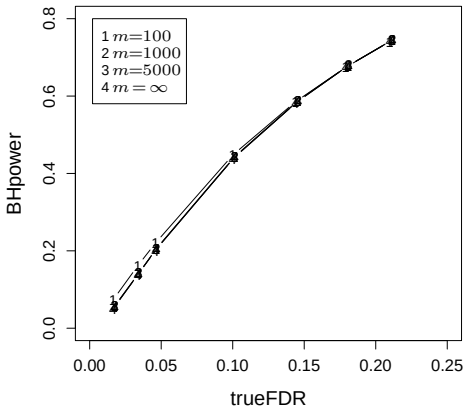
(a) $\mu = 1, \pi_0 = 0.1$



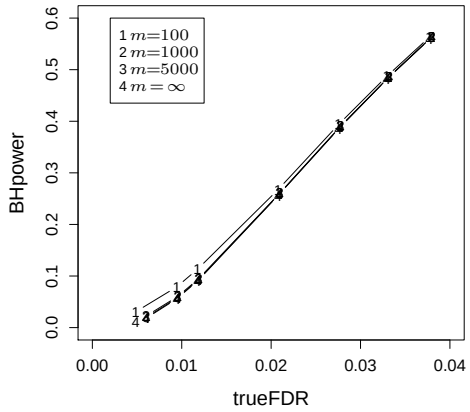
(d) $\mu = 1.5, \pi_0 = 0.1$



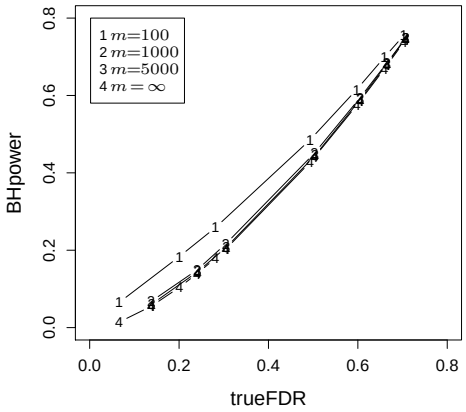
(b) $\mu = 1, \pi_0 = 0.5$



(e) $\mu = 1.5, \pi_0 = 0.5$

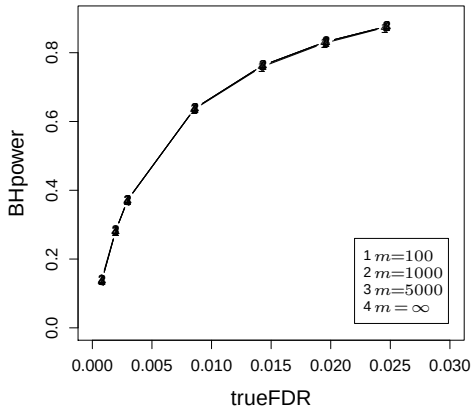


(c) $\mu = 1, \pi_0 = 0.9$

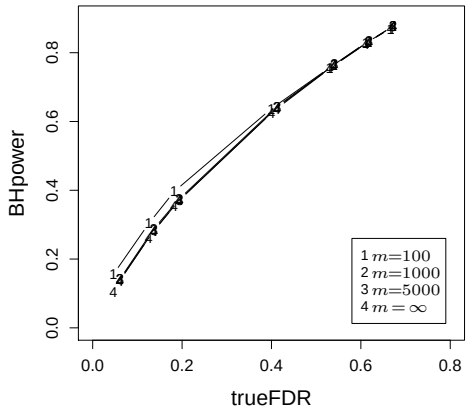


(f) $\mu = 1.5, \pi_0 = 0.9$

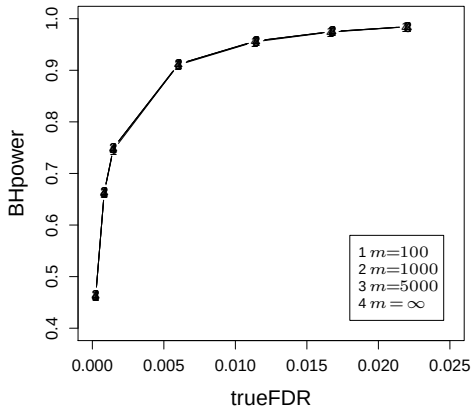
Figure 2.10: Relationship between power and true FDR of BH for different m.



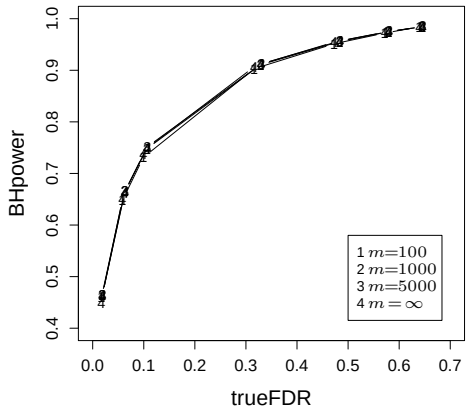
(a) $\mu = 2, \pi_0 = 0.1$



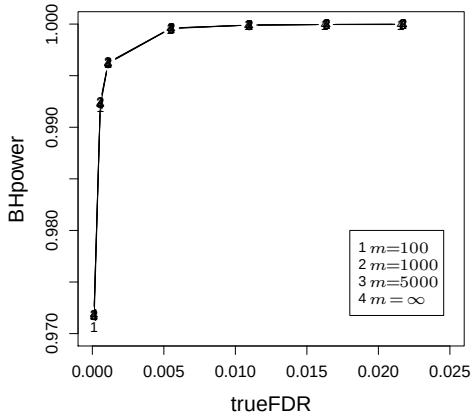
(b) $\mu = 2, \pi_0 = 0.9$



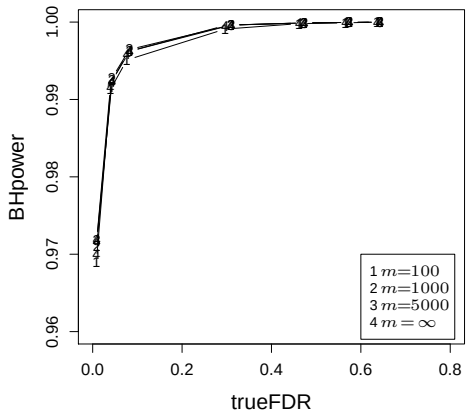
(c) $\mu = 3, \pi_0 = 0.1$



(d) $\mu = 3, \pi_0 = 0.9$



(e) $\mu = 5, \pi_0 = 0.1$



(f) $\mu = 5, \pi_0 = 0.9$

Figure 2.11: Relationship between power and true FDR of BH for different m.

whose distributions are close to those of the reference genes. The capability of detecting such genes is one criterion for assessing the performance of a multiple hypothesis testing procedure. In the microarray context, only a few genes are expected to be differentially expressed, *i.e.* π_0, the proportion of true nulls is close to 1. The fair comparison simulation results in Section 2.5 imply that in this situation the performance of **BH** is better than that of **FSL** as long as one incorporates a good estimate of π_0.

The focus of this chapter has been on the case where the multiple hypothesis tests are independent. This simplifying assumption is unrealistic in many applications, including the analysis of microarray data. Indeed, **BH** has been shown to validly control FDR under certain types of dependence (Benjamini and Yekutieli [11]).

Chapter 3

Multiple deletion diagnostics

3.1 Introduction

Regression analysis is a commonly used statistical tool for modelling and analyzing several variables. The goal of regression analysis is to investigate the relationship between the response variable and one or more explanatory variables. When certain model assumptions are valid, regression models are well developed and the standard results of regression models are introduced in many books, such as [26, 72, 73, 74]. Some useful results of linear regression models were presented in Chapter 1. However, as mentioned there, the model assumptions are often violated for real data because of the existence of outliers. For example, the real dataset given in the next chapter is obtained from Kusalik [58]. This dataset is also given in Kanduc *et al.* [56]. As mentioned in Section 1.1, Kanduc *et al.* showed that the level of viral overlaps to the human proteome has a strong linear relationship to the length of viral proteome based on the analysis of thirty viral proteomes and the human proteomes. One interesting problem is then to identify viruses that show more amino acid sequence similarity to the human proteome than the others, *i.e.* identify outliers (expected to be more than one) when modelling the level of overlaps and the size of viral proteome. My goal is thus to develop a method to identify multiple atypical observations of a dataset.

A powerful method of detecting a single outlier, introduced in Cook and Weisberg [25] and Atkinson [1], is based on the deletion of single observations [6]. This method was introduced in Section 1.4.2. As discussed there, the algebra of the deletion of single observation can be generalized to the deletion of multiple observations, but the methods based on the deletion of multiple observations are time-consuming, when the number of observations and the number of outliers are large. Hence one may consider to use multi-step methods instead. I gave a review of multi-step methods in Section 1.4.2. One problem

of multi-step methods that needs to be emphasized here is the masking problem, especially with the approaches based on the least square method [4, 5, 45]. The residual of an outlier may be small due to the existence of the other outliers, and hence its effect on the residual is said to be masked. The masking problem can be conquered if all observations are tested simultaneously.

As mentioned in Section 1.1, the multiple outlier identification problem in linear regression analysis can be viewed as a problem of multiple hypothesis testing. Each observation can be assigned a null hypothesis that this observation does follow the assumed regression model, and an alternative hypothesis that it is an outlier. By assuming appropriate distributions for null and outliers and prior distributions for distribution parameters, I propose a Bayesian multiple testing approach to identify multiple outliers. In the proposed Bayesian model, it is assumed that outliers have a mean shift, and that the proportion and the mean shift of outliers respectively follow a Beta prior distribution and a normal prior distribution. The proposed Bayesian multiple testing approach is based on the deletion residual. The deletion residual for ith observation is obtained by deleting the ith observation, and follows a central t distribution under the null hypothesis given that it is the only outlier. However, when there is more than one outlier, I prove that the null distribution of the ith deletion residual is no longer a central or noncentral t distribution, and becomes a doubly noncentral t distribution under the proposed Bayesian model. Consequently, the square of the deletion residual has a doubly noncentral F distribution under the same assumptions. The non-central parameters depend on the total number of outliers, which is usually unknown, and therefore marginal p-values cannot be calculated. Thus the frequentist methods based on marginal p-values, such as that of Benjamini and Hochberg [9], are not available for testing more than one outlier. The usual Bayesian models for regression specify prior distributions on the parameters of regression models (coefficients and variance) [14], but the proposed Bayesian model avoids assuming any prior distribution on model parameters, since the distribution of the deletion residuals does not depend on these parameters. In order to find the posterior distribution of the parameters, we need to know the distribution of the estimates. The distributions of some robust estimates, such as the least median of squares and the least trimmed squares [78], are hard to calculate. Although the least square estimates are criticized for lack of robustness, they are still used in the proposed Bayes model since the effects of other outliers, which appear in the noncentrality

of the distribution of the deletion residuals, are included in the model.

In order to identify outliers, I need to calculate the marginal posterior probability that the ith observation is an outlier given all deletion residuals. This requires computing the joint distribution of all deletion residuals, which is complicated. Since the outlyingness of an observation depends more on its own deletion residual than the deletion residual of other observations, the marginal posterior probability that the ith observation is an outlier given the ith deletion residual can be used to measure the outlyingness of the ith observation. In this chapter, I also propose an importance sampling method to calculate this marginal posterior probability. The calculation of the marginal posterior probability requires the computation of the density of deletion residual squares that is achieved by using Patnaik's approximation. Johnson and Kotz [55, p. 139] did not consider Patnaik's approximation to be the best approximation to the density of the doubly noncentral F distribution, but they thought it may be the simplest one. Other approximations to the density of the doubly noncentral F distribution can be found in [19, 20, 55, 96]. The main purpose of Patnaik [68] proposing his approximation is not to compute the density of the doubly noncentral F distribution. The main application of Patnaik's approximation is calculating the CDF of doubly noncentral F distribution [20, 55, 96]. The comparison between Patnaik's method and other approximations for the CDF is presented in [20, 55, 96]. Since the accuracy of Patnaik's approximation to the density of the doubly noncentral F distribution is not examined elsewhere, I also propose an algorithm to calculate the density, and compare the results computed by the proposed algorithm and those by Patnaik's method. The comparison for chosen variable and parameter values shows that using Patnaik's approximation can save considerable computing time without losing much accuracy.

Next I carry out a simulation study on the proposed Bayesian outlier identification method. Simulation parameters vary among a range of values. I also select various prior distributions for the distribution parameters. The marginal posterior probability that the ith observation is an outlier given the ith deletion residual is calculated for all observations in the simulated datasets with various prior parameters. In order to study how sensitive the posterior is to the different priors, I calculate the area under ROC curves, which is introduced in Section 1.4.4, for each combination of simulation and prior parameters. I firstly simulate two single datasets with a fixed number of outliers, different variance of the mean shift, and the explanatory variable arising from two different distributions,

and I compare the calculated AUC for various prior parameters for two datasets. Next I simulate two multiple datasets with a fixed number of outliers, different variances of the mean shift and different distributions of the explanatory variable, and then compare the average AUC for various prior parameters. At last, to include more levels of simulation and prior parameters, I perform a factorial design analysis to compare AUC for a wider range of the simulation parameters and prior parameters. The resulting AUC values are high for various parameters, indicating that the proposed method can identify a majority of the outliers with tolerable error. The results of the factorial design analysis suggest that the priors do not have much effect on the marginal posterior probability that the ith observation is an outlier given the ith deletion residual as long as the sample size is not too small.

The structure of this chapter is described as follows. In Section 3.2, I prove the new result that the marginal distribution of the ith deletion residual is a doubly noncentral t, assuming that there is a mean shift for outliers and more than one outlier exists in a dataset. The proposed Bayesian method is introduced in Section 3.3. In Section 3.4, I propose an algorithm to calculate the density of doubly noncentral F distribution and compare the results calculated by this algorithm with those by Patnaik's method. In Section 3.5, a simulation study is presented. The conclusions of this chapter is given in the last section.

3.2 Distribution of the deletion residual when there is more than one outlier

The standard results from least squares for the linear regression model and those of identifying a single outlier are presented in Section 1.4.1 and 1.4.2. In this section, I use Result 1.4.2 and Lemma 1.4.1 to obtain the new result that the marginal distribution of r_i^* is a doubly noncentral t under appropriate assumptions, when there is more than one outlier in a dataset.

The linear regression model has the form

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3.1}$$

where $\boldsymbol{y}$ is the $m \times 1$ vector of responses, X is the $m \times k$ full-rank design matrix of $k-1$ known vectors explanatory variables, with all elements of the first column equal to 1 and

ith row $\boldsymbol{x}_i^T$, $\boldsymbol{\beta}$ is a vector of k unknown parameters, and $\boldsymbol{\varepsilon}$ is a vector of m unknown random errors. When the model assumptions 1.4.1 are valid, the linear regression model is true for the given dataset.

The deletion residual defined in 1.105 in Section 1.4.2 has the form

$$r_i^* = \frac{y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{(i)}}{s_{(i)} \sqrt{1 + \boldsymbol{x}_i^T (X_{(i)}^T X_{(i)})^{-1} \boldsymbol{x}_i}} = \frac{r_i}{s_{(i)} \sqrt{(1 - g_i)}}. \tag{3.2}$$

It is shown in Result 1.4.4 that the distribution of r_i^* is a central t on $m-k-1$ degrees of freedom if there is no outlier. I prove that however the null distribution of the ith deletion residual is no longer a central t distribution, and becomes a doubly noncentral t distribution when there is more than one outlier. To show this result I first make the following assumptions about the observations when Assumption 1.4.1 is invalid.

Assumption 3.2.1 *Assume that* $\boldsymbol{H}$ *is a fixed unknown vector, and hence the number of outliers* $\sum_{i=1}^{m} H_i = m_1 > 1$ *is also fixed. Let* $m_0 = m - m_1$.

Let $\mathcal{I} = \{1, \cdots, m\}$ be the set of indices of all observations, $\mathcal{I}_1 = \{i_1, i_2, \cdots, i_{m_1}\}$ be that of the atypical observations, and $\mathcal{I}_0 = \mathcal{I} - \mathcal{I}_1$ be that of the typical observations.

Assumption 3.2.2 *Suppose the random errors corresponding to outliers are i.i.d.* $N(\mu_i, \sigma^2)$ *and those of nulls are i.i.d.* $N(0, \sigma^2)$, *where* $\mu_i \neq 0$, $i = i_1, i_2, \cdots, i_{m_1}$. *Let* $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_m)$, *where* $\mu_i = 0$ *if* $H_i = 0$. *All nulls and alternatives are independent. The null and alternative hypotheses for the ith observation can be written:*

$$h_{i,0} : \left(y_i - \boldsymbol{x}_i^T \beta \mid H_i = 0\right) \sim N(0, \sigma^2) \text{ v.s. } h_{i,1} : \left(y_i - \boldsymbol{x}_i^T \beta \mid H_i = 1\right) \sim N(\mu_i, \sigma^2). \tag{3.3}$$

Remark 1 *In Assumption 3.2.2, (1) in Assumption 1.4.1 is violated while (2)-(4) still hold.*

In this section I prove the distribution of r_i^* under the null and alternative hypotheses (3.3) are both doubly noncentral t distributions with different noncentrality parameters. The central t distribution with ν degrees of freedom is defined as the distribution of the ratio of a standard normal variable U to the square root of a central χ^2 variable, *i.e.* $U(\chi_\nu^2/\nu)^{-\frac{1}{2}}$. The random variable with the doubly noncentral t distribution $t''_\nu(\xi, \eta)$ with ν degrees of freedom and noncentrality parameters ζ and η is defined as

$$\frac{U + \xi}{\sqrt{\chi_\nu'^2(\eta)/\nu}}, \tag{3.4}$$

where $\chi^{2\prime}_{\nu}(\eta)$ has a noncentral χ^2 distribution with degrees of freedom ν and noncentrality parameter $\eta = \sum_{i=1}^{\nu}\tau_i^2$, which is defined as the sum of squares of independent standard normal variables U_i plus some constant deviations τ_i, *i.e.*

$$\sum_{i=1}^{\nu}(U_i + \tau_i)^2. \tag{3.5}$$

When $\eta = 0$, the distribution of $\frac{U+\xi}{\sqrt{\chi^{2\prime}_{\nu}(\eta)/\nu}}$ is called the singly noncentral t distribution with ν degrees of freedom and noncentrality parameter ξ, and the variable is denoted by $t'_{\nu}(\xi)$.

Theorem 3.2.1 *Under the Assumptions 3.2.1 and 3.2.2, the distributions of the deletion residual r_i^* are doubly noncentral t distributions $t''_{m-k-1}(\xi_l, \eta_l)$ under both the null ($l = 0$) and alternative ($l = 1$) hypotheses (3.3). When $h_{i,0}$ is true, $\xi_0 = -\frac{1}{\sigma\sqrt{1-g_i}}\sum\limits_{j=i_1,j\neq i}^{i_{m_1}} g_{ij}\mu_j$ and $\eta_0 = \frac{1}{\sigma^2}\boldsymbol{\mu}_{(i)}^T(I - G_{(i)})\boldsymbol{\mu}_{(i)}$, where $\boldsymbol{\mu}_{(i)} = \boldsymbol{\mu}\diagup\{\mu_i\} = (\mu_1, \cdots, \mu_{i-1}, \mu_{i+1}, \cdots, \mu_m)$. When $h_{i,1}$ is true, $\xi_1 = \frac{1}{\sigma}\mu_i\sqrt{1-g_i} - \frac{1}{\sigma\sqrt{1-g_i}}\sum\limits_{j=i_1,j\neq i}^{i_{m_1}} g_{ij}\mu_j$ and $\eta_1 = \frac{1}{\sigma^2}\boldsymbol{\mu}_{(i)}^T(I - G_{(i)})\boldsymbol{\mu}_{(i)}$. When $\eta_l = 0$, the distribution of r_i^* becomes singly noncentral t, while for both $\xi_l = 0$ and $\eta_l = 0$, r_i^* has the central t distribution.*

Remark 2 *Note that the values of the parameter η_k under the null hypothesis and the alternative hypothesis are different. Since the number of outliers is fixed in the assumption, when $h_{i,0}$ is true, $\sum\limits_{j=1,j\neq i}^{m} H_j = m_1$, while when $h_{i,1}$ is true $\sum\limits_{j=1,j\neq i}^{m} H_j = m_1 - 1$, and therefore the number of nonzero elements in $\boldsymbol{\mu}_{(i)}$ under the null hypothesis is m_1 and that under the alternative hypothesis is $m_1 - 1$.*

Proof. [of Theorem 3.2.1]

Under the above assumptions, the vector of observations $\boldsymbol{y}$ has a multivariate normal distribution

$$\boldsymbol{y} \sim \text{MVN}(X\boldsymbol{\beta} + \boldsymbol{\mu}, \sigma^2 I), \tag{3.6}$$

and the vector of observations after deleting y_i still has a multivariate normal distribution

$$\boldsymbol{y}_{(i)} \sim \text{MVN}(X_{(i)}\boldsymbol{\beta}_{(i)} + \boldsymbol{\mu}_{(i)}, \sigma^2 I), \tag{3.7}$$

where $\boldsymbol{H}_{(i)}$ denotes the vector of indices of outliers with the ith observation being deleted. Consider the following two cases.

(a). The null hypothesis $h_{i,0}$ is true.

Then the numerator in (3.2), $y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{(i)} = y_i - \boldsymbol{x}_i^T(X_{(i)}^T X_{(i)})^{-1}X_{(i)}^T\boldsymbol{y}_{(i)}$, still has a normal

distribution under the null hypothesis since all y_i's are independent normal variables, but the mean does not equal to 0. It is easy to calculate the mean of $r_{(i)} = y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{(i)}$. Since $r_{(i)} = \frac{r_i}{1-g_i}$ and

$$\begin{aligned} \mathrm{E}[\boldsymbol{r}] &= \mathrm{E}\left[(I-G)\boldsymbol{y}\right] && (3.8)\\ &= (I-G)(X\boldsymbol{\beta}+\boldsymbol{\mu}) && (3.9)\\ &= (I - X(X^TX)^{-1}X^T)X\,\boldsymbol{\beta} + (I-G)\boldsymbol{\mu} && (3.10)\\ &= (I-G)\boldsymbol{\mu}, && (3.11) \end{aligned}$$

then

$$\mathrm{E}\left[r_i\right] = \mu_i - \sum_{j=i_1}^{i_{m_1}} g_{ij}\mu_j, \tag{3.12}$$

and

$$\mathrm{E}\left[r_i \mid H_i = 0\right] = -\sum_{j=i_1}^{i_{m_1}} g_{ij}\mu_j. \tag{3.13}$$

Thus

$$\mathrm{E}\left[y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{(i)} \mid H_i = 0\right] = -\frac{1}{1-g_i}\sum_{j=i_1, j\neq i}^{i_{m_1}} g_{ij}\mu_j. \tag{3.14}$$

Because

$$\begin{aligned} \mathrm{Var}[\boldsymbol{r}] &= \mathrm{Var}[(I-G)\boldsymbol{y}] && (3.15)\\ &= (I-G)Var[\boldsymbol{y}] && (3.16)\\ &= (I-G)\sigma^2, && (3.17) \end{aligned}$$

then

$$\mathrm{Var}[r_i] = (1-g_i)\sigma^2, \tag{3.18}$$

and hence the variance of $y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{(i)}$ equals

$$\begin{aligned} \mathrm{Var}[y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{(i)}] &= \mathrm{Var}\left[\frac{r_i}{1-g_i}\right] && (3.19)\\ &= \frac{\sigma^2}{1-g_i}. && (3.20) \end{aligned}$$

Therefore, $z_i = \frac{1}{\sigma} r_{(i)}\sqrt{1-g_i}$ has a normal distribution since

$$(z_i \mid H_i = 0) \sim N(\xi_0, 1), \tag{3.21}$$

where $\xi_0 = -\frac{1}{1-g_i}\sum_{j=i_1, j\neq i}^{i_{m_1}} g_{ij}\mu_j \Big/ \frac{\sigma}{\sqrt{1-g_i}} = -\frac{1}{\sigma\sqrt{1-g_i}}\sum_{j=i_1, j\neq i}^{i_{m_1}} g_{ij}\mu_j$. Then we need to calculate the distribution of $s_{(i)}$. If the ith observation is the only candidate for an outlier, then

$(m-k-1)s^2_{(i)}/\sigma^2$ has a central chi-square distribution with degrees of freedom $m-k-1$. However, if we assume there is more than one outlier, then we can show under the above assumptions that $(m-k-1)s^2_{(i)}/\sigma^2$ has a noncentral chi-square distribution with $m-k-1$ degrees of freedom. First note that $(m-k-1)s^2_{(i)}/\sigma^2 = \boldsymbol{y}^T_{(i)}(I-G_{(i)})\boldsymbol{y}_{(i)}/\sigma^2$, where $G_{(i)}$ is the hat matrix with the ith observation being deleted. By (3.7), we have $\boldsymbol{y}_{(i)}/\sigma \sim MVN(X_{(i)}\boldsymbol{\beta}_{(i)} + \boldsymbol{\mu}_{(i)}, I)$. Since $G_{(i)}$ is $(m-1)\times(m-1)$, idempotent and symmetric, by Lemma 1.4.1 we have that

$$\left\{ \frac{1}{\sigma^2}(m-k-1)s^2_{(i)} = \frac{1}{\sigma^2}\boldsymbol{y}^T_{(i)}(I-G_{(i)})\boldsymbol{y}_{(i)} \mid H_i = 0 \right\} \sim \chi^2_{m-k-1}(\eta_0) \tag{3.22}$$

where

$$\begin{aligned}
\eta_0 &= \frac{1}{\sigma^2}\left(X_{(i)}\boldsymbol{\beta}_{(i)}+\boldsymbol{\mu}_{(i)}\right)^T\left(I-G_{(i)}\right)\left(X_{(i)}\boldsymbol{\beta}_{(i)}+\boldsymbol{\mu}_{(i)}\right) && (3.23)\\
&= \frac{1}{\sigma^2}\left(X_{(i)}\boldsymbol{\beta}_{(i)}\right)^T\left(I-G_{(i)}\right)X_{(i)}\boldsymbol{\beta}_{(i)} + \frac{1}{\sigma^2}\left(X_{(i)}\boldsymbol{\beta}_{(i)}\right)^T\left(I-G_{(i)}\right)\boldsymbol{\mu}_{(i)} && (3.24)\\
&+ \frac{1}{\sigma^2}\boldsymbol{\mu}^T_{(i)}\left(I-G_{(i)}\right)X_{(i)}\boldsymbol{\beta}_{(i)} + \frac{1}{\sigma^2}\left(\boldsymbol{\mu}_{(i)}\right)^T\left(I-G_{(i)}\right)\boldsymbol{\mu}_{(i)} && (3.25)\\
&= \frac{1}{\sigma^2}\left(X_{(i)}\boldsymbol{\beta}_{(i)}\right)^T\left(X_{(i)}\boldsymbol{\beta}_{(i)} - X_{(i)}(X^T_{(i)}X_{(i)})^{-1}X^T_{(i)}X_{(i)}\boldsymbol{\beta}_{(i)}\right) && (3.26)\\
&+ \frac{1}{\sigma^2}\left(\boldsymbol{\beta}^T_{(i)}X^T_{(i)} - \boldsymbol{\beta}^T_{(i)}X^T_{(i)}X_{(i)}(X^T_{(i)}X_{(i)})^{-1}X^T_{(i)}\right)\boldsymbol{\mu}_{(i)} && (3.27)\\
&+ \frac{1}{\sigma^2}\boldsymbol{\mu}^T_{(i)}\left(X_{(i)}\boldsymbol{\beta}_{(i)} - X_{(i)}(X^T_{(i)}X_{(i)})^{-1}X^T_{(i)}X_{(i)}\boldsymbol{\beta}_{(i)}\right) + \frac{1}{\sigma^2}\boldsymbol{\mu}^T_{(i)}(I-G_{(i)})\boldsymbol{\mu}_{(i)} && (3.28)\\
&= \frac{1}{\sigma^2}\boldsymbol{\mu}^T_{(i)}(I-G_{(i)})\boldsymbol{\mu}_{(i)} && (3.29)
\end{aligned}$$

and the degrees of freedom is $\text{trace}(I-G_{(i)}) = m-k-1$. Since the covariance matrix of $\boldsymbol{y}$ is still $\sigma^2 I$ and the least square estimate $\hat{\boldsymbol{\beta}}$ has the same form as in Section 1.4.1, then Result 1.4.2 (3) is true for $\hat{\boldsymbol{\beta}}$ and $s^2_{(i)}$. Therefore z_i and $s^2_{(i)}$ are independent, and hence the marginal distribution of $r_i^* = \frac{z_i\sigma^2}{(m-k-1)s^2_{(i)}}$ under the null hypothesis is the doubly noncentral t with the degrees of freedom equal to $m-1-k$ and noncentrality parameters $-\frac{1}{\sigma\sqrt{1-g_i}}\sum_{j=i_1,j\neq i}^{i_{m_1}} g_{ij}\mu_j$ and $\frac{1}{\sigma^2}\boldsymbol{\mu}^T_{(i)}(I-G_{(i)})\boldsymbol{\mu}_{(i)}$.

(b). The alternative hypothesis h_{i1} is true.

In this case,

$$\begin{aligned}
\mathrm{E}\left[r_i \mid H_i = 1\right] &= \mu_i - \left(g_i\mu_i - \sum_{j=i_1,j\neq i}^{i_{m_1}} g_{ij}\mu_j\right) && (3.30)\\
&= (1-g_i)\mu_i - \sum_{j=i_1,j\neq i}^{i_{m_1}} g_{ij}\mu_j, && (3.31)
\end{aligned}$$

and hence the mean of $y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{(i)}$ is

$$\mathrm{E}\left[y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{(i)} \mid H_i = 1\right] = \mu_i - \frac{1}{1-g_i} \sum_{j=i_1, j\neq i}^{i_{m_1}} g_{ij}\mu_j. \tag{3.32}$$

Hence we have

$$(z_i \mid H_i = 1) \sim N(\xi_1, 1) \tag{3.33}$$

where $\xi_1 = \left\{\mu_i - \frac{1}{1-g_i} \sum_{j=i_1, j\neq i}^{i_{m_1}} g_{ij}\mu_j\right\} \Big/ \frac{\sigma}{\sqrt{1-g_i}} = \frac{1}{\sigma}\mu_i\sqrt{1-g_i} - \frac{1}{\sigma\sqrt{1-g_i}} \sum_{j=i_1, j\neq i}^{i_{m_1}} g_{ij}\mu_j$. Similarly,

$$\left\{\frac{1}{\sigma^2}(m-k-1)s_{(i)}^2 \mid H_i = 1\right\} \sim \chi_{m-k-1}^2(\eta_1) \tag{3.34}$$

where $\eta_1 = \frac{1}{\sigma^2}\boldsymbol{\mu}_{(i)}^T(I - G_{(i)})\boldsymbol{\mu}_{(i)}$. Because $H_i = 1$, there are only $m_1 - 1$ elements in $\boldsymbol{H}_{(i)}$ equal to 1. Therefore the marginal distribution of r_i^* under h_{i1} is the doubly noncentral t with the degrees of freedom equal to $m-k-1$ and noncentrality parameters $\frac{1}{\sigma}\mu_i\sqrt{1-g_i} - \frac{1}{\sigma\sqrt{1-g_i}} \sum_{j=i_1, j\neq i}^{i_{m_1}} g_{ij}\mu_j$ and $\frac{1}{\sigma^2}\boldsymbol{\mu}_{(i)}^T(I - G_{(i)})\boldsymbol{\mu}_{(i)}$. ∎

3.3 A Bayesian approach for multiple deletion diagnostics

3.3.1 Model description

In this section, I propose a Bayesian approach for identifying outliers in linear models. Usually the total number of outliers $\sum_{i=1}^m H_i$ is unknown, but I will assume $\boldsymbol{H}$ is random. Let $\boldsymbol{h}$ be a realization of $\boldsymbol{H}$ when $\boldsymbol{H}$ is assumed random.

I start with the description of the Bayesian model. Suppose we observe m observations $\boldsymbol{y}^T = (y_1, \cdots, y_m)$, where y_i arises independently from a normal density, $\phi_{y_i}(y)$ with mean $\boldsymbol{x}_i^T\boldsymbol{\beta} + \sigma\mu_i$ and variance σ^2, where $\boldsymbol{x}_i^T$ is known, but σ^2, $\boldsymbol{\beta}$ and μ_i are unknown. When $\mu_i = 0$, y_i arises from the linear model (1.58) and satisfies the model assumptions in Assumption 1.4.1, while an outlier with nonzero μ_i follows a linear model with a different mean and the same variance as that of the true model. The difference between the mean of the outliers and that of the other observations is assumed to be proportional to the unknown constant σ. The reason I choose the difference to be proportional to σ is because then the distribution parameters of the resulting deletion residuals do not include σ. So I can avoid specifying a prior distribution for σ and hence the posterior probabilities are simpler. My goal is to identify which μ_i are nonzero. The indicators of outliers are

$$H_i = \begin{cases} 0 & \text{if } \mu_i = 0 \\ 1 & \text{if } \mu_i \neq 0 \end{cases}. \tag{3.35}$$

Hence we can rewrite the hypotheses as

$$\begin{array}{l} h_{i,0} : (y_i \mid H_i = 0) \sim N(\mathbf{x}_i^T \beta,\ \sigma^2) \\ h_{i,1} : (y_i \mid H_i = 1) \sim N(\mathbf{x}_i^T \beta + \sigma\mu_i,\ \sigma^2) \end{array} . \tag{3.36}$$

Under the model assumptions, the vector of observations $\boldsymbol{y}$ has a multivariate normal distribution

$$\boldsymbol{y} \sim \text{MVN}(X\boldsymbol{\beta} + \sigma\boldsymbol{\mu}, \sigma^2 I), \tag{3.37}$$

where $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_m)$.

Next I investigate the marginal distribution of the deletion residuals under this model. Using similar arguments as in the proof of Lemma 3.2.1, I can obtain the conditional distribution of r_i^* given the H_i's.

Corollary 3.3.1 *For testing hypotheses (3.36), given $\boldsymbol{\mu}$ and $\boldsymbol{H}_{(i)}$, the conditional distributions of the deletion residual r_i^* under the null and alternative hypotheses are both doubly noncentral t distributions, i.e.*

$$(r_i^* \mid H_i = 0, \boldsymbol{H}_{(i)}, \boldsymbol{\mu}_{(i)}) \sim t''_{m-k-1}(\xi_0,\ \eta) \tag{3.38}$$

and

$$(r_i^* \mid H_i = 0, \boldsymbol{H}_{(i)}, \boldsymbol{\mu}) \sim t''_{m-k-1}(\xi_1,\ \eta), \tag{3.39}$$

where the noncentrality parameter of the denominator in r_i^* is

$$\eta = \boldsymbol{\mu}_{(i)}^T (I - G_{(i)}) \boldsymbol{\mu}_{(i)}, \tag{3.40}$$

and the noncentrality parameters of the numerator in r_i^* for the null and alternative are respectively

$$\xi_0 = -\frac{1}{\sqrt{1 - g_i}} \sum_{j=1, j\neq i}^{m} g_{ij}\mu_j \tag{3.41}$$

and

$$\xi_1 = \sqrt{1 - g_i}\mu_i - \frac{1}{\sqrt{1 - g_i}} \sum_{j=1, j\neq i}^{m} g_{ij}\mu_j. \tag{3.42}$$

Here $\boldsymbol{\mu}_{(i)}$ denotes the vector $\boldsymbol{\mu}$ with the ith observation being deleted. When $\eta = 0$, the distribution of r_i^* becomes singly noncentral t, while when both $\xi_i = 0$, $i = 0, 1$ and $\eta = 0$, r_i^* has the central t distribution.

Proof.

By a similar argument to that in the proof of Theorem 3.2.1, the mean of $y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{(i)}$ under h_{i0} is

$$\mathrm{E}\left[y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{(i)} \mid H_i = 0\right] = -\frac{\sigma}{(1-g_i)} \sum_{j=i_1, j\neq i}^{i_{m_1}} g_{ij}\mu_j, \tag{3.43}$$

and under h_{i1} is

$$\mathrm{E}\left[y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{(i)} \mid H_i = 1\right] = \sigma\mu_i - \frac{\sigma}{1-g_i} \sum_{j=1, j\neq i}^{m} g_{ij}\mu_j. \tag{3.44}$$

Therefore the noncentrality parameter of the numerator in r_i^* under h_{i0} is

$$\xi_0 = -\frac{\sigma}{(1-g_i)} \sum_{j=i_1, j\neq i}^{i_{m_1}} g_{ij}\mu_j \Big/ \frac{\sigma}{\sqrt{1-g_i}} \tag{3.45}$$

$$= -\frac{1}{\sqrt{1-g_i}} \sum_{j=1, j\neq i}^{m} g_{ij}\mu_j, \tag{3.46}$$

and under h_{i1} is

$$\xi_1 = \left[\sigma\mu_i - \frac{\sigma}{(1-g_i)} \sum_{j=i_1, j\neq i}^{i_{m_1}} g_{ij}\mu_j\right] \Big/ \frac{\sigma}{\sqrt{1-g_i}} \tag{3.47}$$

$$= \sqrt{1-g_i}\mu_i - \frac{1}{\sqrt{1-g_i}} \sum_{j=1, j\neq i}^{m} g_{ij}\mu_j. \tag{3.48}$$

The noncentrality parameter of the denominator in r_i^* under h_{i0} and that under h_{i1} are the same and have the form:

$$\eta = \left(X_{(i)}\boldsymbol{\beta}_{(i)} + \sigma\boldsymbol{\mu}_{(i)}\right)^T \left(I - G_{(i)}\right) \left(X_{(i)}\boldsymbol{\beta}_{(i)} + \sigma\boldsymbol{\mu}_{(i)}\right) \tag{3.49}$$

$$= \boldsymbol{\mu}_{(i)}^T (I - G_{(i)})\boldsymbol{\mu}_{(i)}. \tag{3.50}$$

■

Note that the conditional distributions of the square of the deletion residual r_i^* under the null and alternative hypotheses are both doubly noncentral F with the numerator degrees of freedom 1, the denominator degrees of freedom $m-k-1$, the denominator noncentrality parameter η, and the numerator noncentrality parameters $\zeta_0 = \xi_0^2$ and $\zeta_1 = \xi_1^2$, respectively. That is,

$$(r_i^{*2} \mid H_i = 0, \boldsymbol{H}_{(i)}, \boldsymbol{\mu}_{(i)}) \sim F''_{1,m-k-1}(\zeta_0,\ \eta) \tag{3.51}$$

and

$$(r_i^{*2} \mid H_i = 1, \boldsymbol{H}_{(i)}, \boldsymbol{\mu}) \sim F''_{1,m-k-1}(\zeta_1,\ \eta), \tag{3.52}$$

where $F''_{\nu_1,\nu_2}(\zeta,\ \eta)$ denotes a random variable following the doubly noncentral F distribution with degrees of freedom ν_1 and ν_2 and the noncentrality parameter ζ and η. When either $\zeta_i = 0$, $i = 0, 1$ or $\eta = 0$, the distribution of r_i^{*2} becomes singly noncentral F, while for both $\zeta_i = 0$, $i = 0, 1$ and $\eta = 0$, r_i^{*2} has the central F distribution.

Next I need to choose appropriate priors for the parameters of the distribution of r_i^{*2}. A common assumption for the indicators H_i is that they are i.i.d. random Bernoulli variables with probability π_1 [29, 88, 89, 90]. As mentioned before, the proportion of outliers is smaller than or equal to $m/2$, otherwise the assumed model is not proper for that dataset. Let $\pi_0 = 1- \pi_1$ and assume π_0 follows a Beta distribution Beta(a, b) with hyper-parameters a and b, that is

$$\pi_0 \sim \text{Beta}(a, b). \tag{3.53}$$

Scott and Berger [88] used a prior of Beta$(a, 1)$ and calculated a by specifying a prior median. However, the choice of $b = 1$ makes the Beta prior have the mode equal to 1 and puts a lot of mass near 1, which implies that there is no outlier at all. Usually we have knowledge of the existence of some outliers before we start a test for outliers, so we are not interested in the case that there is no outlier. When we think there will be some outliers, it might be more reasonable to choose a and b such that the mean of the Beta distribution is close to one and the density of the Beta distribution is maximized around a value close to 1, for example Beta$(8, 2)$. Beta$(8, 2)$ has mean 0.8 and variance 0.0145. This prior distribution is appropriate if one believes the mean of the proportion of outliers is around 0.2. The hyper-parameters of Beta distribution can be chosen according to the prior information about the mean and variance of π_0. For example, if one strongly believes the mean of π_0 is about 0.2, then a stronger prior distribution Beta$(80, 20)$ can be used, whereas one may choose a weaker prior distribution Beta$(0.8, 0.2)$ if the belief of $\text{E}(\pi_0) = 0.2$ is weak. The prior on $\boldsymbol{\mu}$ is also unknown, and it is convenient to assume those nonzero μ_i are i.i.d. normal with density $\phi_{\mu_i}(u)$, with mean 0 and known variance V, *i.e.*

$$(\mu_i \mid H_i = 1) \sim N(0, V). \tag{3.54}$$

In Section 3.5, the sensitivity of the marginal posterior distribution of H_i to the Beta prior of π_0 and the normal prior of μ_i is studied.

3.3.2 Posterior distributions

As discussed in the previous section, I am interested in identifying nonzero μ_i, which can be measured by the marginal posterior probability $\Pr(H_i = 1 \mid r_1^{*2}, \cdots r_m^{*2})$. The calculation of this marginal posterior probability involves the computation of the joint distribution of all deletion residual squares, which is difficult because of the complex dependence structure of $r_1^{*2}, \cdots, r_m^{*2}$. One may think that whether the ith null hypothesis is true or not depends more on its own deletion residual square than the others, and hence there might not be a large difference between the marginal posterior probability of $\{H_i = 1\}$ given all deletion residuals and that given only the ith deletion residual. The marginal posterior probability $\Pr(H_i = 1 \mid r_i^{*2})$ depends only on the distribution of the ith deletion residual, which is given in Theorem 3.3.1. This marginal posterior can be calculated from the joint posterior distribution of $\boldsymbol{\theta} = (\boldsymbol{H}, \boldsymbol{\mu}, \pi_0)$, $\boldsymbol{\theta} \in \Theta$. Let $f_{r_i^{*2}}(r^2|\boldsymbol{H}, \boldsymbol{\mu}, \pi_0)$ and $f_{r_i^{*2}}(r^2)$ denote respectively the doubly noncentral F density function of r_i^{*2} given parameters $\boldsymbol{\theta}$ and the marginal density of r_i^{*2}. Let $p_{\pi_0}(\pi)$, $\phi_{\mu_i}(u)$ and $p_{\boldsymbol{H}}(\boldsymbol{\omega} \mid \pi_0)$ denote respectively the prior densities of continuous π_0 and μ_i and discrete $\boldsymbol{H}$, and $p_{(\boldsymbol{H},\boldsymbol{\mu},\pi_0)}(\boldsymbol{\omega}, \boldsymbol{u}, \pi)$ denote the joint density of $\boldsymbol{H}, \boldsymbol{\mu}, \pi_0$. Under the model assumptions given in Section 3.3.1, the joint posterior density for $\boldsymbol{\theta}$ is

$$p_{(\boldsymbol{H},\boldsymbol{\mu},\pi_0)}(\boldsymbol{\omega}, \boldsymbol{u}, \pi | r_i^{*2}) = \frac{f_{r_i^{*2}}(r^2|\boldsymbol{H}, \boldsymbol{\mu}, \pi_0) \cdot \prod_{j=1}^{m} \phi_{\mu_j}(u) \cdot p_{\boldsymbol{H}}(\boldsymbol{\omega} \mid \pi_0) \cdot p_{\pi_0}(\pi)}{f_{r_i^{*2}}(r^2)}, \tag{3.55}$$

where $p_{\boldsymbol{H}}(\boldsymbol{\omega} \mid \pi_0) = \prod_{j=1}^{m} \pi_0^{(1-\omega_j)}(1-\pi_0)^{\omega j}$. The marginal density of r_i^{*2} can be calculated as

$$f_{r_i^{*2}}(r^2) = \Pr(H_i = 0) f_{r_i^{*2}}(r^2|H_i = 0) + \Pr(H_i = 1) f_{r_i^{*2}}(r^2|H_i = 1) \tag{3.56}$$

$$= \int_{-\infty}^{\infty} \cdots \int_{0}^{1} \sum_{\boldsymbol{\omega}_{(i)} \in \{0,1\}^{m-1}} \{\Pr(H_i = 0|\pi_0 = \pi) \tag{3.57}$$

$$f_{r_i^{*2}}(r^2|H_i = 0, \boldsymbol{H}_{(i)} = \boldsymbol{\omega}_{(i)}, \boldsymbol{\mu} = \boldsymbol{u}, \pi_0 = \pi) +$$

$$[1 - \Pr(H_i = 0|\pi_0 = \pi)] f_{r_i^{*2}}(r^2|H_i = 1, \boldsymbol{H}_{(i)} = \boldsymbol{\omega}_{(i)}, \boldsymbol{\mu} = \boldsymbol{u}, \pi_0 = \pi)\}$$

$$p_{\boldsymbol{H}_{(i)}}(\boldsymbol{\omega}_{(i)} \mid \pi_0) p_{\pi_0}(\pi) \prod_{j=1}^{m} \phi_{\mu_j}(u) d\pi d\boldsymbol{u}$$

$$= \int_{-\infty}^{\infty} \cdots \int_{0}^{1} \sum_{\boldsymbol{\omega}_{(i)} \in \{0,1\}^{m-1}} \{\pi f_{r_i^{*2}}(r^2|H_i = 0, \boldsymbol{H}_{(i)} = \boldsymbol{\omega}_{(i)}, \boldsymbol{\mu} = \boldsymbol{u}, \pi_0 = \pi) + \tag{3.58}$$

$$
\begin{aligned}
&(1-\pi) f_{r_i^{*2}}(r^2 | H_i = 1, \boldsymbol{H}_{(i)} = \boldsymbol{\omega}_{(i)}, \boldsymbol{\mu} = \boldsymbol{u}, \pi_0 = \pi)\} \\
&p_{\boldsymbol{H}_{(i)}}(\boldsymbol{\omega}_{(i)} |\ \pi_0) p_{\pi_0}(\pi) \prod_{j=1}^{m} \phi_{\mu_j}(u) d\pi d\boldsymbol{u} \\
&= \int_{-\infty}^{\infty} \cdots \int \int_{0}^{1} \sum_{\boldsymbol{\omega}_{(i)} \in \{0,1\}^{m-1}} \pi f_{r_i^{*2}}(r^2 | H_i = 0, \boldsymbol{H}_{(i)} = \boldsymbol{\omega}_{(i)}, \boldsymbol{\mu} = \boldsymbol{u}, \pi_0 = \pi) \qquad (3.59) \\
&p_{\boldsymbol{H}_{(i)}}(\boldsymbol{\omega}_{(i)} |\ \pi_0) dF_{\pi_0}(\pi) dF_{\boldsymbol{\mu}_{(i)}}(\boldsymbol{u}_{(i)}) \\
&+ \int_{-\infty}^{\infty} \cdots \int \int_{0}^{1} \textstyle\sum_{\boldsymbol{\omega}_{(i)} \in \{0,1\}^{m-1}} (1-\pi) f_{r_i^{*2}}(r^2 | H_i = 1, \boldsymbol{H}_{(i)} = \boldsymbol{\omega}_{(i)}, \boldsymbol{\mu} = \boldsymbol{u}, \pi_0 = \pi) \\
&p_{\boldsymbol{H}_{(i)}}(\boldsymbol{\omega}_{(i)} |\ \pi_0) dF_{\pi_0}(\pi) dF_{\boldsymbol{\mu}}(\boldsymbol{u}),
\end{aligned}
$$

where $\int_{-\infty}^{\infty} \cdots \int h(\boldsymbol{u}) d\boldsymbol{u}$ denotes a m-dimensional integral for $h(\boldsymbol{u})$ with respect to the vector of variables $\boldsymbol{u}$, and $F_{\pi_0}(\pi)$ and $F_{\boldsymbol{\mu}}(\boldsymbol{u})$ denote respectively the distribution functions for π_0 and the joint distribution function for $\boldsymbol{\mu}$.

Lemma 3.3.1 *The marginal posterior probability that the ith observation is an outlier given its deletion residual is*

$$
\Pr(H_i = 0 \mid r_i^{*2} = r^2) = \frac{\mathrm{E}_{(\boldsymbol{H}_{(i)}, \boldsymbol{\mu}, \pi_0)}[h_1(r^2, \boldsymbol{\omega}_{(i)}, \boldsymbol{u}, \pi)]}{\mathrm{E}_{(\boldsymbol{H}_{(i)}, \boldsymbol{\mu}, \pi_0)}[h_1(r^2, \boldsymbol{\omega}_{(i)}, \boldsymbol{u}, \pi)] + \mathrm{E}_{(\boldsymbol{H}_{(i)}, \boldsymbol{\mu}, \pi_0)}[h_2(r^2, \boldsymbol{\omega}_{(i)}, \boldsymbol{u}, \pi)]}, \qquad (3.60)
$$

where $\mathrm{E}_{(\boldsymbol{H}_{(i)}, \boldsymbol{\mu}, \pi_0)}[\cdot]$ *represents the expectation with respect to the joint prior distribution of* $\boldsymbol{H}_{(i)}$, $\boldsymbol{\mu}$ *and* π_0, *and* $h_1(r^2, \boldsymbol{\omega}_{(i)}, \boldsymbol{u}, \pi) = \pi f_{F''_{1,m-k-1}(\zeta_0,\ \eta)}(r^2)$ *and* $h_2(r^2, \boldsymbol{\omega}_{(i)}, \boldsymbol{u}, \pi) = (1-\pi) f_{F''_{1,m-k-1}(\zeta_1,\ \eta)}(r^2)$.

Proof.

$$
\begin{aligned}
&\Pr(H_i = 0 \mid r_i^{*2} = r^2) \\
&\quad = \int_{-\infty}^{\infty} \cdots \int \int_{0}^{1} \sum_{\boldsymbol{\omega}_{(i)} \in \{0,1\}^{m-1}} \Pr(H_i = 0 | \pi_0 = \pi) \qquad (3.61) \\
&\quad f_{r_i^{*2}}(r^2\ | H_i = 0,\ \boldsymbol{H}_{(i)} = \boldsymbol{\omega}_{(i)}, \boldsymbol{\mu}_{(i)} = \boldsymbol{u}_{(i)}, \pi_0 = \pi) p_{(\boldsymbol{H}, \boldsymbol{\mu}, \pi_0)}(\boldsymbol{\omega}, \boldsymbol{u}, \pi | r_i^{*2}) d\pi d\boldsymbol{u} \qquad (3.62) \\
&\quad \int_{-\infty}^{\infty} \cdots \int \int_{0}^{1} \sum_{\boldsymbol{\omega}_{(i)} \in \{0,1\}^{m-1}} \pi f_{r_i^{*2}}(r^2\ | H_i = 0,\ \boldsymbol{H}_{(i)} = \boldsymbol{\omega}_{(i)}, \boldsymbol{\mu}_{(i)} = \boldsymbol{u}_{(i)}, \pi_0 = \pi) \\
&= \frac{p_{(\boldsymbol{H}, \boldsymbol{\mu}, \pi_0)}(\boldsymbol{\omega}, \boldsymbol{u}, \pi) d\pi d\boldsymbol{u}}{f_{r_i^{*2}}(r^2)} \qquad (3.63)
\end{aligned}
$$

$$= \frac{\int_{-\infty}^{\infty} \cdots \int \int_{0}^{1} \sum_{\boldsymbol{\omega}_{(i)}} \pi f_{F''_{1,m-k-1}(\zeta_0,\,\eta)}(r^2) p_{(\boldsymbol{H},\boldsymbol{\mu},\pi_0)}(\boldsymbol{\omega},\boldsymbol{u},\pi) d\pi d\boldsymbol{u}}{\int_{-\infty}^{\infty} \cdots \int \int_{0}^{1} \sum_{\boldsymbol{\omega}_{(i)}} \left\{ \begin{array}{c} \pi f_{F''_{1,m-k-1}(\zeta_0,\,\eta)}(r^2) + \\ (1-\pi) f_{F''_{1,m-k-1}(\zeta_1,\,\eta)}(r^2) \end{array} \right\} p_{(\boldsymbol{H},\boldsymbol{\mu},\pi_0)}(\boldsymbol{\omega},\boldsymbol{u},\pi) d\pi d\boldsymbol{u}} \tag{3.64}$$

$$= \frac{\int \cdots \int \sum_{\boldsymbol{\omega}_{(i)}} h_1(r^2, \boldsymbol{\omega}_{(i)}, \boldsymbol{u}, \pi) p_{(\boldsymbol{H},\boldsymbol{\mu},\pi_0)}(\boldsymbol{\omega},\boldsymbol{u},\pi) d\pi d\boldsymbol{u}}{\int_{-\infty}^{\infty} \cdots \int \int_{0}^{1} \sum_{\boldsymbol{\omega}_{(i)}} \left[h_1(r^2, \boldsymbol{\omega}_{(i)}, \boldsymbol{u}, \pi) + h_2(r^2, \boldsymbol{\omega}_{(i)}, \boldsymbol{u}, \pi) \right] p_{(\boldsymbol{H},\boldsymbol{\mu},\pi_0)}(\boldsymbol{\omega},\boldsymbol{u},\pi) d\pi d\boldsymbol{u}} \tag{3.65}$$

$$= \frac{E_{(\boldsymbol{H}_{(i)},\boldsymbol{\mu},\pi_0)}[h_1(r^2, \boldsymbol{\omega}_{(i)}, \boldsymbol{u}, \pi)]}{E_{(\boldsymbol{H}_{(i)},\boldsymbol{\mu},\pi_0)}[h_1(r^2, \boldsymbol{\omega}_{(i)}, \boldsymbol{u}, \pi)] + E_{(\boldsymbol{H}_{(i)},\boldsymbol{\mu},\pi_0)}[h_2(r^2, \boldsymbol{\omega}_{(i)}, \boldsymbol{u}, \pi)]}. \tag{3.66}$$

■

3.3.3 Computational implementation

The computation of the marginal posterior probability $\Pr(H_i = 0 \mid r_i^{*2})$ involves evaluating high dimensional integrals, which can be implemented numerically. Importance sampling can be used here to calculate this posterior probability. *Importance sampling* is a Monte Carlo method to calculate high dimensional integrals, which is introduced in many books, for example [12, 36, 75]. Let $\boldsymbol{\theta} = (\boldsymbol{H}, \boldsymbol{\mu}, \pi_0)$. Suppose we can generate a sequence of i.i.d. random variables $\{\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \cdots \boldsymbol{\theta}^n\}$ from a density $g(\boldsymbol{\theta}) > 0$ on the parameter space Θ. Then

$$\int \cdots \int h(\boldsymbol{\theta}) f(r_i^{*2}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} = \mathrm{E}_g \left[\frac{h(\boldsymbol{\theta}) f(r_i^{*2}|\boldsymbol{\theta}) p(\boldsymbol{\theta})}{g(\boldsymbol{\theta})} \right], \tag{3.67}$$

and by the strong law of large numbers,

$$\int \cdots \int h(\boldsymbol{\theta}) f(r_i^{*2}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} = \lim_{\boldsymbol{n}\to\boldsymbol{\infty}} \frac{1}{n} \sum_{j=1}^{n} \left[\frac{h(\boldsymbol{\theta}^j) f(r_i^{*2}|\boldsymbol{\theta}^j) p(\boldsymbol{\theta}^j)}{g(\boldsymbol{\theta}^j)} \right]. \tag{3.68}$$

Let $E_{\boldsymbol{\theta}|r_i^{*2}}[\cdot]$ denote the expectation with respect to the posterior distribution of $\boldsymbol{\theta} = (\boldsymbol{H}, \boldsymbol{\mu}, \pi_0)$. Hence $E_{\boldsymbol{\theta}|r_i^{*2}}[h(\boldsymbol{\theta})]$ can be approximated by

$$\mathrm{E}_{\boldsymbol{\theta}|r_i^{*2}}[h(\boldsymbol{\theta})] = \frac{\int \cdots \int h(\boldsymbol{\theta}) f(r_i^{*2}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}}{f(r_i^{*2})} \tag{3.69}$$

$$= \frac{\int \cdots \int h(\boldsymbol{\theta}) f(r_i^{*2}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int \cdots \int f(r_i^{*2}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}} \tag{3.70}$$

$$\simeq \frac{\sum_{j=1}^{n} h(\boldsymbol{\theta}^j) w(\boldsymbol{\theta}^j)}{\sum_{j=1}^{n} w(\boldsymbol{\theta}^j)}, \tag{3.71}$$

where $w(\boldsymbol{\theta}^j) = f(r_i^{*2}|\boldsymbol{\theta}^j)p(\boldsymbol{\theta}^j)/g(\boldsymbol{\theta}^j)$ are called *importance ratios* or *importance weights* [37] and $g(\boldsymbol{\theta}^j)$ is called the *importance function* [12]. The choice of an importance function is the crucial step in the importance sampling method as discussed in Berger [12]. On the one hand, the importance function g is desirable to be chosen so that the approximation in (3.71) is accurate for as small a number of random samples n as possible. On the other hand, we want to choose g so that the generation of the random samples from g is not time-consuming. The two goals are usually against each other and finding a balance between them is not easy. One suggestion given by Berger [12] is to choose the importance function equal to the prior density, so that $w(\boldsymbol{\theta}^j) = f(r_i^{*2}|\boldsymbol{\theta}^j)$, and the calculation of $p(\boldsymbol{\theta}^j)$ is avoided. Since all priors are proper, sampling from the joint prior density is easy. There may be problems to choose the joint prior density as the importance function. Importance sampling works poorly if the importance ratios are small with high probabilities but are very large with a low probability [37]. If we choose $g = p$, then $w(\boldsymbol{\theta}^j) = f(r_i^{*2}|\boldsymbol{\theta}^j)$. The estimates would be poor if the likelihood is strongly informative, that is some $f(r_i^{*2}|\boldsymbol{\theta}^j)$ are much larger than the others. Gelman *et al.* [37] suggest to examine the distribution of sampled importance ratios to detect possible problems. It is helpful to plot a histogram of the logarithms of the large importance ratios. Such histograms were examined for some situations and exceedingly large importance ratios were not observed.

The joint prior distribution is

$$p(\boldsymbol{H}_{(i)}, \boldsymbol{\mu}, \pi_0) = p(\boldsymbol{\mu}|\boldsymbol{H}_{(i)})p(\boldsymbol{H}_{(i)}|\pi_0)p(\pi_0). \tag{3.72}$$

Thus we can draw n samples $(\boldsymbol{H}_{(i)}^j, \boldsymbol{\mu}^j, \pi_0^j)$, $j = 1, \cdots, n$, from $p(\boldsymbol{H}_{(i)}, \boldsymbol{\mu}, \pi_0)$, and then approximate the marginal posterior probability $\Pr(H_i = 0 \mid r_i^*)$ as

$$\begin{aligned}
&\frac{E_{(\boldsymbol{H}_{(i)}, \boldsymbol{\mu}, \pi_0)}[h_1(r_i^{*2}, \boldsymbol{H}_{(i)}, \boldsymbol{\mu}, \pi_0)]}{E_{(\boldsymbol{H}_{(i)}, \boldsymbol{\mu}, \pi_0)}[h_1(r_i^{*2}, \boldsymbol{H}_{(i)}, \boldsymbol{\mu}, \pi_0)] + E_{(\boldsymbol{H}_{(i)}, \boldsymbol{\mu}, \pi_0)}[h_2(r_i^{*2}, \boldsymbol{H}_{(i)}, \boldsymbol{\mu}, \pi_0)]} \\
&\simeq \frac{\sum_{j=1}^n h_1(r_i^{*2}, \boldsymbol{H}_{(i)}^j, \boldsymbol{\mu}^j, \pi_0^j)}{\sum_{j=1}^n h_1(r_i^{*2}, \boldsymbol{H}_{(i)}^j, \boldsymbol{\mu}^j, \pi_0^j) + \sum_{j=1}^n h_2(r_i^{*2}, \boldsymbol{H}_{(i)}^j, \boldsymbol{\mu}^j, \pi_0^j)} \qquad (3.73) \\
&= \frac{1}{1 + \frac{\sum_{j=1}^n h_2(r_i^{*2}, \boldsymbol{H}_{(i)}^j, \boldsymbol{\mu}^j, \pi_0^j)}{\sum_{j=1}^n h_1(r_i^{*2}, \boldsymbol{H}_{(i)}^j, \boldsymbol{\mu}^j, \pi_0^j)}} \qquad (3.74) \\
&= \frac{1}{1 + \frac{\sum_{j=1}^n (1-\pi_0^j) f_{F''_{1,m-k-1}(\zeta_{1i}^j,\ \eta)}(r_i^{*2})}{\sum_{j=1}^n \pi_0^j f_{F''_{1,m-k-1}(\zeta_{0i}^j,\ \eta)}(r_i^{*2})}}. \qquad (3.75)
\end{aligned}$$

The algorithm for computing all posterior probabilities $\Pr(H_i = 0 \mid r_i^{*2}),\ i = 1, \cdots, m$, is given as follows.

Algorithm 3.3.1 *1. For a given set of observations $(y_1, x_1), (y_2, x_2), \cdots, (y_m, x_m)$, calculate the hat matrix G and the deletion residuals $r_1^*, r_2^*, \cdots, r_m^*$.*

2. ***For*** *j from 1 to n:*

2.1 Generate a π_0^j from Beta(a, b)*;*

2.2 Generate $\boldsymbol{H}_{(i)}^j = (H_1^1, H_2^1, \cdots H_{i-1}, H_{i+1}, \cdots H_m)$ from Binomial$(m-1,\ \pi_0^j)$*;*

2.3 ***If*** *$H_s^j = 1,\ s = 1, \cdots, i-1, i+1, \cdots m,$*

then *generate μ_s^j from $N(0, V)$.*

3. ***For*** *i from 1 to m:*

3.1 ***For*** *j from 1 to n:*

3.1.1 Generate μ_i^j from $N(0, V)$;

3.1.2 Calculate the noncentral parameters $\zeta_{1i}^j = \left(\sqrt{1-g_i}\mu_i - \frac{1}{\sqrt{1-g_i}} \sum_{j=1, j\neq i}^{m} g_{ij}\mu_j\right)^2$ and $\zeta_{0i}^j = \left(\frac{1}{\sqrt{1-g_i}} \sum_{j=1, j\neq i}^{m} g_{ij}\mu_j\right)^2$, where g_{ij} is the ijth element of G, and $\eta = \boldsymbol{\mu}_{(i)}^T (I - G_{(i)})\boldsymbol{\mu}_{(i)}$ with $G_{(i)} = X_{(i)}(X^TX)^{-1}\left\{I + \frac{1}{(1-g_i)}\boldsymbol{x}_i\boldsymbol{x}_i^T(X^TX)^{-1}\right\}X_{(i)}^T$;

3.1.3 Calculate $f_{F''_{1,m-k-1}(\zeta_{0i}^1,\ \eta)}(r_i^{*2})$ and $f_{F''_{1,m-k-1}(\zeta_{1i}^1,\ \eta)}(r_i^{*2})$.

3.2 Calculate $\Pr(H_i = 0 \mid r_i^{*2}) = 1 \Big/ \left(1 + \frac{\sum_{j=1}^n (1-\pi_0^j) f_{F''_{1,m-k-1}(\zeta_{1i}^j,\ \eta)}(r_i^{*2})}{\sum_{j=1}^n \pi_0^j f_{F''_{1,m-k-1}(\zeta_{0i}^j,\ \eta)}(r_i^{*2})}\right)$.

3.4 Computing doubly noncentral F density

Algorithm 3.3.1 in Section 3.3.3 involves calculating the density of a random variable that follows the doubly noncentral F distribution with ν_1, ν_2 degrees of freedom and noncentrality parameters ζ and η, which is denoted by $F''_{\nu_1,\nu_2}(\zeta, \eta)$. The density of $F''_{\nu_1,\nu_2}(\zeta, \eta)$ can be expressed as a doubly infinite series. I have not encountered a function which can calculate the density of $F''_{\nu_1,\nu_2}(\zeta, \eta)$ in any package in "R" [71] so far, so I have written my own function to do so. In order to calculate the density of $F''_{\nu_1,\nu_2}(\zeta, \eta)$, I need to evaluate

the double sum until convergence. However, the computation is burdensome, especially when the sample size n in Algorithm 3.3.1 is very large. Therefore I used an approximation proposed by Patnaik [68] in my simulation study for computing the density of $F''_{\nu_1,\nu_2}(\zeta,\eta)$. Patnaik [68] approximated the noncentral χ^2 distribution by a scaled central χ^2 distribution with degrees of freedom related to the noncentrality parameter. Since the doubly noncentral F random variable can be written as the ratio of two noncentral χ^2 random variables, Patnaik's method can also be used to approximate the doubly noncentral F distribution. In this section, I introduce this approximation and compare the approximated values with the actual values, which are computed by an algorithm I propose in this section. The expressions for the densities of noncentral χ^2, noncentral t and noncentral F distribution can be found in Johnson and Kotz [55].

First note that the doubly noncentral F distribution of $F''_{\nu_1,\nu_2}(\zeta,\eta)$ is defined as the distribution of the ratio of two noncentral χ^2 variables, that is

$$F''_{\nu_1,\nu_2}(\zeta,\eta) = \frac{\chi'^2_{\nu_1}(\zeta)/\nu_1}{\chi'^2_{\nu_2}(\eta)/\nu_2}. \tag{3.76}$$

The noncentral χ^2 random variable was defined in (3.5) in Section 3.2. The distribution function of the noncentral $\chi'^2_{\nu_1}(\zeta)$ can be expressed as a weighted infinite sum of central χ^2 distributions with weights equal to a Poisson mass having mean $\zeta/2$, *i.e.*

$$\Pr(\chi'^2_{\nu_1}(\zeta) \le x) = \sum_{i=0}^{\infty} \left[\frac{(\frac{1}{2}\zeta)^i}{i!} e^{-\frac{1}{2}\zeta} \right] \Pr(\chi^2_{\nu_1+2i} \le x). \tag{3.77}$$

The simplest approximation to the distribution of $\chi'^2_{v_1}(\zeta)$ is a scaled central χ^2 distribution of $c\chi^2_\nu$. Patnaik [68] suggested to choose c and ν such that the distribution of $\chi'^2_{\nu_1}(\zeta)$ and the approximated distribution of $c\chi^2_\nu$ have the same first and second moments, and he obtained

$$c = \frac{\nu_1 + 2\zeta}{\nu_1 + \zeta},\ \nu = \frac{(\nu_1+\zeta)^2}{\nu_1+2\zeta}. \tag{3.78}$$

It can be shown that, for fixed x and ν, the maximum error of Patnaik's approximation to the CDF of $\chi'^2_{v_1}(\zeta)$, *i.e.* the maximum difference between the CDF and Patnaik's approximation, is $O(\zeta^2)$ as $\zeta \to 0$ and $O(\zeta^{-\frac{1}{2}})$ as $\zeta \to \infty$ [55].

Next consider the singly noncentral F distribution where the denominator in (3.76) has a central χ^2 distribution. The singly noncentral F random variable, indicated by $F'_{\nu_1,\nu_2}(\zeta)$, has $\eta = 0$ but $\zeta > 0$. If instead, $\zeta = 0$ but $\eta > 0$, then the doubly noncentral F variable is

the reciprocal of a singly noncentral F variable

$$F''_{\nu_1,\nu_2}(0,\eta) = 1 \ / \ F'_{\nu_1,\nu_2}(\eta), \tag{3.79}$$

and the density function of $F''_{\nu_1,\nu_2}(0,\eta)$ is

$$f_{F''_{\nu_1,\nu_2}(0,\eta)}(x) = \frac{1}{x^2} f_{F'_{\nu_1,\nu_2}(\eta)}\left(\frac{1}{x}\right). \tag{3.80}$$

Although the "R" function "df( )" in package "stats" [71] can be used to compute the density of $F'_{\nu_1,\nu_2}(\zeta)$, in order to compute the density of $F''_{\nu_1,\nu_2}(\zeta,\eta)$, I firstly write an algorithm to calculate the density of $F'_{\nu_1,\nu_2}(\zeta)$ in order to develop a method to compute the more complex density of $F''_{\nu_1,\nu_2}(\zeta,\eta)$. The density of $F'_{\nu_1,\nu_2}(\zeta)$ given in [54] is

$$f_{F'_{\nu_1,\nu_2}(\zeta)}(x) = \sum_{j=0}^{\infty}\left(\frac{(\frac{1}{2}\zeta)^j}{j!}e^{-\frac{1}{2}\zeta}\right)\frac{1}{B(\frac{\nu_1}{2}+j,\frac{\nu_2}{2})}\left(\frac{\nu_1 x}{\nu_2+\nu_1 x}\right)^{\frac{\nu_1}{2}+j}\left(\frac{\nu_2}{\nu_2+\nu_1 x}\right)^{\frac{\nu_2}{2}} x^{-1} \tag{3.81}$$

$$\equiv \sum_{j=0}^{\infty} S_j \tag{3.82}$$

where $B(\frac{\nu_1}{2}+j,\frac{\nu_2}{2})$ is the Beta function and S_j is defined as the jth term in the sum in (3.81).

Lemma 3.4.1 *There exists a j^* such that $R_{j^*} = \frac{S_{j^*+1}}{S_{j^*}} < 1$, and $f_{F'_{\nu_1,\nu_2}(\zeta)}(x)$ is bounded by $\left(\sum_{j=0}^{j^*} S_j\right) + \frac{S_{j^*}}{1-R_{j^*}} = \left(\sum_{j=0}^{j^*} S_j\right) + \frac{S_{j^*}^2}{S_{j^*}^* - S_{j^*+1}}$.*

Proof.

It is shown next that the term S_j in (3.82) either decreases in j or increases to a maximum and then decreases in j by showing that the ratio of any two consecutive terms decreases in j. The ratio of S_{j+1} to S_j is

$$R_j = \frac{S_{j+1}}{S_j} \tag{3.83}$$

$$= \frac{\frac{1}{2}\eta}{j+1}\frac{B(\frac{\nu_1}{2}+j,\frac{\nu_2}{2})}{B(\frac{\nu_1}{2}+j+1,\frac{\nu_2}{2})}\frac{\nu_1 x}{\nu_2+\nu_1 x} \tag{3.84}$$

$$= \frac{\eta\nu_1 x}{2(j+1)(\nu_2+\nu_1 x)}\frac{\Gamma(\frac{\nu_1}{2}+j)\Gamma(\frac{\nu_2}{2})}{\Gamma(\frac{\nu_1}{2}+\frac{\nu_2}{2}+j)}\frac{\Gamma(\frac{\nu_1}{2}+\frac{\nu_2}{2}+j+1)}{\Gamma(\frac{\nu_1}{2}+j+1)\Gamma(\frac{\nu_2}{2})} \tag{3.85}$$

$$= \frac{\eta\nu_1 x}{2(j+1)(\nu_2+\nu_1 x)}\frac{\Gamma(\frac{\nu_1}{2}+j)}{\Gamma(\frac{\nu_1}{2}+\frac{\nu_2}{2}+j)}\frac{\Gamma(\frac{\nu_1}{2}+\frac{\nu_2}{2}+j)(\frac{\nu_1}{2}+\frac{\nu_2}{2}+j)}{\Gamma(\frac{\nu_1}{2}+j)(\frac{\nu_1}{2}+j)} \tag{3.86}$$

$$= \frac{\eta\nu_1 x}{2(j+1)(\nu_2+\nu_1 x)}\frac{\frac{\nu_1}{2}+\frac{\nu_2}{2}+j}{\frac{\nu_1}{2}+j} \tag{3.87}$$

$$= \frac{\eta\nu_1 x}{2(j+1)(\nu_2+\nu_1 x)}\left(1+\frac{\frac{\nu_2}{2}}{\frac{\nu_1}{2}+j}\right), \tag{3.88}$$

and therefore R_j decreases in j with $R_{j+1}/R_j < 1$. If $R_0 < 1$, then S_j decreases in j for any nonnegative integer j. On the other hand, if $R_0 > 1$, then R_j becomes less than 1 eventually, and S_j increases to a maximum and then decreases. To prove this, note that

$$R_j < 1 \tag{3.89}$$

$$\Leftrightarrow \frac{\eta\nu_1 x}{2(j+1)(\nu_2+\nu_1 x)}\left(1+\frac{\frac{\nu_2}{2}}{\frac{\nu_1}{2}+j}\right) < 1 \tag{3.90}$$

$$\Leftrightarrow j^2 + \left(\frac{\nu_1}{2}+1-\frac{\eta\nu_1 x}{2(\nu_2+\nu_1 x)}\right)j + \frac{\nu_1}{2} - \frac{\eta\nu_1(\nu_2+\nu_1)x}{4(\nu_2+\nu_1 x)} > 0, \tag{3.91}$$

and the left hand side of (3.91) is the function of a convex parabola, and hence there exists some j satisfying the inequality (3.91). Let j^* be an index of j such that $R_{j^*} < 1$. Then S_j decreases for $j \geq j^*$. Hence

$$\sum_{j=j^*+1}^{\infty} S_j = S_{j^*}(R_{j^*} + R_{j^*}R_{j^*+1} + R_{j^*}R_{j^*+1}R_{j^*+2} + \cdots) \tag{3.92}$$

$$\leq S_{j^*}(R_{j^*} + R_{j^*}^2 + R_{j^*}^3 + \cdots) \tag{3.93}$$

$$= \frac{S_{j^*}R_{j^*}}{1-R_{j^*}} \leq \frac{S_{j^*}}{1-R_{j^*}} < \infty, \tag{3.94}$$

so the sum of the infinite series S_j is convergent. ∎

Then we can set an error bound E beforehand, and find $k \geq j^*$ such that $\frac{S_k}{1-R_k} < E$. Then $\sum_{j=0}^{\infty} S_j - \sum_{j=0}^{k} S_j$ is also controlled below this error bound.

Algorithm 3.4.1 *1. Starting from $j = 0$, calculate S_0, S_1 and $R_0 = S_1/S_0$.*

2. ***If*** *$R_0 \leq 1$,* ***then***

2.1 $k = 0$;

2.2 ***while*** *$\frac{S_k}{1-R_k} > E$,*

2.2.1 calculate S_k and R_k;

2.2.2 $k = k + 1$;

2.3 Calculate $\sum_{j=0}^{k} S_j$.

3. ***If*** *$R_0 > 1$,* ***then***

3.1 $j^* = 0$;

3.2 ***while*** $R_{j^*} > 1$,

3.2.1 calculate R_{j^*} *and* S_{j^*};

3.3.2 $j^* = j^* + 1$;

3.3 $k = j^*$;

3.4. ***while*** $\frac{S_k}{1-R_k} > E$,

3.4.1 calculate S_k *and* R_k;

3.4.2 $k = k + \dot{1}$;

3.5 Calculate $\sum_{j=0}^{k} S_j$.

Remark 3 *The "R" function "*$\mathsf{df}()$*" in package "stats" [71] can be used to compute the density of* $F'_{\nu_1,\nu_2}(\zeta)$ *instead of Algorithm 3.4.1.*

We can also use Patnaik's approximation to calculate the distribution of $F'_{\nu_1,\nu_2}(\zeta) = \frac{\chi^2_{\nu_1}(\zeta)/\nu_1}{\chi^2_{\nu_2}/\nu_2}$ by that of $\frac{c\,\nu}{\nu_1} F_{\nu,\nu_2}$.

At last I give an algorithm to calculate the density of the doubly noncentral F distribution. The density of $F''_{\nu_1,\nu_2}(\zeta,\eta)$ is given in Johnson and Kotz [55] as

$$f_{F''_{\nu_1,\nu_2}(\zeta,\,\eta)}(x) = f_{F\nu_1,\nu_2}(x) \sum_{l=0}^{\infty}\sum_{h=0}^{\infty} \frac{B(\frac{\nu_1}{2},\frac{\nu_2}{2})}{B(\frac{\nu_1}{2}+l,\frac{\nu_2}{2}+h)} \tag{3.95}$$

$$\frac{(\frac{1}{2}\zeta)^l e^{-\frac{1}{2}\zeta}}{l!} \frac{(\frac{1}{2}\eta)^h e^{-\frac{1}{2}\eta}}{h!} \left(\frac{\nu_1 x}{\nu_2+\nu_1 x}\right)^l \left(\frac{\nu_2}{\nu_2+\nu_1 x}\right)^h$$

$$= f_{F\nu_1,\nu_2}(x) \sum_{l=0}^{\infty}\sum_{h=0}^{\infty} W_{l,h}, \tag{3.96}$$

where $W_{l,h} = \frac{B(\frac{\nu_1}{2},\frac{\nu_2}{2})}{B(\frac{\nu_1}{2}+l,\frac{\nu_2}{2}+k)} \frac{(\frac{1}{2}\zeta)^l e^{-\frac{1}{2}\zeta}}{l!} \frac{(\frac{1}{2}\eta)^h e^{-\frac{1}{2}\eta}}{h!} \left(\frac{\nu_1 x}{\nu_2+\nu_1 x}\right)^l \left(\frac{\nu_2}{\nu_2+\nu_1 x}\right)^h$.

Consider l and h to be a row and a column indices, respectively. For any fixed h, let $r_h(l)$ be the ratio of any successive terms $W_{l+1,h}$ to $W_{l,h}$ on the column h. By the expression of $f_{F''_{\nu_1,\nu_2}(\zeta_1,\,\eta)}(x)$ in (3.95), this ratio is calculated as:

$$r_h(l) = \frac{W_{l+1,h}}{W_{l,h}} \tag{3.97}$$

$$= \frac{B(\frac{\nu_1}{2}+l,\frac{\nu_2}{2}+h)}{B(\frac{\nu_1}{2}+l+1,\frac{\nu_2}{2}+h)} \frac{1}{l+1} \frac{\frac{1}{2}\zeta\nu_1 x}{\nu_2+\nu_1 x} \tag{3.98}$$

$$= \frac{\Gamma(\frac{\nu_1}{2}+l)\Gamma(\frac{\nu_2}{2}+h)}{\Gamma(\frac{\nu_1}{2}+\frac{\nu_2}{2}+l+h)} \frac{\Gamma(\frac{\nu_1}{2}+\frac{\nu_2}{2}+l+h+1)}{\Gamma(\frac{\nu_1}{2}+l+1)\Gamma(\frac{\nu_2}{2}+h)} \frac{1}{l+1} \frac{\frac{1}{2}\zeta\nu_1 x}{\nu_2+\nu_1 x} \tag{3.99}$$

$$= \frac{\frac{\nu_1}{2} + \frac{\nu_2}{2} + l + h}{\frac{\nu_1}{2} + l} \frac{\frac{1}{2}\zeta\nu_1 x}{\nu_2 + \nu_1 x} \frac{1}{l+1} \tag{3.100}$$

$$= \left(1 + \frac{\frac{\nu_2}{2} + h}{\frac{\nu_1}{2} + l}\right) \frac{1}{l+1} \frac{\frac{1}{2}\zeta\nu_1 x}{\nu_2 + \nu_1 x}, \tag{3.101}$$

where the fourth equality is obtained by using $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$, $\alpha \in \mathcal{R}^+$. It is easy to show that $r_h(l+1) < r_h(l)$ for any fixed h, so the ratio $r_h(l)$ decreases in l. Note that $r_h(l)$ increases in h with $r_h(l) < r_{h+1}(l)$ for any fixed l.

For any fixed l, let $t_l(h)$ be the ratio of any successive terms $W_{l,h+1}$ to $W_{l,h}$ on the row l. By (3.95), we have

$$t_l(h) = \frac{W_{l,h+1}}{W_{l,h}} \tag{3.102}$$

$$= \frac{B(\frac{\nu_1}{2} + l, \frac{\nu_2}{2} + h)}{B(\frac{\nu_1}{2} + l, \frac{\nu_2}{2} + h + 1)} \frac{1}{h+1} \frac{\frac{1}{2}\eta\nu_2}{\nu_2 + \nu_1 x} \tag{3.103}$$

$$= \frac{\frac{\nu_1}{2} + \frac{\nu_2}{2} + l + h}{\frac{\nu_2}{2} + h} \frac{1}{h+1} \frac{\frac{1}{2}\eta\nu_2}{\nu_2 + \nu_1 x} \tag{3.104}$$

$$= \left(1 + \frac{\frac{\nu_1}{2} + l}{\frac{\nu_2}{2} + h}\right) \frac{1}{h+1} \frac{\frac{1}{2}\eta\nu_2}{\nu_2 + \nu_1 x}. \tag{3.105}$$

Similarly, the ratio $t_l(h)$ decreases in h for any fixed l, with $t_l(h+1) < t_l(h)$, and increases in l for any fixed h since $t_l(h) < t_{l+1}(h)$.

Let $s(l,h)$ be the ratio of any diagonally successive terms $W_{l+1,h+1}$ to $W_{l,h}$. By (3.95), $s(l,h)$ has the following form,

$$s(l,h) = \frac{W_{l+1,h+1}}{W_{l,h}} \tag{3.106}$$

$$= \frac{B(\frac{\nu_1}{2} + l, \frac{\nu_2}{2} + h)}{B(\frac{\nu_1}{2} + l + 1, \frac{\nu_2}{2} + h + 1)} \frac{1}{(l+1)(h+1)} \frac{\lambda_1\lambda_2\nu_1\nu_2 x}{4(\nu_2 + \nu_1 x)^2} \tag{3.107}$$

$$= \frac{\lambda_1\lambda_2\nu_1\nu_2 x}{4(\nu_2 + \nu_1 x)^2} \frac{1}{(l+1)(h+1)} \frac{(\frac{\nu_1}{2} + \frac{\nu_2}{2} + l + h + 1)(\frac{\nu_1}{2} + \frac{\nu_2}{2} + l + h)}{(\frac{\nu_1}{2} + l)(\frac{\nu_2}{2} + h)} \tag{3.108}$$

$$= \frac{\lambda_1\lambda_2\nu_1\nu_2 x}{4(\nu_2 + \nu_1 x)^2} \frac{1}{l+1} \left[1 + \frac{\frac{\nu_1}{2} + l + 1}{\frac{\nu_2}{2} + h}\right] \frac{1}{h+1} \left[1 + \frac{\frac{\nu_2}{2} + h}{\frac{\nu_1}{2} + l}\right] \tag{3.109}$$

$$= \frac{\lambda_1\lambda_2\nu_1\nu_2 x}{4(\nu_2 + \nu_1 x)^2} \left[\frac{1}{l+1} + \frac{1}{\frac{\nu_2}{2} + h} + \frac{\frac{\nu_1}{2}}{(l+1)(\frac{\nu_2}{2} + h)}\right] \tag{3.110}$$

$$\left[\frac{1}{h+1} + \frac{1}{\frac{\nu_1}{2} + l} + \frac{\frac{\nu_2}{2} - 1}{(h+1)(\frac{\nu_1}{2} + l)}\right].$$

It is easy to show that for any fixed l, $s(l,\dot{h})$ decreases in h with $s(l,\dot{h+1}) < s(l,\dot{h})$, and for any fixed h, $s(l,\dot{h})$ decreases in l with $s(\dot{l+1},h) < s(\dot{l},h)$. It can be shown that $s(l,h)$

decreases in both l and h with $s(l+1,h+1) < s(l,h)$ if $\nu_2 = m-k-1 \geq 2 \Leftrightarrow m \geq k+3$, where ν_2 is the degrees of freedom of the doubly noncentral t distribution of r_i^*. This is usually true for most datasets encountered.

Note that the ratio $s(l,h)$ has the relationship with $r_h(l)$ and $t_l(h)$:

$$s(l,h) = \frac{W_{l+1,h+1}}{W_{l,h}} \tag{3.111}$$

$$= \frac{W_{l+1,h+1}}{W_{l+1,h}} \frac{W_{l+1,h}}{W_{l,h}} \tag{3.112}$$

$$= t_{l+1}(h) r_h(l) \tag{3.113}$$

$$= t_l(h) r_{h+1}(l). \tag{3.114}$$

The lemma given next show some properties of the ratios defined in (3.101) and (3.105).

Lemma 3.4.2 *(1). For any fixed l, there exists some h such that $r_h(l) < 1$.*

(2). For any fixed h, there exists some l such that $t_l(h) < 1$.

(3). There exist some l and h such that both $r_h(l) < 1$ and $t_l(h) < 1$.

The proof of this lemma is given in Appendix A.

Remark 4 *For a finite h, if $r_h(0) \leq 1$, then $W_{l,h}$ decreases in l; if $r_h(0) > 1$, then $W_{l,h}$ increases in l until a maximum and then starts decreasing in l once $r_h(l)$ becomes less than one.*

For any finite l, if $t_l(h) \leq 1$, then $W_{l,h}$ decreases in h; if $t_l(h) > 1$, then $W_{l,h}$ increases in h until a maximum and then starts decreasing in l once $r_h(l)$ becomes less than one.

By Lemma 3.4.2, there exist l^* and h^* such that both $r_{h^*+1}(l^*) < 1$ and $t_{l^*}(h^*) < 1$. Since $r_{h^*}(l^*) \leq r_{h^*+1}(l^*)$, then $r_{h^*}(l^*) < 1$ and $s(l^*,h^*) = t_{l^*}(h^*) r_{h^*+1}(l^*) < 1$.

I show next that the infinite double sum $\sum_{l=0}^{\infty}\sum_{h=0}^{\infty} W_{l,h}$ is bounded, so I can calculate a finite double sum of $W_{l,h}$ until $\sum_{l=0}^{\infty}\sum_{h=0}^{\infty} W_{l,h}$ is convergent.

Lemma 3.4.3 *(1). For any l^* and h^* such that $r_{h^*}(l^*) < r_{h^*+1}(l^*) < 1$ and $t_{l^*}(h^*) < 1$,*

$$f_{F''_{\nu_1,\nu_2}(\zeta_1,\,\eta)}(x) \leq$$

$$f_{F\nu_1,\nu_2}(x) \left\{ \begin{array}{l} \sum_{l=0}^{l^*}\sum_{h=0}^{h^*} W_{l,h} + \sum_{l=0}^{l^*} \frac{W_{l,h^*}^2}{W_{l,h^*} - W_{l,h^*+1}} + \sum_{h=0}^{h^*} \frac{W_{l^*,h}^2}{W_{l^*,h} - W_{l^*+1,h}} + \\ \frac{W_{l^*,h^*}^2}{W_{l^*,h^*} - W_{l^*+1,h^*+1}} \left(\frac{W_{l^*,h^*}}{W_{l^*,h^*} - W_{l^*+1,h^*}} + \frac{W_{l^*,h^*}}{W_{l^*,h^*} - W_{l^*,h^*+1}} \right) \end{array} \right\}. \tag{3.115}$$

(2). Given l^* and h^* that satisfy $r_{h^*}(l^*) < r_{h^*+1}(l^*) < 1$ and $t_{l^*}(h^*) < 1$, for any sequence $\{h_l, l = 0, \cdots, l^*\}$ such that $t_l(h_l) < 1$, $f_{F''_{\nu_1,\nu_2}(\zeta_1,\ \eta)}(x)$ is bounded by

$$f_{F\nu_1,\nu_2}(x) \left\{ \begin{array}{l} \sum_{l=0}^{l^*} \sum_{h=0}^{h_l} W_{l,h} + \sum_{l=0}^{l^*} \frac{W_{l,h_l}^2}{W_{l,h_l} - W_{l,h_l+1}} + \sum_{h=0}^{h^*} \frac{W_{l^*,h}^2}{W_{l^*,h} - W_{l^*+1,h}} + \\ \frac{W_{l^*,h^*}^2}{W_{l^*,h^*} - W_{l^*+1,h^*+1}} \left(\frac{W_{l^*,h^*}}{W_{l^*,h^*} - W_{l^*+1,h^*}} + \frac{W_{l^*,h^*}}{W_{l^*,h^*} - W_{l^*,h^*+1}} \right) \end{array} \right\}. \quad (3.116)$$

(3). Given l^* and h^* that satisfy $r_{h^*}(l^*) < r_{h^*+1}(l^*) < 1$ and $t_{l^*}(h^*) < 1$, for any sequence $\{l_h, h = 0, \cdots, h^*\}$ such that $r_{h^*}(l_h) < 1$, $f_{F''_{\nu_1,\nu_2}(\zeta_1,\ \eta)}(x)$ is bounded by

$$f_{F\nu_1,\nu_2}(x) \left\{ \begin{array}{l} \sum_{h=0}^{h^*} \sum_{l=0}^{l_h} W_{l,h} + \sum_{l=0}^{l^*} \frac{W_{l,h^*}^2}{W_{l,h^*} - W_{l,h^*+1}} + \sum_{h=0}^{h^*} \frac{W_{l_h,h}^2}{W_{l_h,h} - W_{l_h+1,h}} + \\ \frac{W_{l^*,h^*}^2}{W_{l^*,h^*} - W_{l^*+1,h^*+1}} \left(\frac{W_{l^*,h^*}}{W_{l^*,h^*} - W_{l^*+1,h^*}} + \frac{W_{l^*,h^*}}{W_{l^*,h^*} - W_{l^*,h^*+1}} \right) \end{array} \right\}. \quad (3.117)$$

Proof.

I show first that (1) can be obtained by choosing $h_l = h^*$, $\forall\ l = 0, \cdots, l^*$ in (2) or by choosing $l_h = l^*$, $\forall\ h = 0, \cdots, h^*$ in (3).

By Lemma 3.4.2, there exist l^* and h^* such that $r_{h^*+1}(l^*) < 1$ and $t_{l^*}(h^*) < 1$. Fix such l^* and h^*. Since $t_l(h)$ increases in l for any fixed h, then $t_l(h) < t_{l^*}(h) < 1$ for any $0 \le l < l^*$. Therefore $\{h_l = h^*, l = 0, \cdots, l^*\}$ is a sequence such that $t_l(h_l) < 1$. Similarly, since $r_h(l)$ increases in h for any fixed l, then $r_h(l^*) < r_{h^*}(l^*) < 1$ for any $0 \le h < h^*$. Therefore $\{l_h = l^*, h = 0, \cdots, h^*\}$ is a sequence such that $r_h(l_h) < 1$. Hence (1) can be obtained from (2) or (3).

I show next (2) is true. Note that the double sum $\sum_{l=0}^{\infty} \sum_{h=0}^{\infty} W_{l,h}$ can be written as

$$\sum_{l=0}^{\infty} \sum_{h=0}^{\infty} W_{l,h} = \sum_{l=0}^{l^*} \sum_{h=0}^{\infty} W_{l,h} + \sum_{l=l^*+1}^{\infty} \sum_{h=0}^{\infty} W_{l,h} \quad (3.118)$$

$$= \sum_{l=0}^{l^*} \sum_{h=0}^{h_l} W_{l,h} + \sum_{l=0}^{l^*} \left(\sum_{h=h_l+1}^{\infty} W_{l,h} \right) + \quad (3.119)$$

$$\sum_{h=0}^{h^*} \left(\sum_{l=l^*+1}^{\infty} W_{l,h} \right) + \sum_{l=l^*+1}^{\infty} \sum_{h=h^*+1}^{\infty} W_{l,h}.$$

for any $l^* \ge 0$, $h^* \ge 0$ and $h_l \ge 0$, $\forall\ l = 0, \cdots, l^*$.

In order to find an upper bound of $\sum_{l=0}^{\infty} \sum_{h=0}^{\infty} W_{l,h}$, we need to find upper bounds of $\sum_{l=l^*+1}^{\infty} W_{l,h}$, $\sum_{h=h^*+1}^{\infty} W_{l,h}$, and $\sum_{l=l^*+1}^{\infty} \sum_{h=h^*+1}^{\infty} W_{l,h}$. First, for any fixed h, if l_h satisfies $r_h(l_h) < 1$, then we have

$$\sum_{l=l_h+1}^{\infty} W_{l,h} = W_{l_h,h} \left\{ r_h(l_h) + r_h(l_h) r_h(l_h+1) + r_h(l_h) r_h(l_h+1) r_h(l_h+2) + \cdots \right\} \quad (3.120)$$

$$\leq W_{l_h,h}\left\{r_h(l_h)+r_h^2(l_h)+r_h^3(l_h)+\cdots\right\} \tag{3.121}$$

$$=\frac{W_{l_h,h}r_h(l_h)}{1-r_h(l_h)}\leq\frac{W_{l_h,h}}{1-r_h(l_h)}. \tag{3.122}$$

Secondly, for any fixed l, if h_l satisfies $t_l(h_l)<1$, then we have

$$\sum_{h=h_l+1}^{\infty}W_{l,h}=W_{l,h_l}\left\{t_l(h_l)+t_l(h_l)t_l(h_l+1)+t_l(h_l)t_l(h_l+1)t_l(h_l+2)+\cdots\right\} \tag{3.123}$$

$$\leq W_{l,h_l}\left\{t_l(h_l)+t_l^2(h_l)+t_l^3(h_l)+\cdots\right\} \tag{3.124}$$

$$=\frac{W_{l,h_l}t_l(h_l)}{1-t_l(h_l)}\leq\frac{W_{l,h_l}}{1-t_l(h_l)}. \tag{3.125}$$

Thirdly, let l^* and h^* be indices such that both $r_{h^*+1}(l^*)<1$ and $t_{l^*}(h^*)<1$. Since $r_{h^*}(l^*)<r_{h^*+1}(l^*)$, hence $r_{h^*}(l^*)<1$ and $s(l^*,h^*)=t_{l^*}(h^*)r_{h^*+1}(l^*)<1$. The terms $W_{l,h},\ l\geq l^*,h\geq h^*$ are

$$\begin{array}{ccccc}
W_{l^*,h^*} & W_{l^*,h^*+1} & W_{l^*,h^*+2} & W_{l^*,h^*+3} & \cdots \\
W_{l^*+1,h^*} & W_{l^*+1,h^*+1} & W_{l^*+1,h^*+2} & W_{l^*+1,h^*+3} & \cdots \\
W_{l^*+2,h^*} & W_{l^*+2,h^*+1} & W_{l^*+2,h^*+2} & W_{l^*+2,h^*+3} & \cdots \\
W_{l^*+3,h^*} & W_{l^*+3,h^*+1} & W_{l^*+3,h^*+2} & W_{l^*+3,h^*+3} & \cdots \\
\vdots & \vdots & \vdots & \vdots & \ddots
\end{array},$$

and if we add these terms diagonally, then the resulting sum is

$$\sum_{l=l^*}^{\infty}\sum_{h=h^*}^{\infty}W_{l,h}= \tag{3.126}$$

$$(W_{l^*,h^*}+W_{l^*+1,h^*+1}+W_{l^*+2,h^*+2}+\cdots)+ \tag{3.127}$$

$$(W_{l^*,h^*+1}+W_{l^*+1,h^*+2}+W_{l^*+2,h^*+3}+\cdots)+$$

$$(W_{l^*,h^*+2}+W_{l^*+1,h^*+3}+W_{l^*+2,h^*+4}+\cdots)+\cdots+$$

$$(W_{l^*+1,h^*}+W_{l^*+2,h^*+1}+W_{l^*+3,h^*+2}+\cdots)+$$

$$(W_{l^*+2,h^*}+W_{l^*+3,h^*+1}+W_{l^*+4,h^*+2}+\cdots)+\cdots$$

$$=W_{l^*,h^*}\left\{1+s(l^*,h^*)+s(l^*,h^*)s(l^*+1,h^*+1)+\cdots\right\}+ \tag{3.128}$$

$$W_{l^*,h^*+1}\left\{1+s(l^*,h^*+1)+s(l^*,h^*+1)s(l^*+1,h^*+2)+\cdots\right\}+$$

$$W_{l^*,h^*+2}\left\{1+s(l^*,h^*+2)+s(l^*,h^*+2)s(l^*+1,h^*+3)+\cdots\right\}+\cdots+$$

$$W_{l^*+1,h^*}\left\{1+s(l^*+1,h^*)+s(l^*+1,h^*)s(l^*+2,h^*+1)+\cdots\right\}+$$

$$W_{l^*+2,h^*}\left\{1+s(l^*+2,h^*)+s(l^*+2,h^*)s(l^*+3,h^*+1)+\cdots\right\}+\cdots$$

$$\leq W_{l^*,h^*}\left\{1+s(l^*,h^*)+s^2(l^*,h^*)+s^3(l^*,h^*)+\cdots\right\}+ \tag{3.129}$$

$$
\begin{aligned}
&W_{l^*,h^*+1}\left\{1+s(l^*,h^*)+s^2(l^*,h^*)+s^3(l^*,h^*)+\cdots\right\}+\\
&W_{l^*,h^*+2}\left\{1+s(l^*,h^*)+s^2(l^*,h^*)+s^3(l^*,h^*)+\cdots\right\}+\cdots+\\
&W_{l^*+1,h^*}\left\{1+s(l^*,h^*)+s^2(l^*,h^*)+s^3(l^*,h^*)+\cdots\right\}+\\
&W_{l^*+2,h^*}\left\{1+s(l^*,h^*)+s^2(l^*,h^*)+s^3(l^*,h^*)+\cdots\right\}+\cdots\\
&\leq \frac{1}{1-s(l^*,h^*)}\left\{\begin{array}{c}(W_{l^*,h^*}+W_{l^*,h^*+1}+W_{l^*,h^*+2}+\cdots)+\\(W_{l^*,h^*}+W_{l^*+1,h^*}+W_{l^*+2,h^*}+\cdots)\end{array}\right\} &(3.130)\\
&\leq \frac{1}{1-s(l^*,h^*)}\left\{\begin{array}{c}W_{l^*,h^*}\left[1+t_{l^*}(h^*)+t_{l^*}(h^*)t_{l^*}(h^*+1)+\cdots\right]+\\W_{l^*,h^*}\left[1+r_{h^*}(l^*)+r_{h^*}(l^*)r_{h^*}(l^*+1)+\cdots\right]\end{array}\right\} &(3.131)\\
&\leq \frac{1}{1-s(l^*,h^*)}\left\{\begin{array}{c}W_{l^*,h^*}\left[1+t_{l^*}(h^*)+t^2_{l^*}(h^*)+\cdots\right]+\\W_{l^*,h^*}\left[1+r_{h^*}(l^*)+r^2_{h^*}(l^*)+\cdots\right]\end{array}\right\} &(3.132)\\
&= \frac{W_{l^*,h^*}}{1-s(l^*,h^*)}\left\{\frac{1}{1-t_{l^*}(h^*)}+\frac{1}{1-r_{h^*}(l^*)}\right\} &(3.133)\\
&\leq \frac{W_{l^*,h^*}}{\{1-s(l^*,h^*)\}\{1-t_{l^*}(h^*)\}}+\frac{W_{l^*,h^*}}{\{1-s(l^*,h^*)\}\{1-r_{h^*}(l^*)\}}. &(3.134)
\end{aligned}
$$

Since $r_h(l^*) < r_{h^*}(l^*) < 1$ for any $0 \leq h < h^*$ and $t_l(h_l) < 1$ for any $0 \leq l < l^*$. Hence $\sum_{l=l^*+1}^{\infty} W_{l,h} \leq \frac{W_{l^*,h}}{1-r_h(l^*)}$ for any $0 \leq h \leq h^*$ and $\sum_{h=h_l+1}^{\infty} W_{l,h} \leq \frac{W_{l,h_l}}{1-t_l(h_l)}$ for any $0 \leq l \leq l^*$, by (3.122) and (3.125). Therefore by 3.119 and 3.133, we have

$$
\begin{aligned}
\sum_{l=0}^{\infty}\sum_{h=0}^{\infty} W_{l,h} &\leq \sum_{l=0}^{l^*}\sum_{h=0}^{h_l} W_{l,h}+\sum_{l=0}^{l^*}\frac{W_{l,h_l}}{1-t_l(h_l)}+\sum_{h=0}^{h^*}\frac{W_{l^*,h}}{1-r_h(l^*)}+ &(3.135)\\
&\quad \frac{W_{l^*,h^*}}{1-s(l^*,h^*)}\left\{\frac{1}{1-t_{l^*}(h^*)}+\frac{1}{1-r_{h^*}(l^*)}\right\}\\
&\leq \sum_{l=0}^{l^*}\sum_{h=0}^{h_l} W_{l,h}+\sum_{l=0}^{l^*}\frac{W^2_{l,h_l}}{W_{l,h_l}-W_{l,h_l+1}}+\sum_{h=0}^{h^*}\frac{W^2_{l^*,h}}{W_{l^*,h}-W_{l^*+1,h}}+ &(3.136)\\
&\quad \frac{W^2_{l^*,h^*}}{W_{l^*,h^*}-W_{l^*+1,h^*+1}}\left(\frac{W_{l^*,h^*}}{W_{l^*,h^*}-W_{l^*+1,h^*}}+\frac{W_{l^*,h^*}}{W_{l^*,h^*}-W_{l^*,h^*+1}}\right).
\end{aligned}
$$

Then (3.115) results from (3.136) with choosing $h_l = h^*$, $\forall\ l = 0, \cdots, l^*$.

Similarly, (3) follows by

$$
\begin{aligned}
\sum_{l=0}^{\infty}\sum_{h=0}^{\infty} W_{l,h} &= \sum_{h=0}^{h^*}\sum_{l=0}^{\infty} W_{l,h}+\sum_{h=h^*+1}^{\infty}\sum_{l=0}^{\infty} W_{l,h} &(3.137)\\
&= \sum_{h=0}^{h^*}\sum_{l=0}^{l_h} W_{l,h}+\sum_{h=0}^{h^*}\left(\sum_{l=l^*_{h^*}+1}^{\infty} W_{l,h}\right)+ &(3.138)\\
&\quad \sum_{l=0}^{l^*}\left(\sum_{h=h^*+1}^{\infty} W_{l,h}\right)+\sum_{l=l^*+1}^{\infty}\sum_{h=h^*+1}^{\infty} W_{l,h}.
\end{aligned}
$$

■

Remark 5 *I chose to use the bound (3.116) instead of (3.115) because most h_l are much smaller than h^*, so the computation time can be saved by using (3.116). Moreover, since in most cases the denominator noncentral parameter η in (3.40) is much larger than the numerator noncentral parameters ζ_0 (3.41) and ζ_1 (3.42), the value of h^* is much larger than l^*. Therefore I use the bound (3.116) rather than (3.117).*

The last three terms in (3.116) can be bounded below a certain error bound, then I can calculate the finite sum $\sum_{l=0}^{l^*}\sum_{h=0}^{h_l} W_{l,h}$. In my algorithm given below, I use another form of the bound (3.116) as given in (3.135) which has ratios $r_h(l)$, $t_l(h)$ and $s(j,k)$. The last three terms $\sum_{l=0}^{l^*}\frac{W_{l,h_l}}{1-t_l(h_l)}$, $\sum_{h=0}^{h^*}\frac{W_{l^*,h}}{1-r_h(l^*)}$ and $\frac{W_{l^*,h^*}}{1-s(l^*,h^*)}\left\{\frac{1}{1-t_{l^*}(h^*)}+\frac{1}{1-r_{h^*}(l^*)}\right\}$ in (3.135) can be bounded separately. Given an error bound E, find $j \geq l^*$, $k \geq h^*$ and $k_l \geq h_l$, $l = 0, \cdots, j$ with $k_j = k$ for convenience, such that $\frac{W_{j,h}}{1-r_h(j)} \leq E$, $h = 0, \cdots, k$, $\frac{W_{l,k_l}}{1-t_l(k_l)} \leq E$, $l = 0, \cdots, j$, and $\frac{W_{j,k}}{1-s(j,k)}\left\{\frac{1}{1-t_j(k)}+\frac{1}{1-r_k(j)}\right\} \leq E$. Then

$$\begin{aligned}
&\sum_{l=0}^{\infty}\sum_{h=0}^{\infty} W_{l,h} - \sum_{l=0}^{j}\sum_{h=0}^{k_l} W_{l,h} \\
&\leq \sum_{l=0}^{j}\frac{W_{l,k_l}}{1-t_l(k_l)} + \sum_{h=0}^{k}\frac{W_{j,h}}{1-r_h(j)} + \qquad (3.139)\\
&\frac{W_{j,k}}{1-s(j,k)}\left\{\frac{1}{1-t_j(k)}+\frac{1}{1-r_k(j)}\right\} \\
&\leq (j+1)E + (k+1)E + E \\
&= (k+j+3)E. \qquad (3.140)
\end{aligned}$$

The algorithm for computing the density of $F''_{\nu_1,\nu_2}(\zeta,\eta)$ is given below.

Algorithm 3.4.2 *1.* ***If*** $\eta = 0$,

then *use Algorithm 3.4.1 to calculate $f_{F'_{\nu_1,\nu_2}(\zeta)}(x)$.*

2. ***If*** $\zeta = 0$,

then *use Algorithm 3.4.1 to calculate the density $f_{F''_{\nu_1,\nu_2}(0,\eta)}(x) = x^{-1}f_{F'_{\nu_1,\nu_2}(\eta)}(x^{-1})$.*

3. ***Otherwise****, for both $\zeta, \eta > 0$:*

3.1 $l = 0$

3.2 Calculate $W_{l,0}$ *and* $t_l(0)$.

3.3 $\mathrm{mod}_l = 0$.

3.4 ***While*** $t_l(\mathrm{mod}_l) > 1$*:*

3.4.1 $\mathrm{mod}_l = \mathrm{mod}_l + 1$.

3.4.2 Calculate W_{l,mod_l} *and* $t_l(\mathrm{mod}_l)$.

3.5 $k_l = \mathrm{mod}_l$.

3.6 ***While*** $\frac{W_{l,k_l}}{1-t_l(k_l)} > \dot{E}$*:*

3.6.1 $k_l = k_l + 1$.

3.6.2 Calculate W_{l,k_l} *and* $t_l(k_l)$.

3.7 Calculate $\sum_{h=0}^{k_l} W_{l,h}$.

3.8 Calculate $r_{\mathrm{mod}_l+1}(l)$ *and* $s(l, \mathrm{mod}_l)$.

3.9 ***While*** $r_{\mathrm{mod}_l+1}(l) > 1$ *or* $s(l, \mathrm{mod}_l)$ *or* $\frac{W_{l,\mathrm{mod}_l}}{1-s(l,\mathrm{mod}_l)} \left\{ \frac{1}{1-t_l(\mathrm{mod}_l)} + \frac{1}{1-r_{\mathrm{mod}_l}(l)} \right\} > E$*:*

3.9.1 $l = l + 1$.

3.9.2 $\mathrm{mod}_l = 0$

3.9.3 Calculate W_{l,mod_l} *and* $t_l(\mathrm{mod}_l)$.

3.9.4 ***while*** $t_l(\mathrm{mod}_l) > 1$ *:*

$\mathrm{mod}_l = \mathrm{mod}_l + 1$.

Calculate W_{l,mod_l} *and* $t_l(\mathrm{mod}_l)$.

3.9.5 $k_l = \mathrm{mod}_l$.

3.9.6 While $\frac{W_{l,k_l}}{1-t_l(k_l)} > \dot{E}$*:*

$k_l = k_l + 1$.

Calculate W_{l,k_l} *and* $t_l(k_l)$.

3.9.7 Calculate the sum $\sum_{h=0}^{k_l} W_{l,h}$.

3.9.8 Calculate $r_{\mathrm{mod}_l+1}(l)$ *and* $s(l, \mathrm{mod}_l)$.

3.10 Then j *is the value of* l *which breaks out the condition in the previous step, i.e.* $j = l$ *and* $k = \mathrm{mod}_l$.

3.11 Calculate $S = \sum_{l=0}^{j} \left(\sum_{h=0}^{k_l} W_{l,h} \right)$.

3.12 Calculate $f_{F\nu_1,\nu_2}(x)$.

3.13 The density of $F''_{\nu_1,\nu_2}(\zeta,\ \eta)$ *for a fixed* x *is calculated as* $S \cdot f_{F\nu_1,\nu_2}(x)$.

The density function of the central F distribution with degrees of freedom ν_1 and ν_2 for a fixed x can be calculated by using "R" function $\mathsf{df}(x, \nu_1, \nu_2)$ [71].

Using the above algorithm the computational error can be controlled below $(j+k+3)E$. Reset E to be $E/(j+k+3)$, the total computational error is then smaller than the desired error bound E. For example, for $x = 10$, $\nu_1 = 1$, $\nu_2 = 97$, $\zeta = 50$, $\eta = 500$ and $E = 10^{-10}$, the value of $j + k + 3$ is 333, so the actual error is controlled below 3.33×10^{-8}.

The Algorithm 3.4.2 does not converge too slowly for moderate values of x, ν_1, ν_2, ζ and η, so the value of j and k are not very large. For example, for $x = 10$, $\nu_1 = 1$, $\nu_2 = 97$, $\zeta = 50$, $\eta = 500$ and $E = 10^{-10}$, it takes 1.98 seconds on a computer with Intel Pentium D Dual processor of 2.8GHz and 2.79GHz, 2GB of Ram. However, the computational time depends on the value of x, ν_1, ν_2, ζ and η. Moreover, if computation of one density takes 1.98 seconds, then it takes at least $1.98mn$ seconds, where m is the total number of observations and n is the number of random samples generated by using Algorithm 3.3.1, to calculate the marginal posterior probabilities $\Pr(H_i = 0 \mid r_i^{*2})$.

The Algorithm 3.4.2 is time-consuming for large m or large n. Hence I decided to use Patnaik's approximation to calculate the density of doubly noncentral F distribution of $F''_{\nu_1,\nu_2}(\zeta, \eta)$ by that of $\frac{1-\zeta\nu_1^{-1}}{1-\eta\nu_2^{-1}} F_{\nu,\nu'}$ with $\nu = (\nu_1 + \zeta)^2/(\nu_1 + 2\zeta)$ and $\nu' = (\nu_2 + \eta)^2/(\nu_2 + 2\eta)$ [54]. However, we need to examine the error of the approximation, that is, compare the results of the approximation with those of the Algorithm 3.4.2. Since the purpose of using this algorithm is to compare Patnaik's approximation to the density of $F''_{\nu_1,\nu_2}(\zeta, \eta)$, I only need to compute the density of $F''_{\nu_1,\nu_2}(\zeta, \eta)$ with an error below some bound. Though the error bound of Algorithm 3.4.2 depends on the value of x, ν_1, ν_2, ζ and η, it still controls computational error as long as $j + k + 3$ is not too large, which is true for the values of parameters used in my simulation study. It could be a future work to develop a more efficient algorithm or a better approximation to calculate the density of $F''_{\nu_1,\nu_2}(\zeta, \eta)$.

Since the degrees of freedom of the numerator of r^{*2} is always 1, only the value of x, ν_2, ζ and η are varied. The quantile x varies from 0.01 to 10. Since the degrees of freedom of the denominator is $\nu_2 = m-k-1$ and in this thesis I only consider the simple linear regression, then the chosen values of $\nu_2 = 97,\ 47,\ 17$ correspond to the sample size $m = 100,\ 50,\ 20$. The noncentrality parameter of the numerator are set to be $0.01,\ 1,\ 10$ and 50, and the noncentrality parameter of the denominator are chosen as $0.01,\ 1,\ 10,\ 100$ and 500. In fact, in my simulation results given in the next section, the value of η is usually greater

than that of ζ. The density of doubly noncentral F random variable $F''_{\nu_1,\nu_2}(\zeta,\eta)$ calculated by Algorithm 3.4.2 and that computed by Patnaik's approximation are given in Table 3.1 to Table 3.12. Although there are other methods more accurate than Patnaik's method for computing the CDF, I still chose the latter because it is simple and the approximation to the density of the doubly noncentral F is not bad for the values of the parameters encountered when Algorithm 3.3.1 is applied.

$x = 0.001,\ \nu_2 = 97$								
$\eta\backslash\zeta$	0.01		1		10		50	
	True	Difference	True	Difference	True	Difference	True	Difference
0.01	12.51	7.20e-12	7.63	4.39e-12	8.5e-2	7.22e-14	1.79e-10	4.39e-14
1	12.58	6.68e-08	7.67	4.07e-08	8.6e-2	4.73e-10	1.62e-10	1.79e-11
10	13.14	4.94e-06	8.02	3.01e-06	8.9e-2	3.33e-08	1.42e-10	4.70e-11
100	17.84	5.95e-05	10.88	3.62e-05	1.2e-1	3.99e-07	2.51e-10	1.08e-11
500	31.02	3.08e-05	18.97	1.87e-05	2.2e-1	2.02e-07	4.99e-10	2.28e-12

Table 3.1: True value of the density of doubly noncentral F with $x = 0.001$, $\nu_2 = 97$ and various ξ and η, and Difference = (Patnaik's approximation − true value).

$x = 0.01,\ \nu_2 = 97$								
$\eta\backslash\zeta$	0.01		1		10		50	
	True	Difference	True	Difference	True	Difference	True	Difference
0.01	3.94	2.23e-12	2.41	1.39e-12	2.8e-2	5.74e-14	6.88e-11	6.79e-13
1	3.96	2.12e-8	2.43	1.29e-8	2.8e-2	1.48e-10	5.86e-11	1.14e-11
10	4.14	1.57e-6	2.53	9.51e-7	3.0e-2	1.01e-8	5.31e-11	2.15e-11
100	5.59	1.90e-5	3.44	1.14e-5	4.2e-2	1.16e-7	1.16e-10	5.43e-12
500	9.54	1.00e-5	6.00	5.92e-6	8.6e-2	4.99e-8	3.96e-10	1.22e-12

Table 3.2: True value of the density of doubly noncentral F with $x = 0.01$, $\nu_2 = 97$ and various ξ and η, and Difference = (Patnaik's approximation − true value).

All the values in these tables are reported to the second significant digit. The columns named "True" list densities calculated by Algorithm 3.4.2, in which the true error is controlled below 10^{-10}. The columns named "Difference" give the differences between the densities obtained by using Patnaik's approximation and those computed from Algorithm

$x = 1,\ \nu_2 = 97$								
$\eta\backslash\zeta$	0.01		1		10		50	
	True	Difference	True	Difference	True	Difference	True	Difference
0.01	2.4e-1	3.56e-13	2.3e-1	1.56e-13	2.0e-2	1.69e-15	2.20e-09	5.66e-15
1	2.4e-1	3.39e-09	2.3e-1	1.41e-09	2.0e-2	-1.63e-12	2.28e-09	1.63e-13
10	2.4e-1	2.56e-07	2.3e-1	1.09e-07	2.3e-2	-7.26e-10	3.15e-09	8.27e-13
100	2.1e-1	2.96e-06	2.7e-1	2.25e-06	6.4e-2	-3.13e-07	3.89e-08	2.08e-11
500	4.8e-2	-1.40e-06	1.7e-1	1.58e-06	3.9e-2	-3.58e-06	1.46e-05	5.72e-09

Table 3.3: True value of the density of doubly noncentral F with $x = 1$, $\nu_2 = 97$ and various ξ and η, and Difference = (Patnaik's approximation − true value).

$x = 10,\ \nu_2 = 97$								
$\eta\backslash\zeta$	0.01		1		10		50	
	True	Difference	True	Difference	True	Difference	True	Difference
0.01	1.1e-3	-1.03e-13	6.5e-3	-9.03e-14	6.1e-2	6.43e-14	4.45e-05	1.10e-14
1	1.0e-3	-9.79e-10	6.4e-3	-9.08e-10	6.2e-2	7.95e-10	4.76e-05	1.12e-10
10	7.0e-4	-6.39e-08	4.9e-3	-8.75e-08	6.3e-2	1.80e-07	8.57e-05	1.38e-08
100	1.45e-05	-8.40e-08	2.8e-4	-6.90e-07	3.7e-2	6.35e-06	4.2e-3	2.70e-06
500	2.20e-13	-1.13e-14	7.24e-11	-2.54e-12	6.49e-06	-6.77e-08	1.1e-1	2.92e-05

Table 3.4: True value of the density of doubly noncentral F with $x = 10$, $\nu_2 = 97$ and various ξ and η, and Difference = (Patnaik's approximation − true value).

$x = 0.001,\ \nu_2 = 47$								
$\eta\backslash\zeta$	0.01		1		10		50	
	True	Difference	True	Difference	True	Difference	True	Difference
0.01	12.48	1.22e-10	7.61	7.45e-11	8.5e-2	8.52e-13	1.79e-10	4.56e-14
1	12.61	1.14e-06	7.69	6.93e-07	8.6e-2	7.70e-09	1.62e-10	1.83e-11
10	13.74	6.36e-05	8.38	3.88e-05	9.4e-2	4.28e-07	1.47e-10	5.06e-11
100	22.10	2.40e-04	13.49	1.46e-04	1.5e-1	1.60e-06	3.20e-10	1.36e-11
500	42.54	6.04e-05	26.08	3.66e-05	3.1e-1	3.86e-07	7.73e-10	2.72e-12

Table 3.5: True value of the density of doubly noncentral F with $x = 0.001$, $\nu_2 = 47$ and various ξ and η, and Difference = (Patnaik's approximation − true value).

$x = 0.01,\ \nu_2 = 47$								
$\eta\backslash\zeta$	0.01		1		10		50	
	True	Difference	True	Difference	True	Difference	True	Difference
0.01	3.93	3.88e-11	2.41	2.36e-11	2.8e-2	2.99e-13	6.88e-11	7.07e-13
1	3.97	3.61e-07	2.43	2.19e-07	2.9e-2	2.33e-09	5.87e-11	1.18e-11
10	4.32	2.02e-05	2.65	1.23e-05	3.1e-2	1.29e-07	5.59e-11	2.41e-11
100	6.89	7.68e-05	4.27	4.61e-05	5.4e-2	4.46e-07	1.76e-10	7.35e-12
500	12.77	2.02e-05	8.24	1.15e-05	1.4e-1	7.65e-08	1.01e-09	1.48e-12

Table 3.6: True value of the density of doubly noncentral F with $x = 0.01$, $\nu_2 = 47$ and various ξ and η, and Difference = (Patnaik's approximation − true value).

$x = 1,\ \nu_2 = 47$								
$\eta\backslash\zeta$	0.01		1		10		50	
	True	Difference	True	Difference	True	Difference	True	Difference
0.01	2.4e-1	6.15e-12	2.2e-1	2.59e-12	2.0e-2	-6.30e-15	2.46e-09	1.76e-14
1	2.4e-1	5.75e-08	2.3e-1	2.43e-08	2.0e-2	-1.02e-10	2.66e-09	2.16e-13
10	2.4e-1	3.32e-06	2.4e-1	1.49e-06	2.7e-2	-2.78e-08	5.12e-09	2.71e-12
100	1.5e-1	6.34e-06	2.7e-1	1.43e-05	1.4e-2	-5.98e-06	3.60e-07	8.67e-10
500	4.7e-3	-2.22e-06	3.9e-2	-4.04e-06	6.5e-1	1.02e-05	9.6e-4	6.14e-07

Table 3.7: True value of the density of doubly noncentral F with $x = 1$, $\nu_2 = 47$ and various ξ and η, and Difference = (Patnaik's approximation − true value).

$x = 10,\ \nu_2 = 47$								
$\eta\backslash\zeta$	0.01		1		10		50	
	True	Difference	True	Difference	True	Difference	True	Difference
0.01	1.3e-3	-1.76e-12	7.0e-3	-1.36e-12	6.0e-2	1.56e-12	6.47e-05	2.48e-13
1	1.2e-3	-1.60e-08	6.6e-3	-1.38e-08	6.0e-2	1.97e-08	7.43e-05	2.59e-09
10	5.8e-4	-7.04e-07	4.0e-3	-1.17e-06	6.2e-2	3.92e-06	2.3e-4	3.59e-07
100	3.51e-07	-1.91e-08	1.23e-05	-3.62e-07	8.1e-3	-1.72e-05	4.1e-2	-3.50e-05
500	1.72e-22	1.56e-22	2.34e-19	4.34e-19	5.84e-12	-7.63e-13	6.1e-4	-7.77e-06

Table 3.8: True value of the density of doubly noncentral F with $x = 10$, $\nu_2 = 47$ and various ξ and η, and Difference = (Patnaik's approximation − true value).

$x = 0.001,\ \nu_2 = 17$								
$\eta\backslash\zeta$	0.01		1		10		50	
	True	Difference	True	Difference	True	Difference	True	Difference
0.01	12.37	6.51e-09	7.54	3.97e-09	8.4e-2	4.39e-11	1.77e-10	5.17e-14
1	12.72	5.44e-05	7.76	3.32e-05	8.7e-2	3.67e-07	1.63e-10	1.97e-11
10	15.61	1.4e-3	9.52	8.8e-4	1.1e-1	9.71e-06	1.89e-10	3.76e-11
100	32.69	9.0e-4	19.99	5.4e-4	2.3e-1	5.86e-06	5.23e-10	1.56e-11
500	68.12	1.2e-4	42.15	7.41e-05	5.3e-1	7.19e-07	1.77e-09	3.38e-12

Table 3.9: True value of the density of doubly noncentral F with $x = 0.001$, $\nu_2 = 17$ and various ξ and η, and Difference = (Patnaik's approximation − true value).

$x = 0.01,\ \nu_2 = 17$								
$\eta\backslash\zeta$	0.01		1		10		50	
	True	Difference	True	Difference	True	Difference	True	Difference
0.01	3.89	2.07e-09	2.39	1.25e-09	2.8e-2	1.36e-11	6.86e-11	8.05e-13
1	4.00	1.73e-05	2.45	1.05e-05	2.9e-2	1.11e-07	5.91e-11	1.33e-12
10	4.90	4.6e-4	3.01	2.8e-4	3.6e-2	2.89e-06	9.56e-11	3.50e-12
100	10.02	3.0e-4	6.32	1.7e-4	9.3e-2	1.41e-06	4.58e-10	8.97e-12
500	18.80	4.57e-05	13.24	2.30e-05	3.8e-1	9.18e-08	6.57e-09	2.30e-12

Table 3.10: True value of the density of doubly noncentral F with $x = 0.01$, $\nu_2 = 17$ and various ξ and η, and Difference = (Patnaik's approximation − true value).

$x = 1,\ \nu_2 = 17$								
$\eta\backslash\zeta$	0.01		1		10		50	
	True	Difference	True	Difference	True	Difference	True	Difference
0.01	2.4e-1	3.21e-10	2.2e-1	1.41e-10	2.1e-2	-1.58e-12	3.63e-09	1.19e-14
1	2.3e-1	2.71e-06	2.3e-1	1.21e-06	2.3e-2	-1.89e-08	4.55e-09	1.75e-12
10	2.2e-1	7.39e-05	2.5e-1	4.15e-05	4.4e-2	-3.37e-06	2.42e-08	2.36e-10
100	3.9e-2	-4.18e-05	1.4e-1	3.12e-05	4.4e-1	-9.61e-05	4.48e-05	4.38e-07
500	1.35e-06	-1.61e-08	6.84e-05	-4.35e-07	7.7e-2	-3.39e-05	3.4e-1	-3.78e-05

Table 3.11: True value of the density of doubly noncentral F with $x = 1$, $\nu_2 = 17$ and various ξ and η, and Difference = (Patnaik's approximation − true value).

$x = 10,\ \nu_2 = 17$								
$\eta\backslash\zeta$	0.01		1		10		50	
	True	Difference	True	Difference	True	Difference	True	Difference
0.01	2.0e-3	-8.22e-11	8.3e-3	-4.24e-11	5.5e-2	1.20e-10	1.8e-4	2.36e-11
1	1.7e-3	-6.67e-07	7.3e-3	-4.52e-07	5.6e-2	1.45e-06	2.6e-4	2.49e-07
10	3.7e-4	-1.04e-05	2.4e-3	-2.26e-05	5.1e-2	0.00012	2.5e-3	2.45e-05
100	4.41e-11	-2.32e-11	3.27e-09	-1.30e-09	2.25e-05	-3.32e-06	8.0e-2	4.4e-4
500	2.60e-43	-2.49e-43	3.00e-39	-2.68e-39	5.14e-30	-3.38e-30	1.98e-16	-3.83e-17

Table 3.12: True value of the density of doubly noncentral F with $x = 10$, $\nu_2 = 17$ and various ξ and η, and Difference = (Patnaik's approximation − true value).

3.4.2. We can see from Table 3.1 – 3.12, Patnaik's method gives a good approximation to the density of $F''_{\nu_1,\nu_2}(\zeta,\eta)$. Almost all the absolute differences are smaller in order of magnitude than the last significant figure of the true densities except some extremely small differences, though they tend to be larger for small quantiles and small sample size. The largest difference between the density calculated by Patnaik's approximation and that computed by Algorithm 3.4.2 is 0.0014 when $x = 0.001,\ \nu_2 = 17$, $\zeta = 0.01$ and $\eta = 10$ in Table 3.9. These results confirm the use of Patnaik's method in computing the density of doubly noncentral F distribution. I also use both Algorithm 3.4.2 and Patnaik's approximation in my simulation results for some chosen parameters and compare the desired posterior probabilities for the ith observation being an outlier given its deletion residual obtained by the two methods. The results are given in the next section.

3.5 Simulation study

In this section, I apply the proposed Bayesian approach in Section 3.3 to some simulated datasets. For simplicity, I only generate datasets for simple linear regression, so I do not need to consider the problem of variable selection. In this simulation study, datasets are generated for various simulation parameters, including the total number of observations, the proportion of outliers and the variance of the mean shift of outliers. The marginal posterior probabilities $P(H_i = 1 \mid r_i^{*2})$ for each simulated dataset are calculated by Algorithm 3.3.1, which requires generating a number of the importance samples. The hyper-parameters of importance samples are parameters a and b of the Beta prior for π_0 and the variance of the mean shift. In order to study how sensitive the marginal posterior probabil-

ity $P(H_i = 1 \mid r_i^{*2})$ is to the different priors, the hyper-prior parameters of the importance samples also vary over a range of values for each combination of simulation parameters. Then for each combination of simulation and prior parameter levels, the marginal posterior probabilities $P(H_i = 1 \mid r_i^{*2})$ are calculated for all observations and the AUC value is also calculated. In Section 3.5.1, I simulate two single datasets, with $m = 100$ and $\pi_0 = 0.9$, different variances of the mean shift, and the explanatory variable sampled from Bernoulli or normal distribution. The marginal posterior probability $P(H_i = 1 \mid r_i^{*2})$ is plotted as a function of the deletion residual r_i^{*2}. The ROC curves are plotted and the AUC values are calculated for various priors. In Section 3.5.2, two sets of 1000 datasets with $m = 100$ and $\pi_0 = 0.9$ are generated, and the proposed Bayesian method is applied to each dataset with various priors. The variance of the mean shift and the distribution of explanatory variables are different for the two sets of 1000 simulated datasets. For each multiple datasets and various priors, the average TPR and FPR are calculated for selected thresholds, the average TPR is plotted versus the average FPR, and the average AUC are calculated for various priors. In Section 3.5.3, more values of the simulation parameters are considered, and the results for a factorial design analysis with factors equal to simulation and prior parameters are presented. The ANOVA table and the table of means are also presented in Section 3.5.3. I first include a small value of 20 for the factor m. The ANOVA table indicates a significant effect of m, and both the table of means and the residual plot suggest a large difference between $m = 20$ and the other values of m. Hence I remove the data with $m = 20$ and re-perform the factorial design analysis.

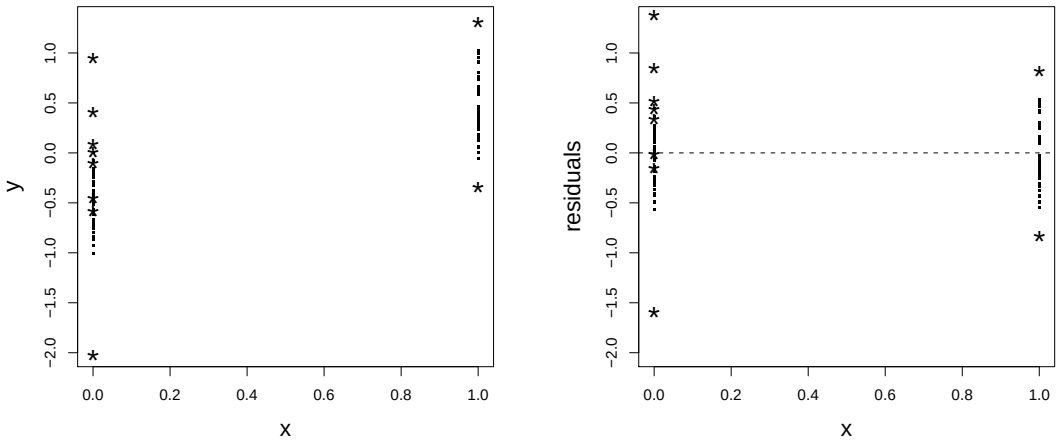
3.5.1 Simulation study of single datasets

I start from the simplest case, binary x. First, m binary $\boldsymbol{x} = (x_1, \cdots, x_m)^T$ are generated from Bernoulli(1/2), that is, x equals 0 or 1 with equal probability 1/2. Secondly, m random errors $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_m)^T$ are generated from $N(0, \sigma^2)$. Then m "clean" data $\boldsymbol{y'}$ are calculated as $X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with all elements of the first column of X equal to 1 and the second column of X equal to $\boldsymbol{x}$. At last, m_1 scaled mean shifts $\mu_1, \cdots, \mu_{m_1}$ are generated from $N(0, V')$, and after being multiplied by σ, they are added to the last m_1 elements of $\boldsymbol{y'}$ to make a vector of responses $\boldsymbol{y}$ with m_1 outliers. In this section, the number of outliers, m_1, is fixed. I use the notation V' to denote the variance of the mean shift μ_i in the simulated datasets, in order to distinguish it from the prior variance of μ_i for the

importance samples, which is denoted by V. I choose the parameters $\boldsymbol{\beta} = (-0.5, 1)^T$. In fact, the choice of $\boldsymbol{\beta}$ does not affect the deletion residuals $r_i^{*2} = \frac{r_i}{s_{(i)}\sqrt{(1-g_i)}}$ because

$$\begin{aligned}
\boldsymbol{r} &= (\ I - G)\boldsymbol{y} && (3.141)\\
&= \left\{I - X(X^TX)^{-1}X^T\right\}(X\boldsymbol{\beta} + \boldsymbol{\varepsilon} + \boldsymbol{\mu}) && (3.142)\\
&= X\boldsymbol{\beta} + \boldsymbol{\varepsilon} + \boldsymbol{\mu} - X(X^TX)^{-1}X^TX\,\boldsymbol{\beta} - X(X^TX)^{-1}X^T(\boldsymbol{\varepsilon} + \boldsymbol{\mu}) && (3.143)\\
&= (I - G)(\boldsymbol{\varepsilon} + \boldsymbol{\mu}), && (3.144)
\end{aligned}$$

where $\boldsymbol{\mu} = (\overbrace{0, \cdots, 0}^{m_0}, \mu_1, \cdots, \mu_{m_1})^T$. By choosing $m = 100$, $m_1 = 10$, $\sigma^2 = 1/16$ and $V' = 16$, the scatter plot and the residual versus explanatory variable plot of the first simulated dataset are shown in Figure 3.1. We can see from the figure that five outliers are apparent outliers, while the other half are hard to distinguish from the nulls.



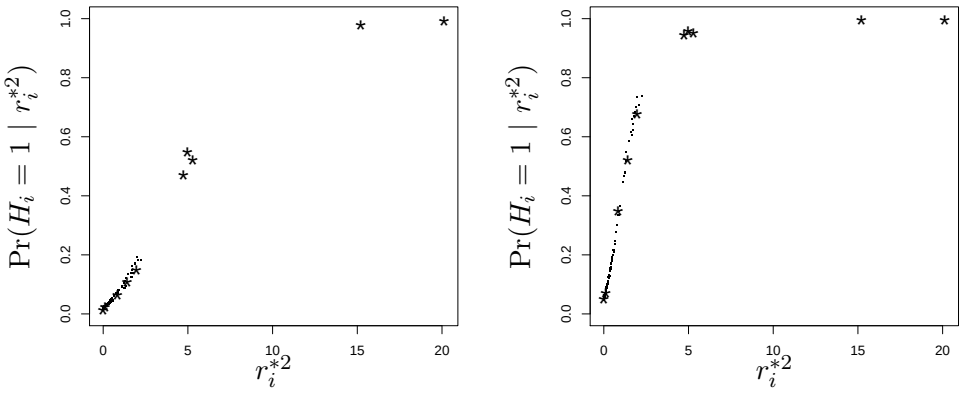
(a) $\boldsymbol{y}$ vs. $\boldsymbol{x}$ (b) Residual vs. $\boldsymbol{x}$

Figure 3.1: The scatter plot and the residual plot of a dataset with 100 points, of which 10 outliers have mean shifts simulated from N(0,1). The observations are shown as dots ($\cdot$) for the nulls and stars (*) for the alternatives.

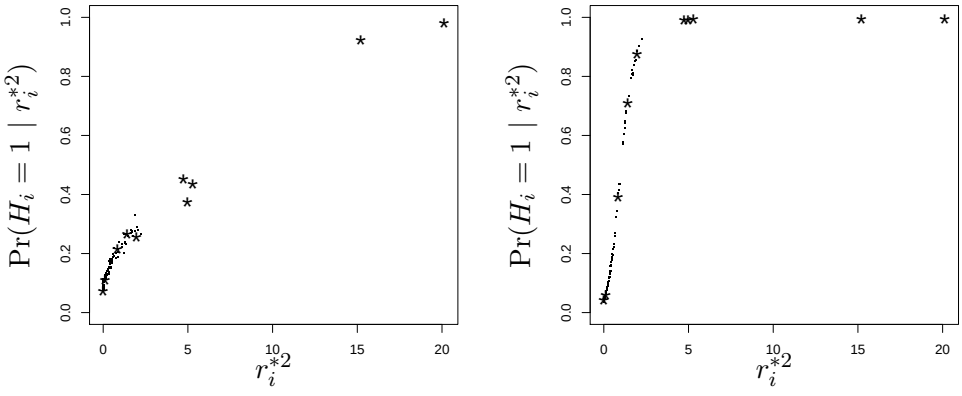
Then I use the proposed method in Algorithm 3.3.1 to calculate the posterior probability $P(H_i = 1 \mid r_i^{*2})$ for those 100 observation. As pointed out by Scott and Berger [88], the Bayes inferences may be sensitive to the choice of the priors. In [88], they suggested to use Beta(a, b) prior with $b = 1$ on the probability of being a null, and in their simulation study, they chose $a = 1$ and $a = 11$ for their model. The former gives a uniform prior for π_0 that means there is no prior information of any outlier, while the latter has a median

of about 0.93. However, those two priors both have a large mass on $\pi_0 = 1$, which does not accord with the problem we are dealing with. Although the proportion of outliers may be small, we expect there to be some outliers. I still use Beta$(11, 1)$ in my simulation as one selection of the prior distribution of π_0. Besides Beta$(11, 1)$, I choose the prior of π_0 to be Beta$(0.8, 0.2)$, Beta$(8, 2)$ and Beta$(80, 20)$ with an equal mean 0.8 and decreasing variances, and Beta$(9.41, 4.03)$ and Beta$(8.09, 8.09)$ that have the same variance as Beta$(8, 2)$ but respectively have mean 0.7 and 0.5. The variance of Beta$(11, 1)$ is between that of Beta$(8, 2)$ and that of Beta$(0.8, 0.2)$. As assumed in the previous section, $\mu_i = 0$ under the null hypothesis but follows a $N(0, V)$ prior distribution under the alternative hypothesis. The alternative μ_i of the simulated dataset is actually simulated from $N(0, V')$ with $V' = 16$ and random errors ε are generated from $N(0, \sigma^2)$ with $\sigma^2 = 1/16$. However, we know neither the value of σ^2 nor V' in the real data. So I also consider different value of V in the prior distributions in order to study the sensitivity of the posterior to this prior choice. The values 9 and 16 of V were used in Scott and Berger [88] as a prior variance of alternatives. I choose $V = 4, 9, 16, 36$ to be the prior variances of the sampled μ_i. For each observation, $n = 1000$ random samples $(\boldsymbol{H}^j_{(i)}, \boldsymbol{\mu}^j, \pi_0^j)$, $j = 1, \cdots, n$, are generated to calculate the posterior probability $P(H_i = 1 \mid r_i^{*2})$. The plots of $P(H_i = 1 \mid r_i^{*2})$ versus r^{*2} for the first dataset by using various priors on π_0 and μ are shown in Figure 3.2 – 3.5. The prior standard deviation of μ_i varies from 6 to 2 in Figure 3.2 – 3.5, where (a) – (f) are for Beta$(11, 1)$, Beta$(8, 2)$, Beta$(0.8, 0.2)$, Beta$(80, 20)$, Beta$(9.41, 4.03)$ and Beta$(8.09, 8.09)$.

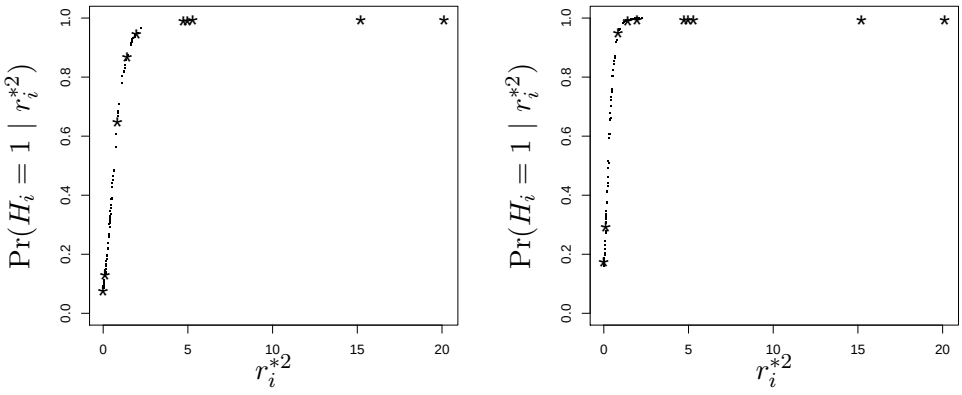
First, I compare the different choice of Beta prior. From Figure 3.2, it can be seen that the posterior probabilities $P(H_i = 1 \mid r^{*2})$ for the five apparent outliers with high values of r^{*2} are greater than those of the other observations for almost all priors. This result can also be observed in Figure 3.3 – 3.5. The posteriors in Figure 3.2 – Figure 3.5, (a) and (c) have a similar pattern and are lower than those in (b), (d), (e) and (f), especially for the three extreme outliers with relatively smaller r^{*2} than the other two extreme observations (between 0.2 and 0.6), but the differences between the posteriors of the two extreme observations with the largest r^{*2} and those of other observations in (a) and (c) are greater than those in (b), (d), (e) and (f). The posteriors in Figure 3.2 – Figure 3.5, (e) and (f) have a similar pattern and are higher than those in (a) – (d). In Figure 3.2 – Figure 3.5 (b) and (d), the curves of posteriors are similar and the five apparent outliers are separated from the other observations because the posteriors of the apparent outliers



(a) Prior $V = 36$, Beta$(11, 1)$ (b) Prior $V = 36$, Beta$(8, 2)$

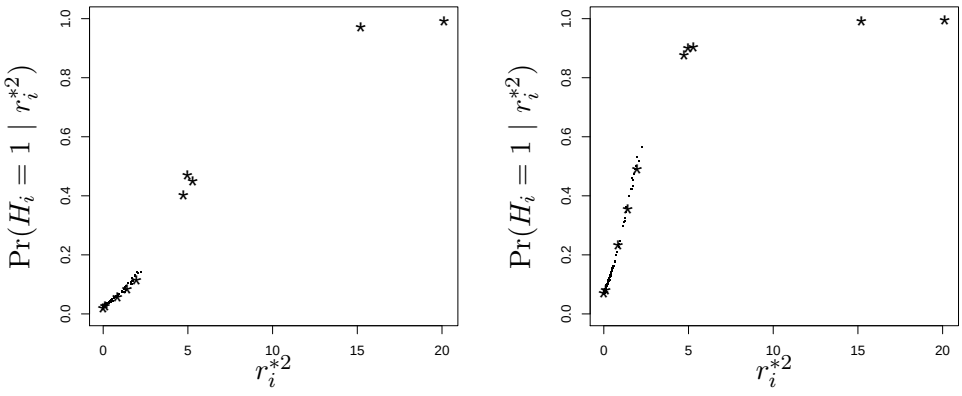


(c) Prior $V = 36$, Beta$(0.8, 0.2)$ (d) Prior $V = 36$, Beta$(80, 20)$

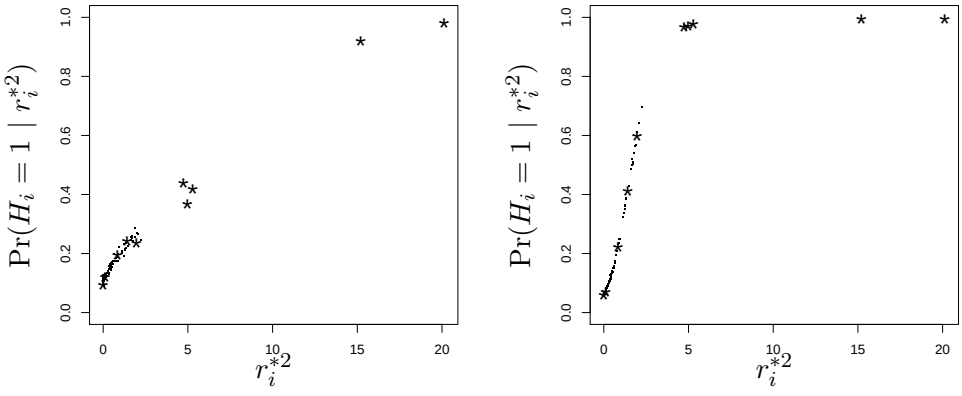


(e) Prior $V = 36$, Beta$(9.41, 4.03)$ (f) Prior $V = 36$, Beta$(8.09, 8.09)$

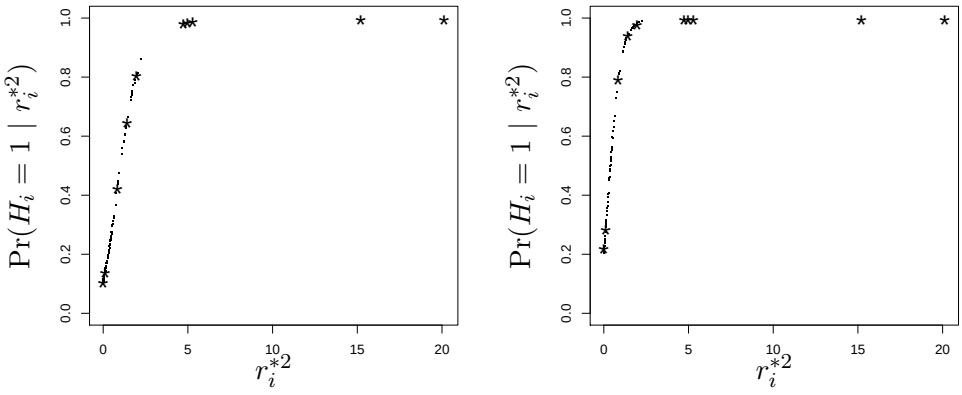
Figure 3.2: The posterior probability $P(H_i = 1 \mid r_i^{*2})$ is plotted as a function of r_i^{*2} for $V = 36$ and six different Beta priors on π_0. The observations are shown as dots $(\cdot)$ for the nulls and stars (*) for the alternatives. The explanatory variable is generated from Bernoulli$(1/2)$.



(a) Prior $V = 16$, Beta$(11, 1)$ (b) Prior $V = 16$, Beta$(8, 2)$

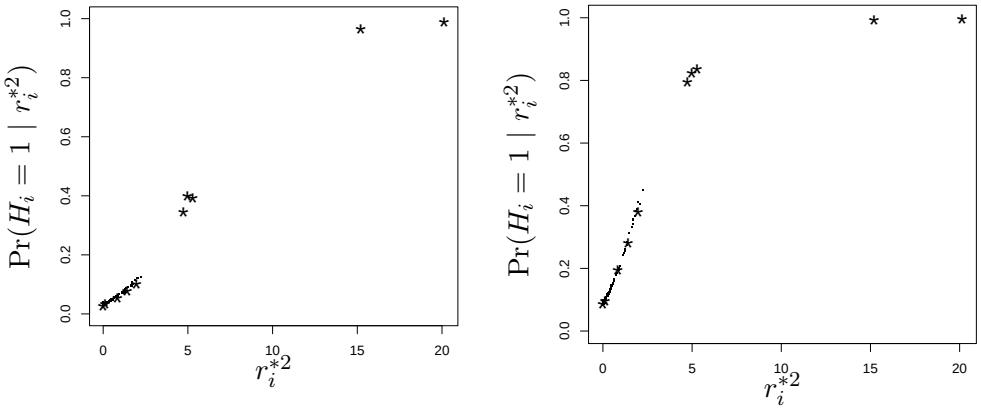


(c) Prior $V = 16$, Beta$(0.8, 0.2)$ (d) Prior $V = 16$, Beta$(80, 20)$

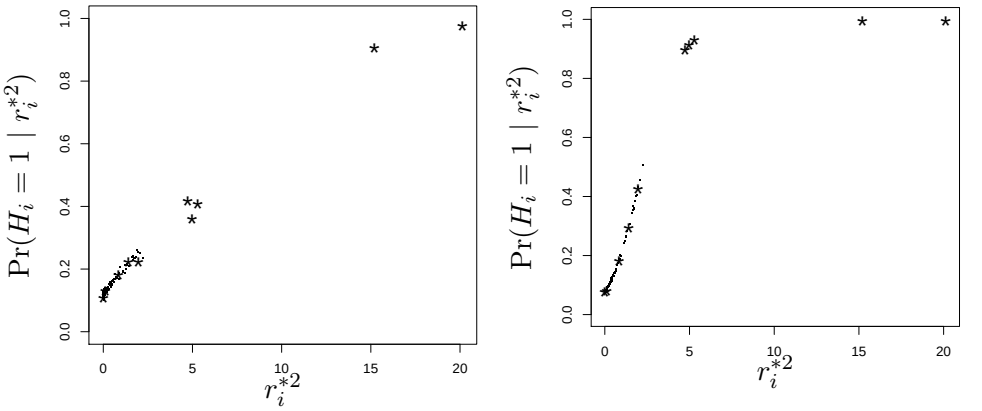


(e) Prior $V = 16$, Beta$(9.41, 4.03)$ (f) Prior $V = 16$, Beta$(8.09, 8.09)$

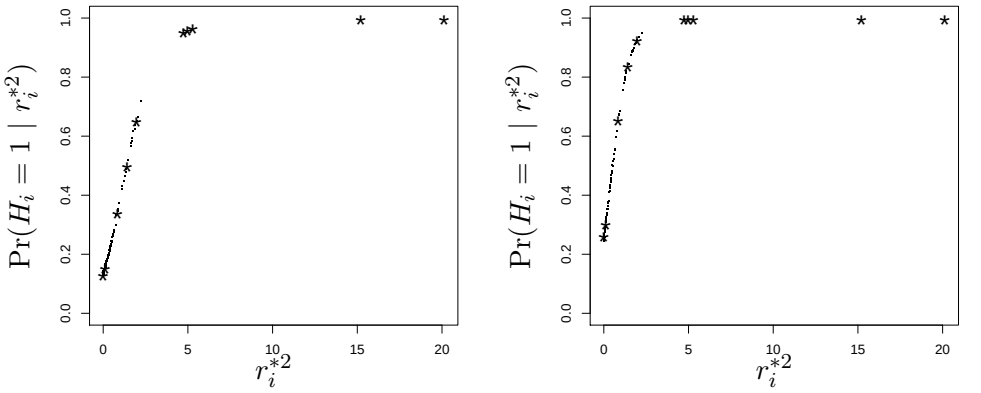
Figure 3.3: The posterior probability $P(H_i = 1 \mid r_i^{*2})$ is plotted as a function of r_i^{*2} for $V = 16$ and six different Beta priors on π_0. The observations are shown as dots $(\cdot)$ for the nulls and stars (*) for the alternatives. The explanatory variable is generated from Bernoulli$(1/2)$.



(a) Prior $V = 9$, Beta$(11, 1)$ (b) Prior $V = 9$, Beta$(8, 2)$

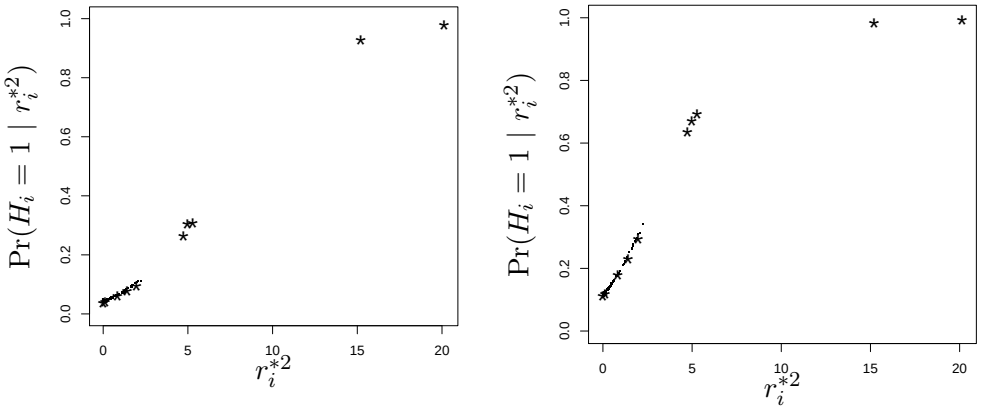


(c) Prior $V = 9$, Beta$(0.8, 0.2)$ (d) Prior $V = 9$, Beta$(80, 20)$

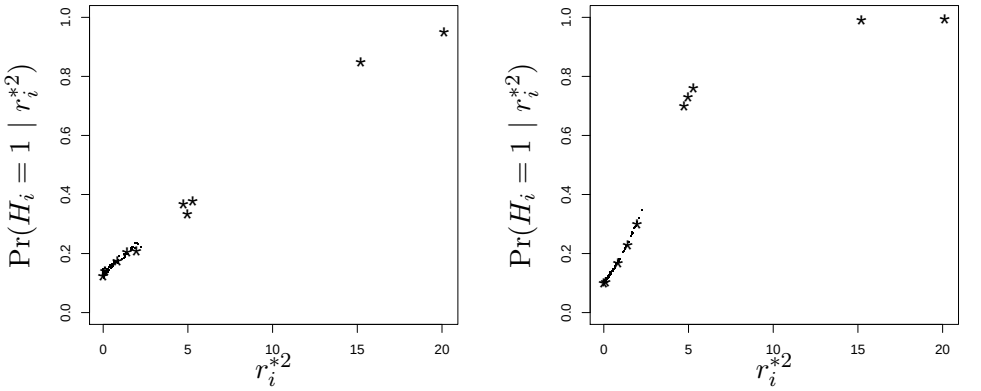


(e) Prior $V = 9$, Beta$(9.41, 4.03)$ (f) Prior $V = 9$, Beta$(8.09, 8.09)$

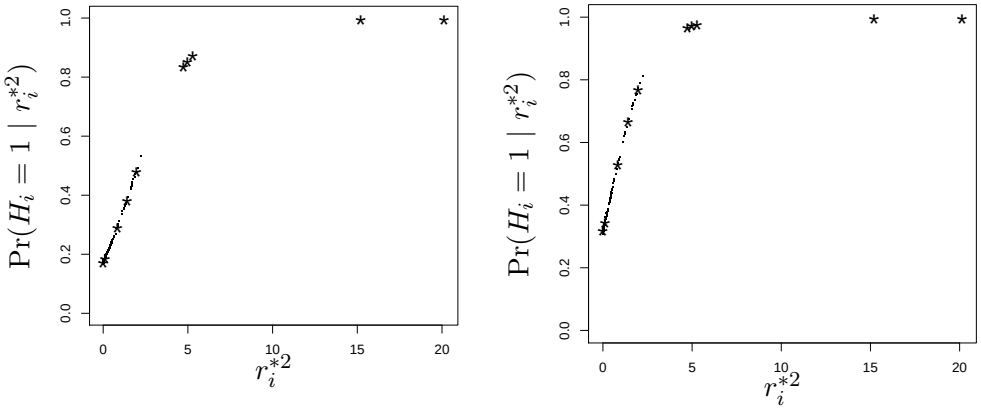
Figure 3.4: The posterior probability$P(H_i = 1 \mid r_i^{*2})$ is plotted as a function of r_i^{*2} for $V = 9$ and six different Beta priors on π_0. The observations are shown as dots $(\cdot)$ for the nulls and stars (*) for the alternatives. The explanatory variable is generated from Bernoulli$(1/2)$.



(a) Prior $V = 4$, Beta$(11, 1)$ (b) Prior $V = 4$, Beta$(8, 2)$



(c) Prior $V = 4$, Beta$(0.8, 0.2)$ (d) Prior $V = 4$, Beta$(80, 20)$



(e) Prior $V = 4$, Beta$(9.41, 4.03)$ (f) Prior $V = 4$, Beta$(8.09, 8.09)$

Figure 3.5: The posterior probability$P(H_i = 1 \mid r_i^{*2})$ is plotted as a function of r_i^{*2} for $V = 4$ and six different Beta priors on π_0. The observations are shown as dots $(\cdot)$ for the nulls and stars (*) for the alternatives. The explanatory variable is generated from Bernoulli$(1/2)$.

are greater than 0.9 and those of the other observations are below 0.8. The Beta prior of Figure 3.2 (c), (b) and (d) has the same mean but decreasing variance. Although the Beta prior with lower variance improves the posteriors of the three extreme observations with relatively smaller r^{*2}, it also increases the posterior of some nulls. A similar trend among (c), (b) and (d) is shown in Figure 3.3 – 3.5. Then I compare Figure 3.2 (b) with (e) and (f), of which the Beta priors have the same variance but decreasing mean. The three plots have similar curves, but the Beta prior with smaller mean produces larger posteriors for all observations. A similar trend of (b), (e) and (f) is shown in Figure 3.3 – 3.5. The Beta prior with smaller variance gives larger posterior for both null and outliers, so does the Beta distribution with smaller mean.

By comparing all (a) – all (f) in Figure 3.2 – 3.5, we can see the curves do not change dramatically across different values of V which indicates that the posterior probability $P(H_i = 1 \mid r^{*2})$ is not affected much by the value of V. As V decreases, all posteriors decrease, the variations in the posteriors of the five apparent observations increase, and the variations in the posteriors of the other observations decrease, though the shape of the posterior curve is not affected noticeably. Then I compare all (b) and all (d) in Figure 3.2 – 3.5, the shape of the posterior curve does not change much and the five apparent outliers are well separated from the other observation in the four plots. The posteriors of the two outliers with the largest deletion residuals change slightly as V varies. Although the smaller V pulls down the posterior probabilities for those observations with high deletion residuals, the posterior probabilities of the observations with low r^{*2} decrease more sharply with smaller V, which means the distance between the extreme observations and the non-extreme observations increases as V decreases. A similar trend can be found for all (e) and (f) in Figure 3.2 – Figure 3.5.

The plots of the posterior probability $P(H_i = 1 \mid r^{*2})$ suggest that the posterior is sensitive to the prior distribution of π_0 and the prior variance of the mean shift μ_i, and the prior of π_0 has a stronger effect on the posterior than the prior of μ_i.

When calculating the posterior probabilities $P(H_i = 1 \mid r_i^{*2})$ above, I used Patnaik's approximation to calculate the density of the doubly noncentral F distribution. In Section 3.4, I developed an algorithm, Algorithm 3.4.2, to compute the actual density of doubly noncentral F distribution. This algorithm is more complicated and time-consuming than Patnaik's approximation for selected quantiles and parameter values (see Section 3.4). In

order to compare Patnaik's approximation with Algorithm 3.4.2, I also use this algorithm to compute the posterior probabilities for the simulated dataset in Figure 3.1. The error bound E in Algorithm 3.4.2 is chosen to be 10^{-10}, but the actual computation error depends on the values of the noncentrality parameters and deletion residuals. Table 3.13 – 3.16 present the comparison between the results calculated by Patnaik's approximation and by Algorithm 3.4.2 for different priors. The values of V are $36, 16, 9, 4$ in Table 3.13 – 3.16. The first column lists various Beta priors; the second column shows the maximum difference between the doubly noncentral F densities calculated by Patnaik's approximation and those by Algorithm 3.4.2; the third column presents the maximum difference between the posterior probabilities computed by Patnaik's approximation and those by Algorithm 3.4.2; the last column presents the maximum computation error on a computer with Intel Pentium D Dual processor of 2.8GHz and 2.79GHz, 2GB of Ram when using Algorithm 3.4.2. The maximum value for each column in Table 3.13 – 3.16 is shown in bold numbers.

$V = 36$			
Beta prior	max difference of densities	max difference of posteriors	max error bound
Beta$(11, 1)$	0.0017	2.36×10^{-5}	1.20×10^{-7}
Beta$(8, 2)$	0.0017	1.30×10^{-5}	1.67×10^{-7}
Beta$(0.8, 0.2)$	0.0017	6.53×10^{-6}	$\mathbf{2.48\times10^{-7}}$
Beta$(80, 20)$	0.0017	1.68×10^{-5}	1.19×10^{-7}
Beta$(9.41, 4.03)$	0.00074	6.37×10^{-6}	1.80×10^{-7}
Beta$(8.09, 8.09)$	**0.071**	1.42×10^{-6}	1.92×10^{-7}

Table 3.13: Comparison between the results calculated by Patnaik's approximation and by Algorithm 3.4.2 for $V = 36$ and various Beta priors. The second column present the maximum difference between the doubly noncentral F densities calculated by Patnaik's approximation and by Algorithm 3.4.2; the third column show the maximum difference between the posterior probabilities computed by Patnaik's approximation and those by Algorithm 3.4.2; the last column present the maximum computation error on a computer with Intel Pentium D Dual processor of 2.8GHz and 2.79GHz, 2GB of Ram when using Algorithm 3.4.2.

The maximum computation error bound is 2.48×10^{-7} when Algorithm 3.4.2 is used in Algorithm 3.3.1 to calculate the posterior probabilities for the simulated dataset in Figure 3.1. The maximum differences of the densities and maximum differences of posteriors in Table 3.13 – 3.16 are not large, with maximums of 0.071 and 6.22×10^{-5}, respectively.

$V = 16$			
Beta prior	max difference of densities	max difference of posteriors	max error bound
$\text{Beta}(11, 1)$	0.0017	2.67×10^{-5}	5.66×10^{-8}
$\text{Beta}(8, 2)$	0.0017	2.89×10^{-5}	7.63×10^{-8}
$\text{Beta}(0.8, 0.2)$	0.0017	8.60×10^{-6}	1.13×10^{-8}
$\text{Beta}(80, 20)$	0.0017	2.57×10^{-5}	5.47×10^{-8}
$\text{Beta}(9.41, 4.03)$	0.0017	1.53×10^{-5}	8.21×10^{-8}
$\text{Beta}(8.09, 8.09)$	0.0017	4.49×10^{-6}	8.76×10^{-8}

Table 3.14: Comparison between the results calculated by Patnaik's approximation and by Algorithm 3.4.2 for $V = 16$ and various Beta priors. The second column present the maximum difference between the doubly noncentral F densities calculated by Patnaik's approximation and by Algorithm 3.4.2; the third column show the maximum difference between the posterior probabilities computed by Patnaik's approximation and those by Algorithm 3.4.2; the last column present the maximum computation error on a computer with Intel Pentium D Dual processor of 2.8GHz and 2.79GHz, 2GB of Ram when using Algorithm 3.4.2.

$V = 9$			
Beta prior	max difference of densities	max difference of posteriors	max error bound
$\text{Beta}(11, 1)$	0.0017	2.38×10^{-5}	3.37×10^{-8}
$\text{Beta}(8, 2)$	0.0017	3.91×10^{-5}	4.42×10^{-8}
$\text{Beta}(0.8, 0.2)$	0.0017	9.06×10^{-6}	6.49×10^{-8}
$\text{Beta}(80, 20)$	0.0017	5.38×10^{-5}	3.18×10^{-8}
$\text{Beta}(9.41, 4.03)$	0.0017	2.52×10^{-5}	4.76×10^{-8}
$\text{Beta}(8.09, 8.09)$	0.0017	9.60×10^{-6}	5.06×10^{-8}

Table 3.15: Comparison between the results calculated by Patnaik's approximation and by Algorithm 3.4.2 for $V = 9$ and various Beta priors. The second column present the maximum difference between the doubly noncentral F densities calculated by Patnaik's approximation and by Algorithm 3.4.2; the third column show the maximum difference between the posterior probabilities computed by Patnaik's approximation and those by Algorithm 3.4.2; the last column present the maximum computation error on a computer with Intel Pentium D Dual processor of 2.8GHz and 2.79GHz, 2GB of Ram when using Algorithm 3.4.2.

$V = 4$			
Beta prior	max difference of densities	max difference of posteriors	max error bound
Beta$(11, 1)$	0.0017	2.79×10^{-5}	1.67×10^{-9}
Beta$(8, 2)$	0.0017	4.1×10^{-5}	2.09×10^{-9}
Beta$(0.8, 0.2)$	0.0017	7.84×10^{-6}	3.02×10^{-8}
Beta$(80, 20)$	0.0017	$\mathbf{6.22\times10^{-5}}$	1.50×10^{-8}
Beta$(9.41, 4.03)$	0.0017	4.73×10^{-5}	2.24×10^{-8}
Beta$(8.09, 8.09)$	0.0017	2.23×10^{-5}	2.39×10^{-8}

Table 3.16: Comparison between the results calculated by Patnaik's approximation and by Algorithm 3.4.2 for $V = 4$ and various Beta priors. The second column present the maximum difference between the doubly noncentral F densities calculated by Patnaik's approximation and by Algorithm 3.4.2; the third column show the maximum difference between the posterior probabilities computed by Patnaik's approximation and those by Algorithm 3.4.2; the last column present the maximum computation error on a computer with Intel Pentium D Dual processor of 2.8GHz and 2.79GHz, 2GB of Ram when using Algorithm 3.4.2.

Patnaik's approximation is always faster than Algorithm 3.3.1. The difference between the CPU time used to calculate the posterior probabilities for the simulated dataset by incorporating Patnaik's approximation with Algorithm 3.3.1 (72.39 seconds) and that by using Algorithm 3.4.2 in Algorithm 3.3.1 (6261.31 seconds) is minimized when $V = 4$ and the Beta prior on π_0 is Beta$(11, 1)$. These results suggest that using Patnaik's approximation in the proposed Bayesian method is acceptable for accuracy and computing time.

Although different priors may result in the same rejections of the observations by choosing an appropriate rejection threshold, the farther the observations are from each other, especially for the extreme observations with high posteriors, the better the decision that can be made, because the observations with very close posterior values may have to be rejected at the same time. In this sense Beta$(8, 2)$, Beta$(80, 20)$ seem to be better than the other Beta priors. In order to compare the performance of the proposed method for different priors, I plot ROC curves and calculate the areas under ROC curves, rather than control an error rate below a specific level and compare the average powers for different choices of priors. In my simulation, the AUC of the proposed Bayes method for simulated datasets is calculated by using the "R" function "roc.area" in the package "verification" which is available on the R website at http://www.r-project.org [65]. As mentioned in Section

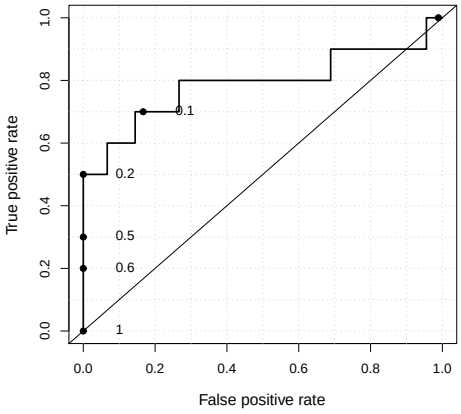
1.4.4, AUC can be interpreted as the probability that the posterior probability of a false null being assigned false is higher than that of a true null being assigned false [46, 62, 69]. Hence, the AUC of the proposed Bayes method is equal to $\Pr[\{p_1|H_i = 1\} > \{p_1|H_i = 0\}]$, where $p_1 = P(H_i = 1|r_*^2)$. AUC can be used to assess the overall performance of a testing method.

The ROC curves for the simulated dataset shown in Figure 3.1 with different values of V and various Beta priors are obtained by using the "R" function "roc.plot" in the "verification" package [65] and are given in Figure 3.6 – 3.9. The value of V, Beta prior and the AUC value for each combination of priors are presented in the subtitles of each graph.

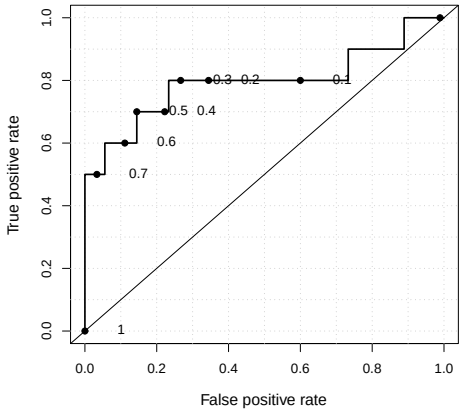
All ROC curves in Figure 3.6 – 3.9 (a) – (f) are similar. The FPRs are the same for large cutoffs in Figure 3.6 – 3.9 (a) and (c), which reflects that most observations have small posterior probabilities as shown in 3.2 – 3.5 (a) and (c), whereas TPRs are similar for small cutoffs in Figure 3.6 – 3.9 (e) and (f), which indicates that the extreme observations have extremely large posteriors as shown in Figure 3.2 – 3.5 (e) and (f). The ROC curves in Figure 3.6 – 3.9 (b) and (d) have less ties than those in the other graphs. As V decreases, the same cutoff tends to give smaller power which reflects the fact that the posteriors decrease as V decreases as shown in Figure 3.2 – Figure 3.5.

Next I compare the AUC for this dataset with various priors. It can be shown that when a method can correctly identify a majority of the typical and atypical observations, a the area under the ROC curve will exceed 0.5 [62]. All AUC values are greater than 0.7, which indicates that the proposed method can identify a majority of the outliers. All values of AUC in Figure 3.6 – 3.9 (a) – (f) are similar, from 0.7833 to 0.8033. The AUC calculated by using Beta$(0.8, 0.2)$ varies from 0.7833 to 0.8033, and has a wider range than it calculated by using other Beta priors. The posterior is more sensitive to Beta$(0.8, 0.2)$ because it has the largest variance among all Beta priors. AUC increases as V decreases in Figure 3.6 – 3.9 except AUC for Beta$(8.09, 8.09)$ decreases from 0.7956 to 0.7933 when V decreases from 9 to 4.

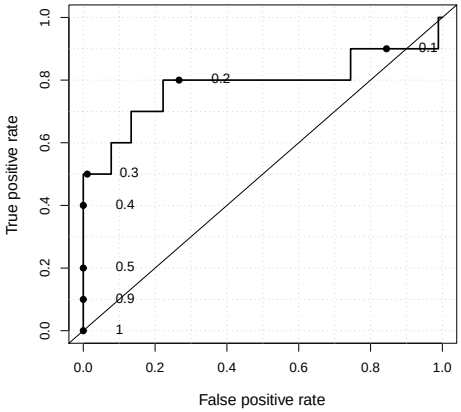
For the first simulated dataset, there is no large difference among the results obtained by using the six Beta priors or by using the four different values of V, though smaller V seems to be slightly better. This may be caused by the error introduced by using Patnaik's approximation since large values of V result in large noncentrality of the distribution of r_i^*.



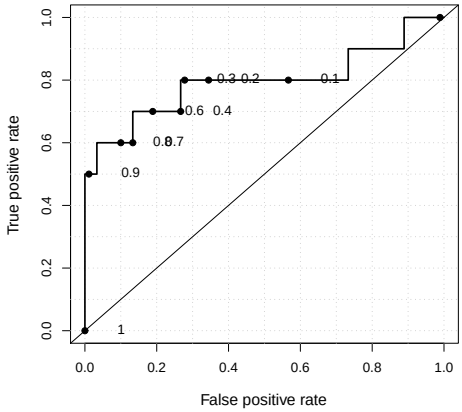
(a) V=36, Beta(11,1), AUC=0.7878



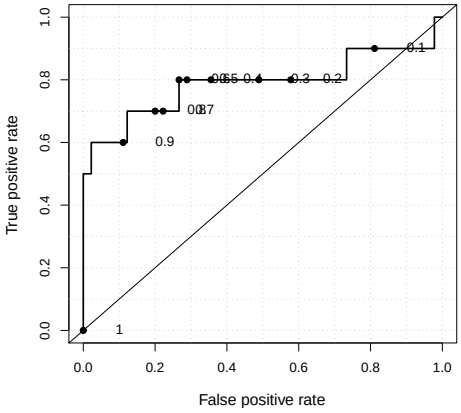
(b) V=36, Beta(8,2), AUC=0.7944



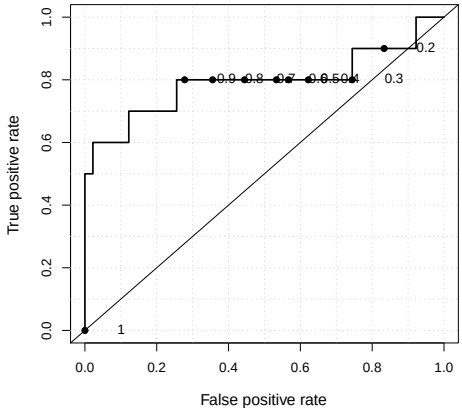
(c) V=36, Beta(0.8,0.2), AUC=0.7833



(d) V=36, Beta(80,20), AUC=0.7944

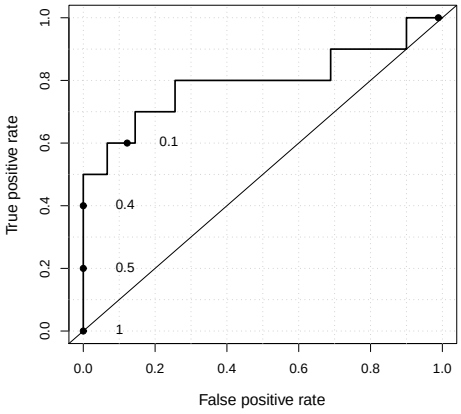


(e) V=36, Beta(9.41,4.03), AUC=0.7878

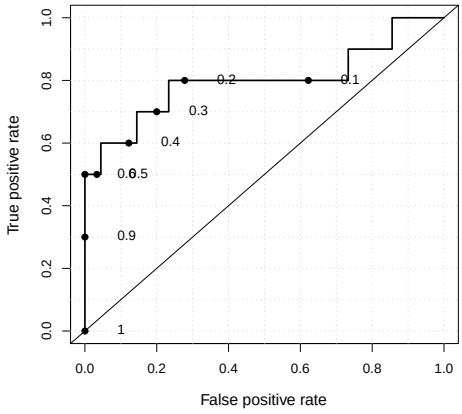


(f) V=36, Beta(8.09,8.09), AUC=0.7933

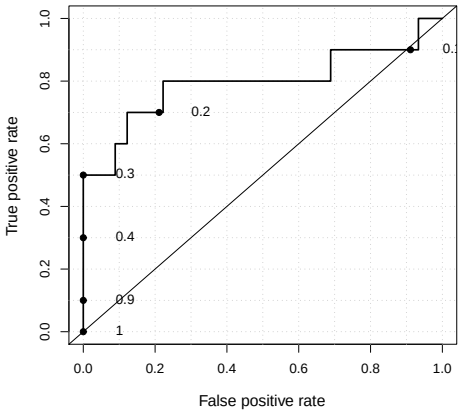
Figure 3.6: True positive rate versus false positive rate for the dataset shown in Figure 3.1 with $V = 36$ and six different Beta priors on π_0. The points are denoted by the corresponding cutting points if applicable, which are from 0 to 1 with increment 0.1. AUC is the area under the curve.



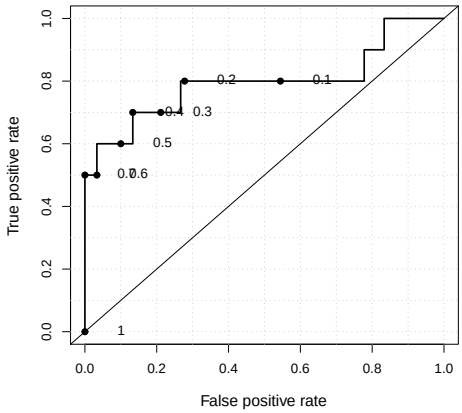
(a) V=16, Beta(11,1), AUC=0.7944



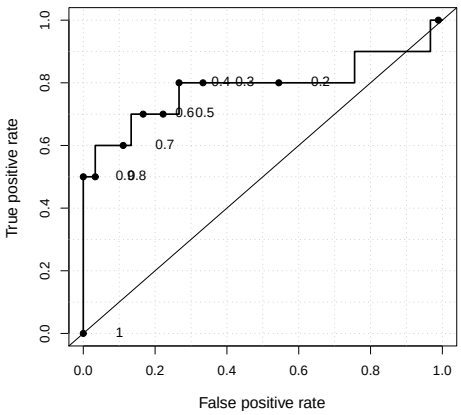
(b) V=16, Beta(8,2), AUC=0.7989



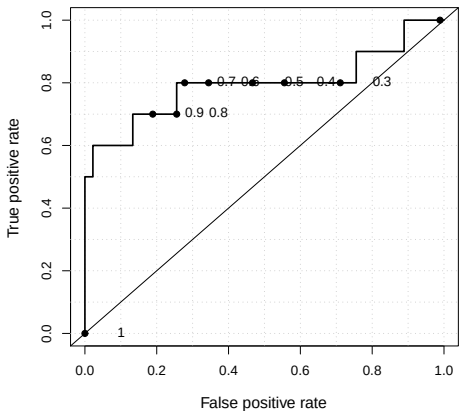
(c) V=16, Beta(0.8,0.2), AUC=0.7944



(d) V=16, Beta(80,20), AUC=0.7956

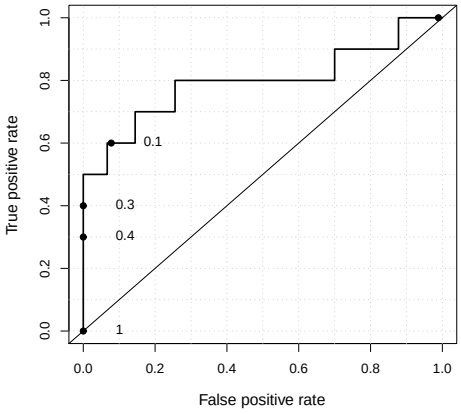


(e) V=16, Beta(9.41,4.03), AUC=0.7844

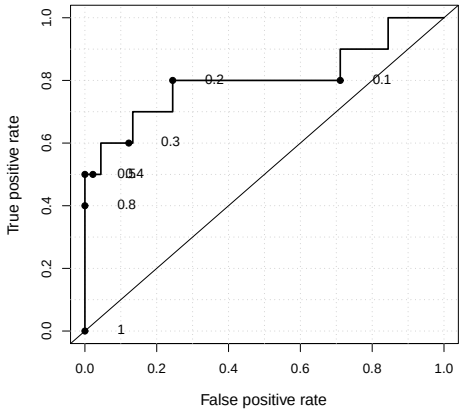


(f) V=16, Beta(8.09,8.09), AUC=0.7944

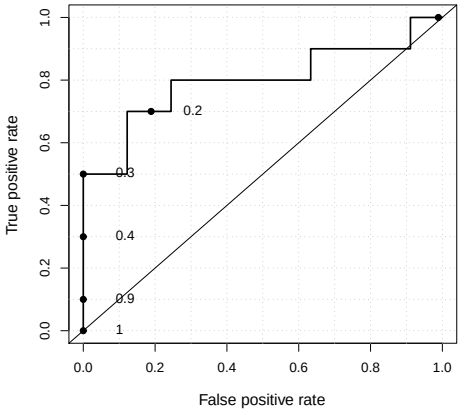
Figure 3.7: True positive rate versus false positive rate for the dataset shown in Figure 3.1 with $V = 16$ and six different Beta priors on π_0. The points are denoted by the corresponding cutting points if applicable, which are from 0 to 1 with increment 0.1. AUC is the area under the curve.



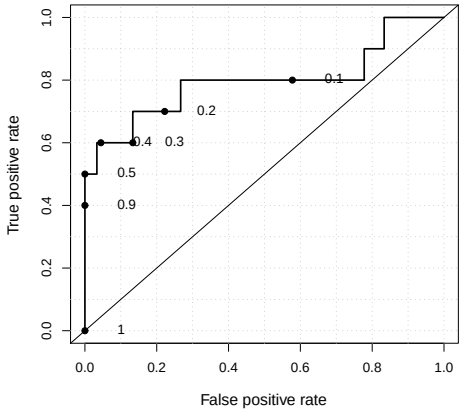
(a) V=9, Beta(11,1), AUC=0.7956



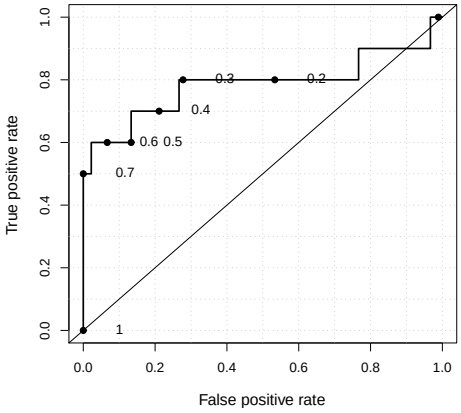
(b) V=9, Beta(8,2), AUC=0.8022



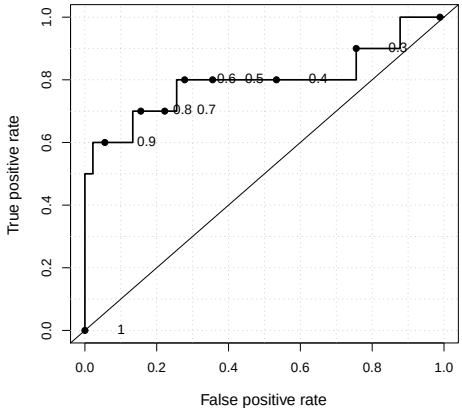
(c) V=9, Beta(0.8,0.2), AUC=0.7967



(d) V=9, Beta(80,20), AUC=0.7956

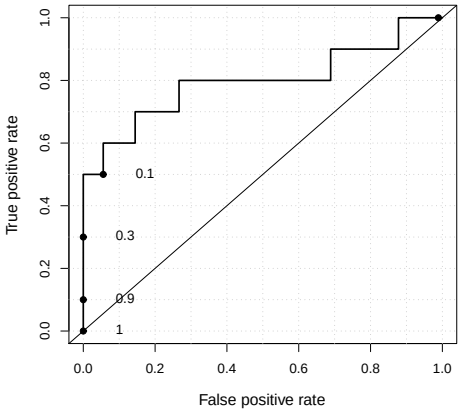


(e) V=9, Beta(9.41,4.03), AUC=0.7844

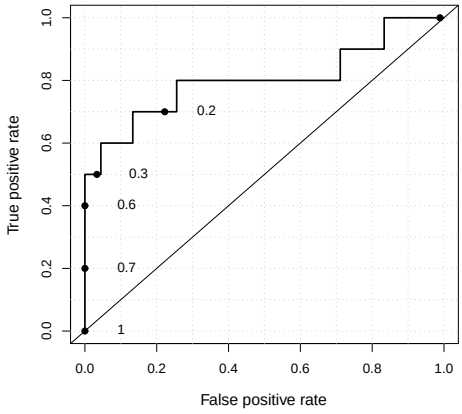


(f) V=9, Beta(8.09,8.09), AUC=0.7956

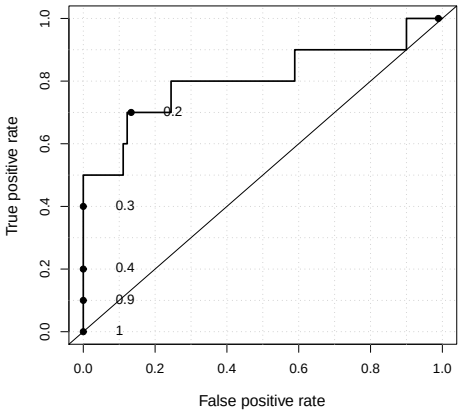
Figure 3.8: True positive rate versus false positive rate for the dataset shown in Figure 3.1 with $V = 9$ and six different Beta priors on π_0. The points are denoted by the corresponding cutting points if applicable, which are from 0 to 1 with increment 0.1. AUC is the area under the curve.



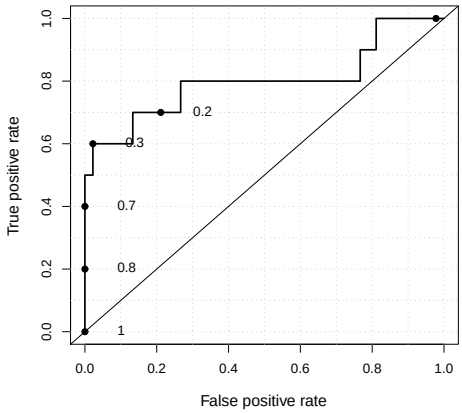
(a) V=4, Beta(11,1), AUC=0.7967



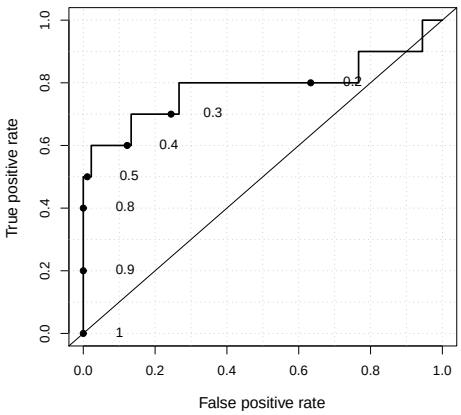
(b) V=4, Beta(8,2), AUC=0.8022



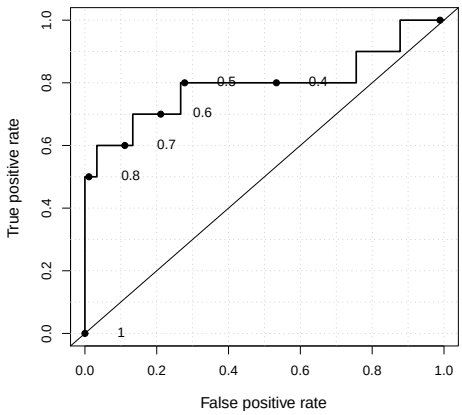
(c) V=4, Beta(0.8,0.2), AUC=0.8033



(d) V=4, Beta(80,20), AUC=0.8



(e) V=4, Beta(9.41,4.03), AUC=0.7867

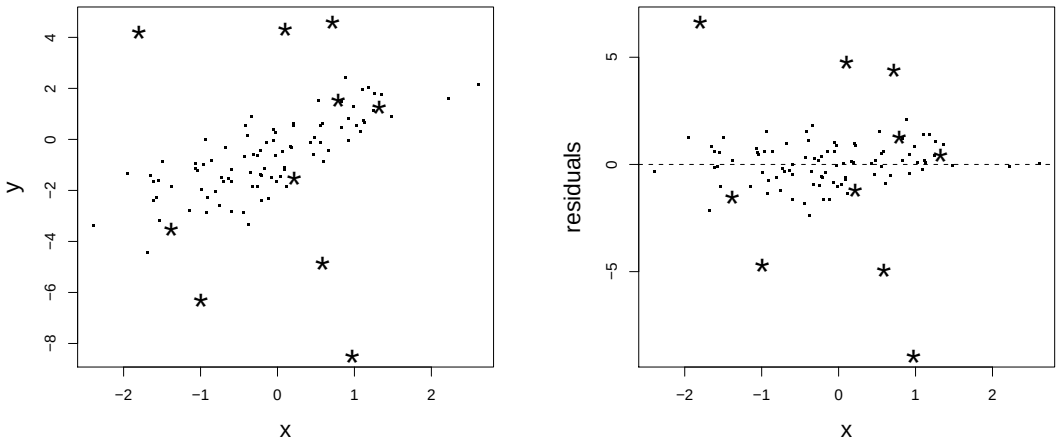


(f) V=4, Beta(8.09,8.09), AUC=0.7933

Figure 3.9: True positive rate versus false positive rate for the dataset shown in Figure 3.1 with $V = 4$ and six different Beta priors on π_0. The points are denoted by the corresponding cutting points if applicable, which are from 0 to 1 with increment 0.1. AUC is the area under the curve.

Also note that the true variance of the mean shifts in the simulated sample is $\sigma^2 V' = 1$.

The results for one simulated dataset might not be true for another one. So I generate another dataset and apply Algorithm 3.3.1 to this dataset. The explanatory variable $\boldsymbol{x} = (x_1, \cdots, x_{100})^T$ and random errors $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_{100})^T$ are independently generated from the standard normal distribution, *i.e.* $\sigma = 1$. Then 10 deviants $\mu_{91}, \cdots, \mu_{100}$ are generated from $N(0, 3)$, *i.e.* $V' = 9$. Then 90 typical observations $y_i = -0.5x_i + \varepsilon_i$, $i = 1, \cdots, 90$, and 10 atypical observations $y_i = -0.5x_i + \sigma\mu_i + \varepsilon_i$, $i = 91, \cdots, 100$. The scatter plot and the residual versus explanatory variable plot of this dataset are shown in Figure 3.10. We can see from the figure that six outliers are apparent outliers, while the other four are hard to distinguished from the nulls.



(a) $\boldsymbol{y}$ vs. $\boldsymbol{x}$ (b) Residual vs. $\boldsymbol{x}$

Figure 3.10: The scatter plot and the residual plot of a dataset with 100 points, of which 10 outliers have mean shifts simulated from N(0,9). The observations are shown as dots ($\cdot$) for the nulls and stars (*) for the alternatives.

Next I use the proposed method in Algorithm 3.3.1 to calculate the posterior probability of being an outlier for those 100 observations. Four values of $V = 4, 9, 16, 36$ are still used for the variances of the sampled μ, and six Beta prior Beta$(11, 1)$, Beta$(0.8, 0.2)$, Beta$(8, 2)$, Beta$(80, 20)$, Beta$(9.41, 4.03)$ and Beta$(8.09, 8.09)$ are used as the prior distributions of π_0. For each observation, $n = 1000$ random samples $(\boldsymbol{H}^j_{(i)}, \boldsymbol{\mu}^j, \pi_0^j)$, $j = 1, \cdots, n$, are generated to calculate the posterior probability $P(H_i = 1 \mid r_i^{*2})$. The plots of $P(H_i = 1 \mid r^{*2})$ versus r^{*2} for the second dataset by using various priors on π_0 and μ are shown in Figure 3.11 – 3.14. The prior standard deviation of μ varies from 6 to 2 in Figure 3.11 – 3.14, where

(a) – (f) are for Beta$(11, 1)$, Beta$(8, 2)$, Beta$(0.8, 0.2)$, Beta$(80, 20)$, Beta$(9.41, 4.03)$ and Beta$(8.09, 8.09)$.

Though the explanatory variables of the two datasets have different distributions, and both σ^2 and V' differ in the two datasets, the curves of $P(H_i = 1 \mid r^{*2})$ versus r^{*2} in Figure 3.11 – 3.14 and are similar to corresponding ones in Figure 3.2 – 3.5. The plots of the second dataset have a similar trend to those of the first dataset. Various values of prior variance V and various Beta priors produce different results, and Beta$(8, 2)$ and Beta$(80, 20)$ are slightly better than the other four Beta priors.

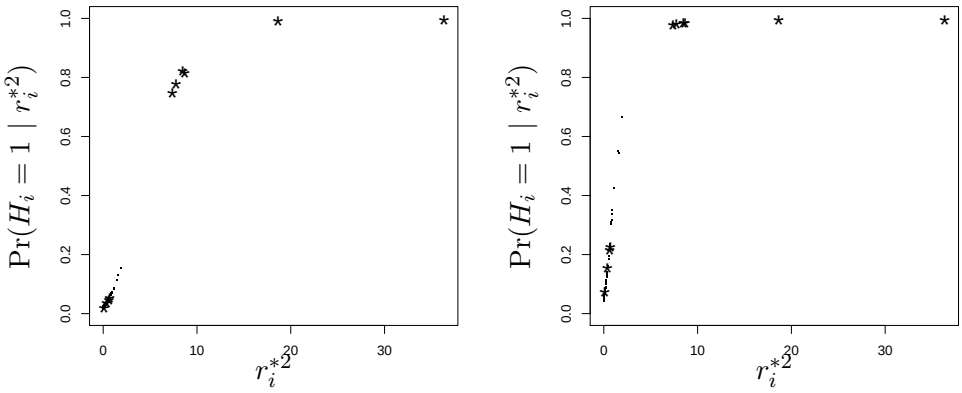
The ROC curves for the simulated dataset shown in Figure 3.10 with different values of V and various Beta prior are presented in Figure 3.15 – 3.18 (a) – (f). The value of V, Beta prior and the AUC value for each combination of priors are given in the subtitles of each graph.

All ROC curves in Figure 3.15 – 3.18 (a) – (f) are similar, and the AUC values are also similar, from 0.8667 to 0.8756. Beta$(0.8, 0.2)$ provides the largest AUC among all six Beta priors for V varying from 4 to 36. The AUC values for the second dataset are greater than those of the first dataset because there are more extreme outliers in the former than in the latter. The trend of AUC of this dataset is different from that of the first dataset. In Figure 3.6 – 3.9, AUC increases as V decreases from 36 to 4 except AUC for Beta$(8.09, 8.09)$. However, AUC decreases as V decreases in 3.15 – 3.18 (c) and (e) with Beta$(0.8, 0.2)$ and Beta$(11, 1)$; AUC of Beta$(8, 2)$ and Beta$(80, 20)$ in (b) and (d) decreases in V; AUC of Beta$(0.8, 0.2)$ and Beta$(8.09, 8.09)$ in (e) and (f) is not monotone in V.

The resulting AUC values for the two simulated datasets are high for various simulation and prior parameters, indicating that the proposed method can identify a majority of the outliers with tolerable error. The plots of ROC curves and the values of AUC indicates that the posterior is not very sensitive to the value of prior variance of the mean shift and the Beta prior of π_0, though there are some difference among the plots of the posterior.

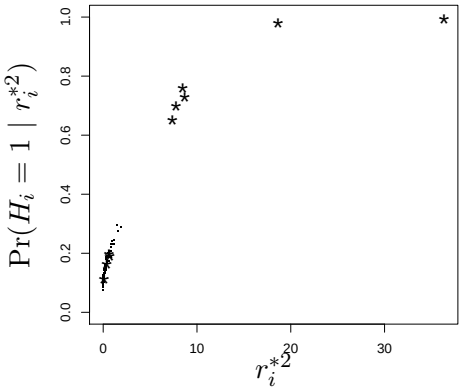
3.5.2 Simulation study of multiple datasets

The results given in the previous section are for two single datasets. I also want to know how sensitive the posterior probability $P(H_i = 1 \mid r^{*2})$ of multiple samples is to the choice of priors. 1000 binary $\boldsymbol{x}$ with $m = 100$ were generated from Benoulli$(1/2)$, and 1000 random errors $\boldsymbol{\varepsilon}$ were generated from $N(0, 1/16)$. Responses $\boldsymbol{y}'$ are calculated as

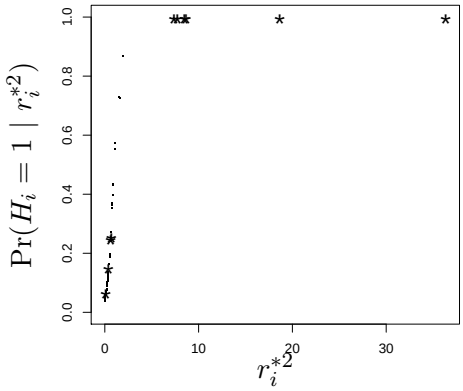


(a) Prior $V = 36$, Beta$(11, 1)$

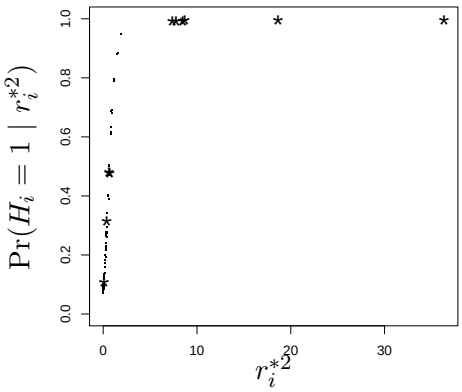
(b) Prior $V = 36$, Beta$(8, 2)$



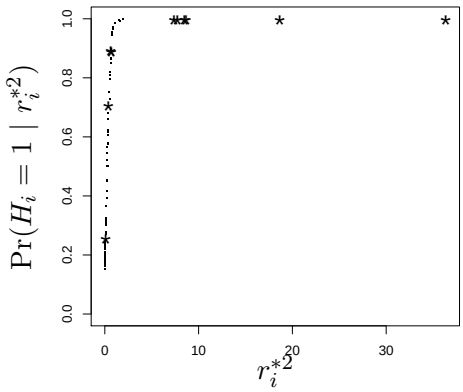
(c) Prior $V = 36$, Beta$(0.8, 0.2)$



(d) Prior $V = 36$, Beta$(80, 20)$

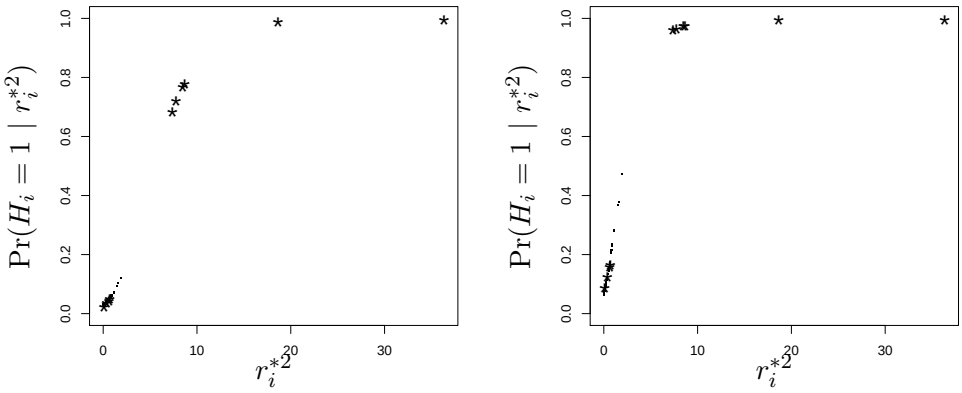


(e) Prior $V = 36$, Beta$(9.41, 4.03)$



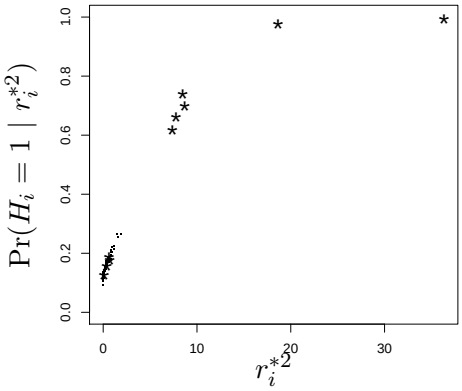
(f) Prior $V = 36$, Beta$(8.09, 8.09)$

Figure 3.11: The posterior probability $P(H_i = 1 \mid r_i^{*2})$ is plotted as a function of r_i^{*2} for $V = 36$ and six different Beta priors on π_0. The observations are shown as dots $(\cdot)$ for the nulls and stars (*) for the alternatives. The explanatory variable is generated from N(0,1).

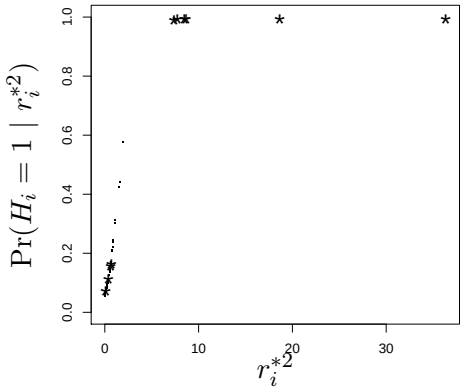


(a) Prior $V = 16$, Beta$(11, 1)$

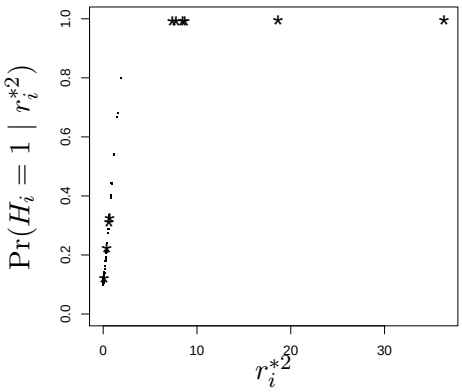
(b) Prior $V = 16$, Beta$(8, 2)$



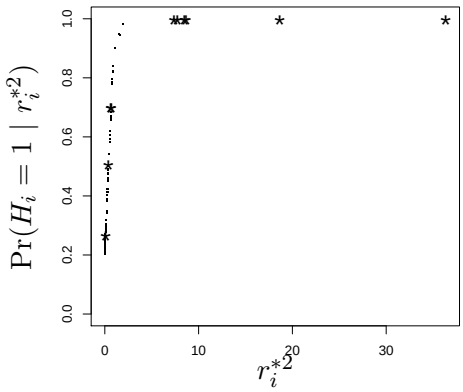
(c) Prior $V = 16$, Beta$(0.8, 0.2)$



(d) Prior $V = 16$, Beta$(80, 20)$

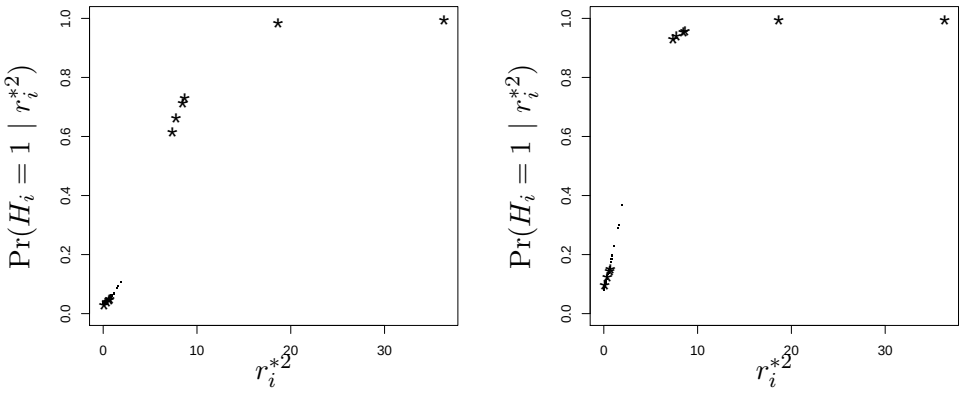


(e) Prior $V = 16$, Beta$(9.41, 4.03)$



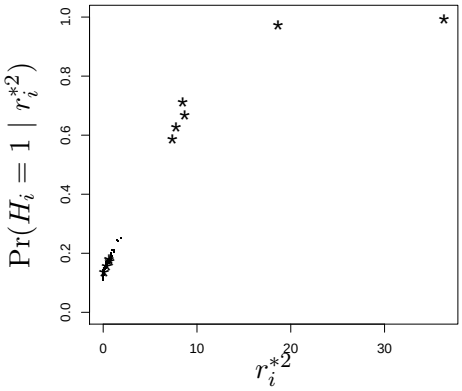
(f) Prior $V = 16$, Beta$(8.09, 8.09)$

Figure 3.12: The posterior probability $P(H_i = 1 \mid r_i^{*2})$ is plotted as a function of r_i^{*2} for $V = 16$ and six different Beta priors on π_0. The observations are shown as dots $(\cdot)$ for the nulls and stars (*) for the alternatives. The explanatory variable is generated from N(0,1).

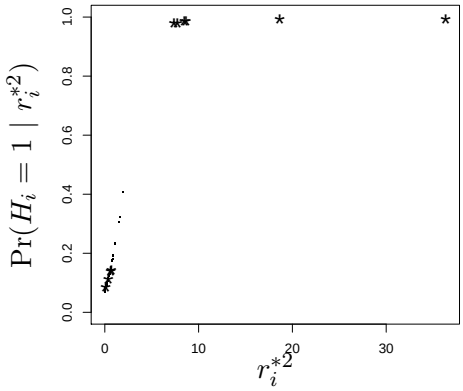


(a) Prior $V = 9$, Beta$(11, 1)$

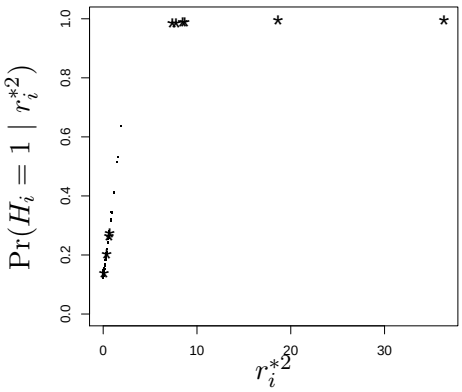
(b) Prior $V = 9$, Beta$(8, 2)$



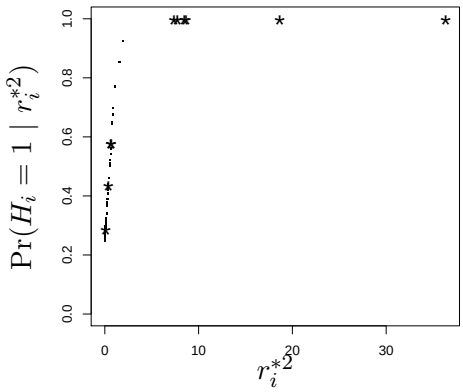
(c) Prior $V = 9$, Beta$(0.8, 0.2)$



(d) Prior $V = 9$, Beta$(80, 20)$

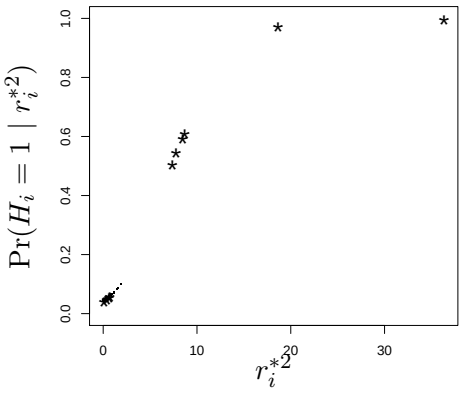


(e) Prior $V = 9$, Beta$(9.41, 4.03)$

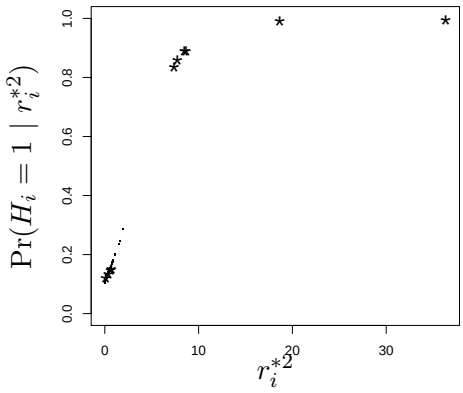


(f) Prior $V = 9$, Beta$(8.09, 8.09)$

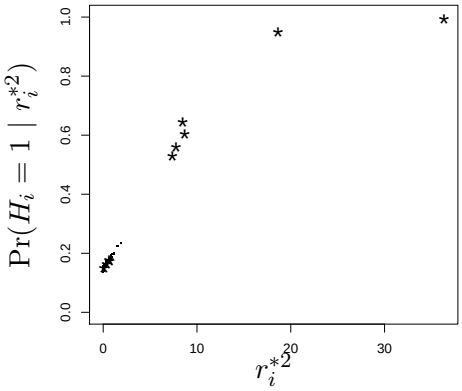
Figure 3.13: The posterior probability $P(H_i = 1 \mid r_i^{*2})$ is plotted as a function of r_i^{*2} for $V = 9$ and six different Beta priors on π_0. The observations are shown as dots $(\cdot)$ for the nulls and stars (*) for the alternatives. The explanatory variable is generated from N(0,1).



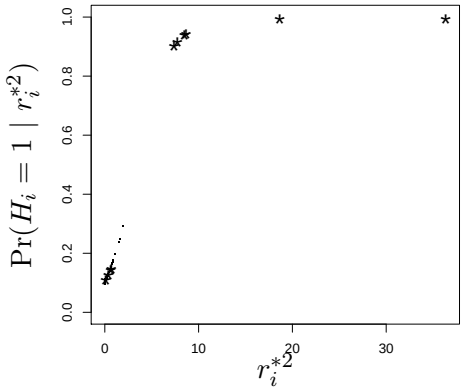
(a) Prior $V = 4$, Beta$(11, 1)$



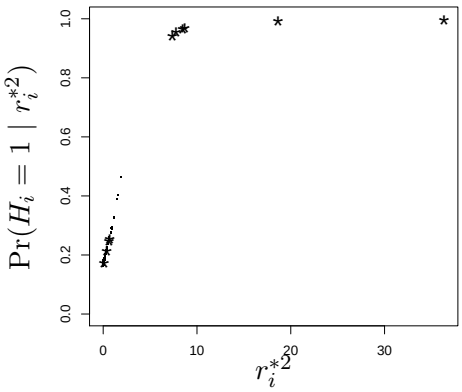
(b) Prior $V = 4$, Beta$(8, 2)$



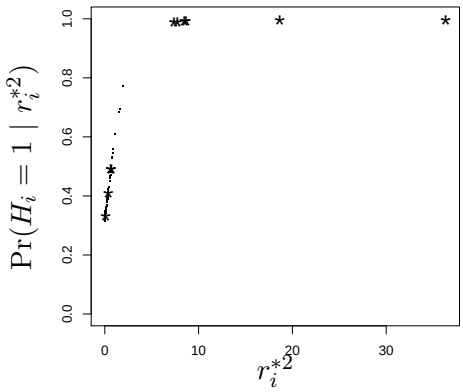
(c) Prior $V = 4$, Beta$(0.8, 0.2)$



(d) Prior $V = 4$, Beta$(80, 20)$

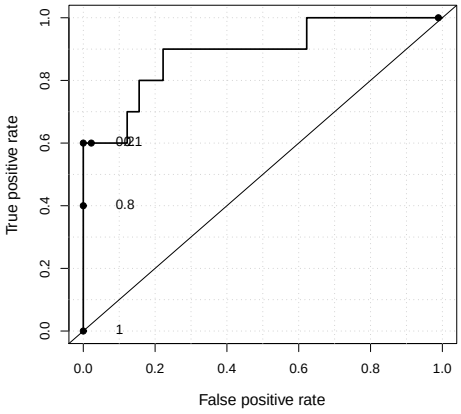


(e) Prior $V = 4$, Beta$(9.41, 4.03)$

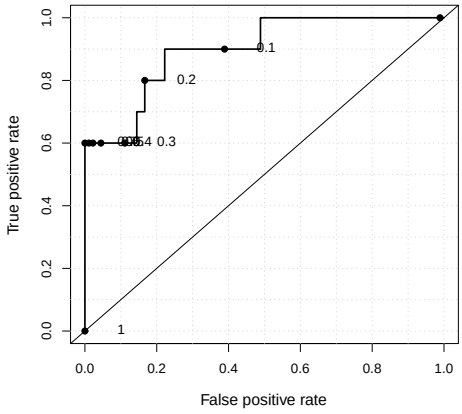


(f) Prior $V = 4$, Beta$(8.09, 8.09)$

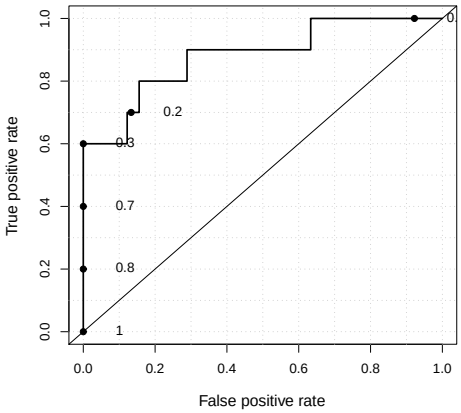
Figure 3.14: The posterior probability $P(H_i = 1 \mid r_i^{*2})$ is plotted as a function of r_i^{*2} for $V = 4$ and six different Beta priors on π_0. The observations are shown as dots $(\cdot)$ for the nulls and stars (*) for the alternatives. The explanatory variable is generated from N(0,1).



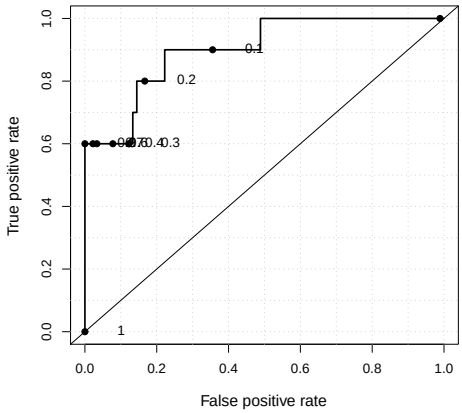
(a) V=36, Beta(11,1), AUC=0.8879



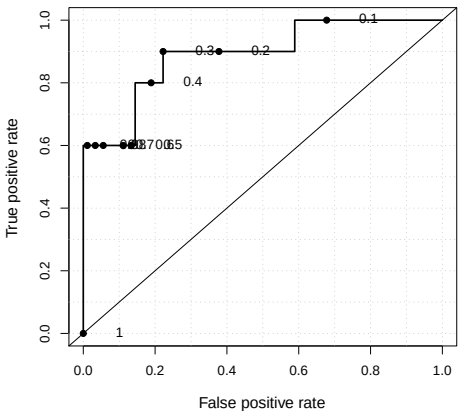
(b) V=36, Beta(8,2), AUC=0.8978



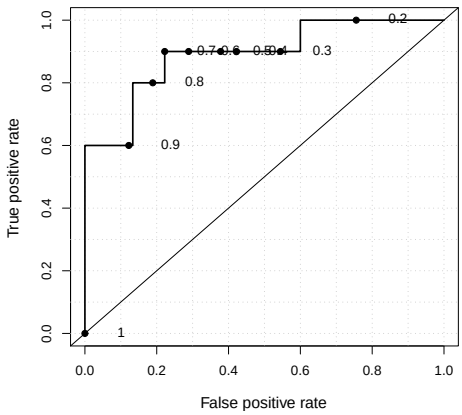
(c) V=36, Beta(0.8,0.2), AUC=0.88



(d) V=36, Beta(80,20), AUC=0.9011

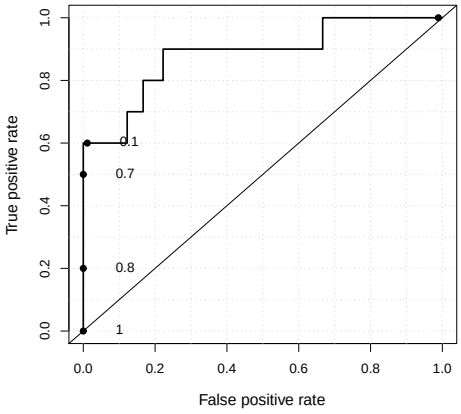


(e) V=36, Beta(9.41,4.03), AUC=0.89

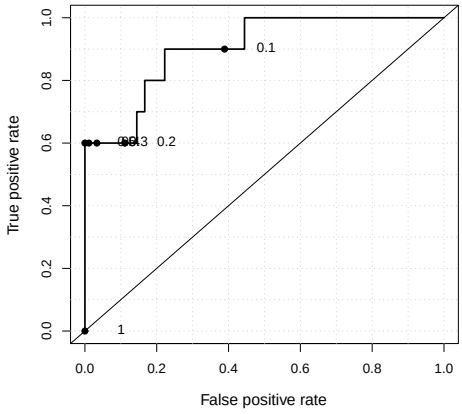


(f) V=36, Beta(8.09,8.09), AUC=0.8911

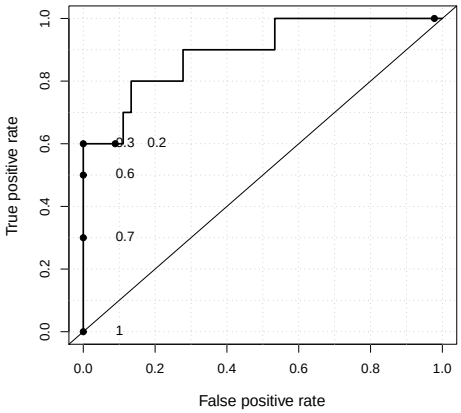
Figure 3.15: True positive rate versus false positive rate for the dataset shown in Figure 3.10 with $V = 36$ and six different Beta priors on π_0. The points are denoted by the corresponding cutting points if applicable, which are from 0 to 1 with increment 0.1. AUC is the area under the curve.



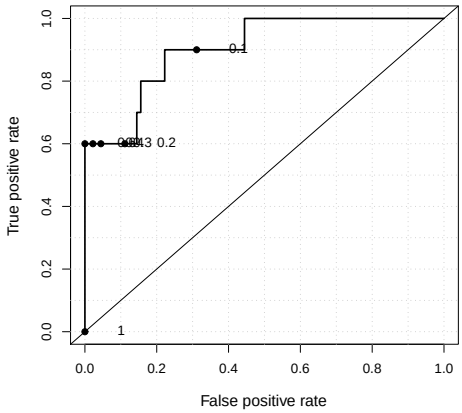
(a) V=16, Beta(11,1), AUC=0.8822



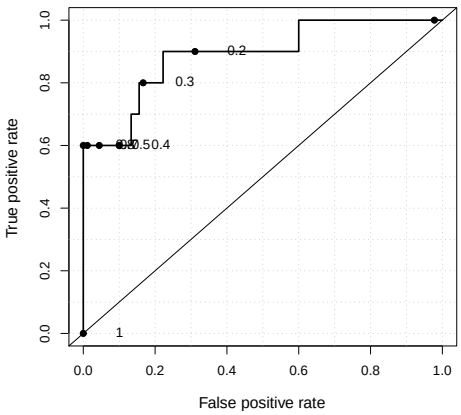
(b) V=16, Beta(8,2), AUC=0.9022



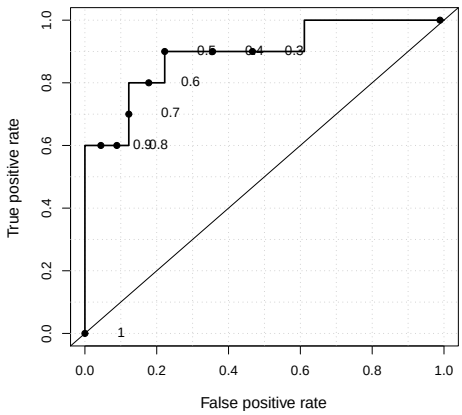
(c) V=16, Beta(0.8,0.2), AUC=0.8944



(d) V=16, Beta(80,20), AUC=0.9033

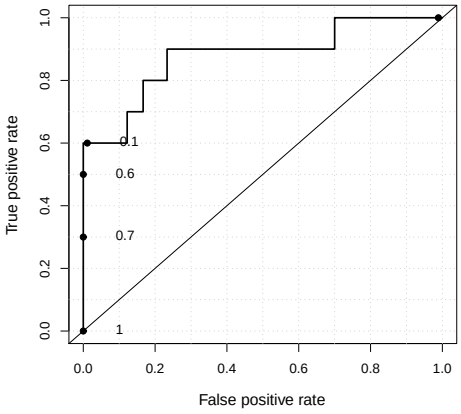


(e) V=16, Beta(9.41,4.03), AUC=0.8889

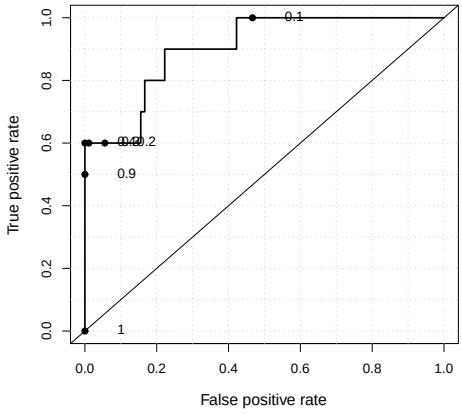


(f) V=16, Beta(8.09,8.09), AUC=0.8922

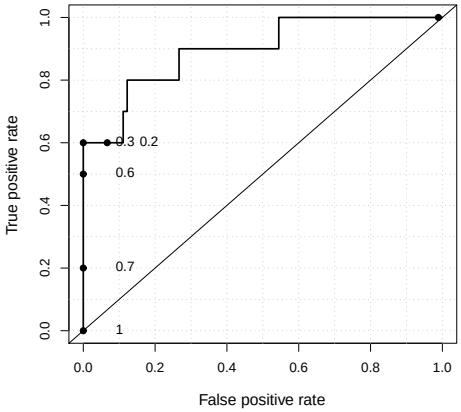
Figure 3.16: True positive rate versus false positive rate for the dataset shown in Figure 3.10 with $V = 16$ and six different Beta priors on π_0. The points are denoted by the corresponding cutting points if applicable, which are from 0 to 1 with increment 0.1. AUC is the area under the curve.



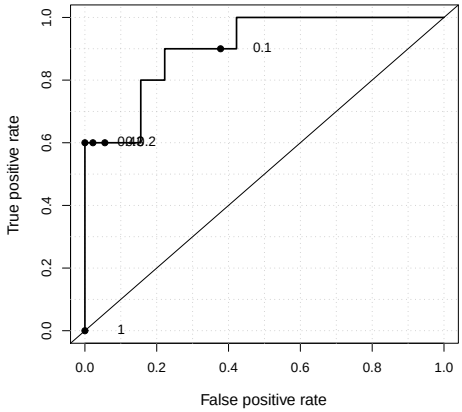
(a) V=9, Beta(11,1), AUC=0.8778



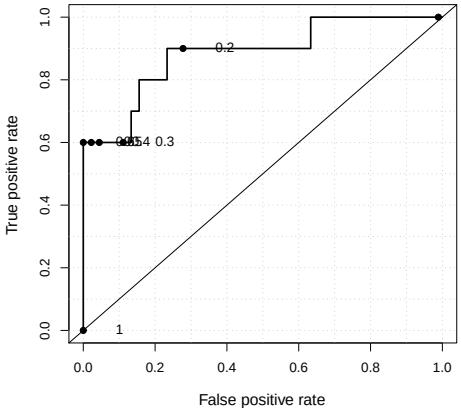
(b) V=9, Beta(8,2), AUC=0.9033



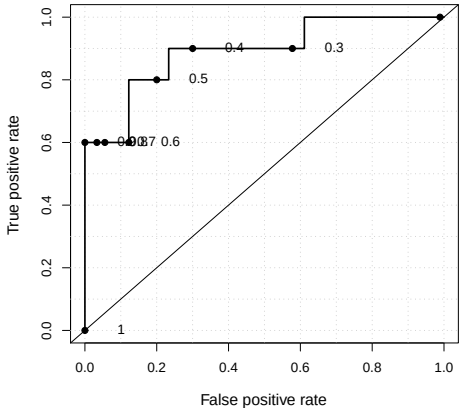
(c) V=9, Beta(0.8,0.2), AUC=0.8956



(d) V=9, Beta(80,20), AUC=0.9044

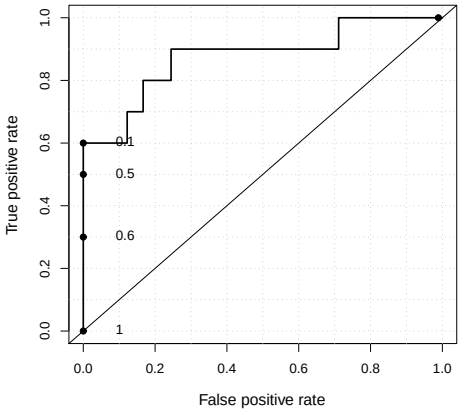


(e) V=9, Beta(9.41,4.03), AUC=0.8844

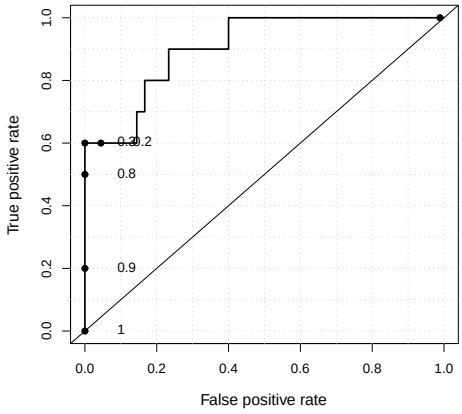


(f) V=9, Beta(8.09,8.09), AUC=0.8911

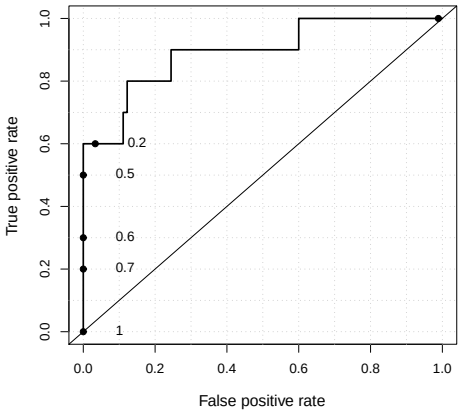
Figure 3.17: True positive rate versus false positive rate for the dataset shown in Figure 3.10 with $V = 9$ and six different Beta priors on π_0. The points are denoted by the corresponding cutting points if applicable, which are from 0 to 1 with increment 0.1. AUC is the area under the curve.



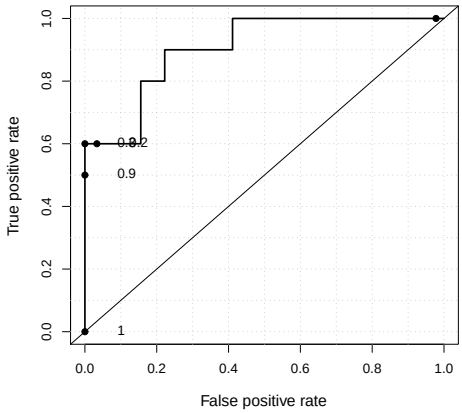
(a) V=4, Beta(11,1), AUC=0.8756



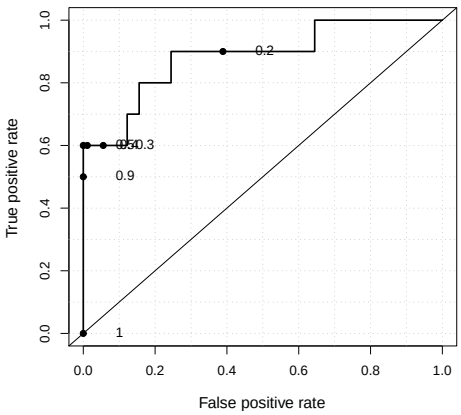
(b) V=4, Beta(8,2), AUC=0.9056



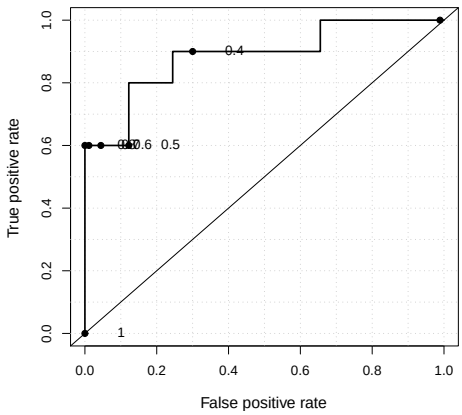
(c) V=4, Beta(0.8,0.2), AUC=0.8922



(d) V=4, Beta(80,20), AUC=0.9056



(e) V=4, Beta(9.41,4.03), AUC=0.8833



(f) V=4, Beta(8.09,8.09), AUC=0.8856

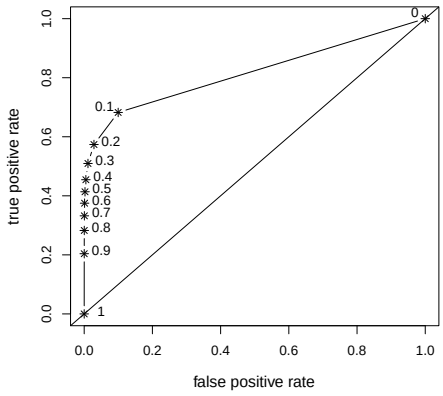
Figure 3.18: True positive rate versus false positive rate for the dataset shown in Figure 3.10 with $V = 4$ and six different Beta priors on π_0. The points are denoted by the corresponding cutting points if applicable, which are from 0 to 1 with increment 0.1. AUC is the area under the curve.

$X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\beta} = (-0.5, 1)$ and all elements of the first column of X equal to 1 and the second column of X equal to $\boldsymbol{x}$. Next 1000 vectors $\boldsymbol{H}$ are generated from Binomial(m, π_0) with $\pi_0 = 0.9$. At last, 1000 sets of scaled mean shifts $\mu_1, \cdots, \mu_{m_1}$ are generated from $N(0, V')$, where m_1 is the number of nonzero elements in $\boldsymbol{H}$ and $V' = 16$, and after being multiplied by $\sigma = 1/4$, each set of deviants is added to the elements of each $\boldsymbol{y}'$ with the corresponding index H_i equal to 1 to generate a vector of responses $\boldsymbol{y}$ with m_1 outliers. Then Algorithm 3.3.1 is applied to these 1000 datasets with $n = 1000$ random sample $(\boldsymbol{H}^j_{(i)}, \boldsymbol{\mu}^j, \pi^j_0)$, $j = 1, \cdots, n$, generated for one observation and one combination of priors. The prior of π_0 is chosen to be Beta$(11, 1)$, Beta$(0.8, 0.2)$, Beta$(8, 2)$, Beta$(80, 20)$, Beta$(9.41, 4.03)$ and Beta$(8.09, 8.09)$. The prior variance of μ_i varies from $4, 9, 16$ to 36.

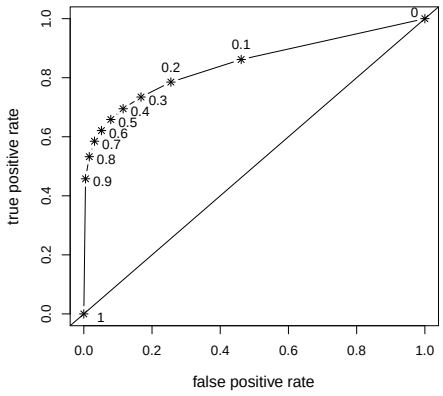
In my simulation, a ROC curve is drawn for each Beta prior and each V, and selected cutting points $0, 0.1, 0.2, \cdots, 1$. For each combination of priors and each cutting points, the average of the true positive rate and the average of the false positive rate over 1000 iteration are calculated and plotted. The plots of the true positive rate averaged over 1000 datasets versus the false positive rate averaged over 1000 datasets for various priors on π_0 and μ are presented in Figure 3.19 – 3.22 (a) to (f). The prior variance of μ varies from 36 to 4 in Figure 3.19 – 3.22, where plots (a) – (f) are for six different Beta priors on π_0. The average area under the ROC curve over 1000 iterations for each combination of priors on π_0 and μ is calculated, with AUC for each iteration computed by using the "R" package "verification" [65], and the average AUC is given in the caption of each plot in Figure 3.21.

All Beta priors result in close values of AUC, which are greater than 0.8, in Figure 3.19 – 3.22. For each value of V, the AUC values of all Beta priors are close to each other. There is no one Beta prior that provides a uniformly larger AUC than the other Beta priors for all values of V. As V increases, AUC increases, except for the AUC values of Beta$(0.8, 0.2)$, Beta$(80, 20)$ and Beta$(9.41, 4.03)$ which decrease slightly (about 0.0001) as V increase from 16 to 36 in Figure 3.19 – 3.20, (c)–(e). Moreover, AUC increases slower as V become larger. The AUC of Beta$(0.8, 0.2)$, which has smallest variation among all Beta priors, increases faster than that of other Beta priors. This observation indicates that the weakest Beta prior is more sensitive to the prior variance of the outliers. The AUC values of (b), (e), (d), (f) in Figure 3.19 and Figure 3.20 are almost the same, which indicates that V does not need to be greater than 36.

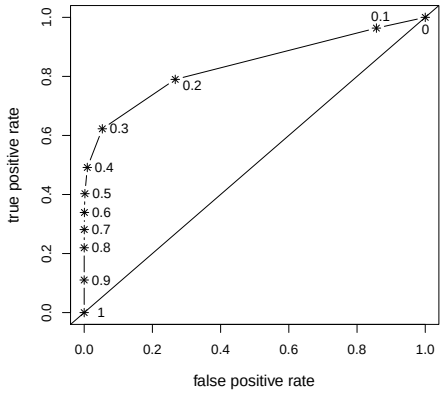
The ROC curves are similar in (b) and (d) of Figure 3.19 – 3.22, where the true positive



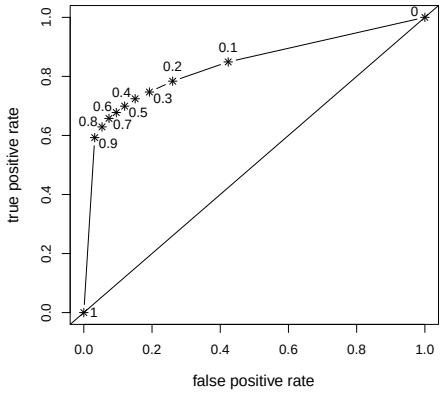
(a) V=36, Beta(11,1), AUC=0.8444



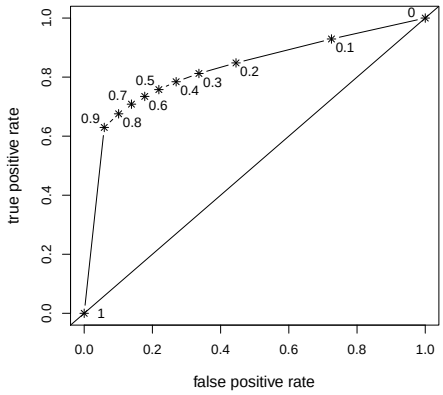
(b) V=36, Beta(8,2), AUC=0.8459



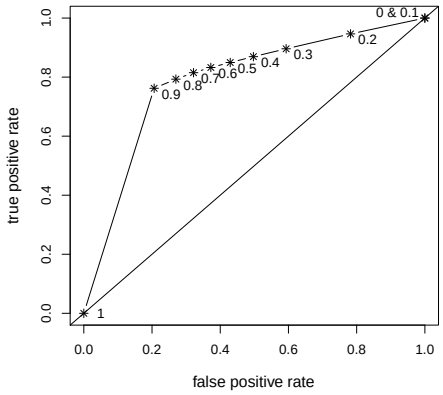
(c) V=36, Beta(0.8,0.2), AUC=0.8473



(d) V=36, Beta(80,20), AUC=0.8461

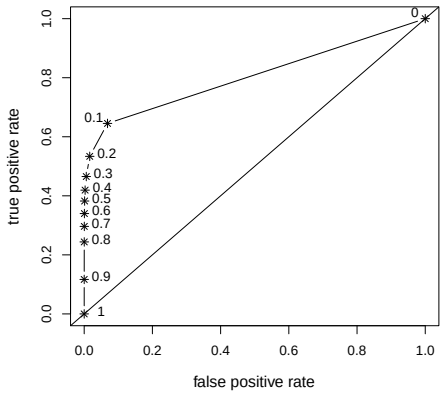


(e) V=36, Beta(9.41,4.03), AUC=0.8416

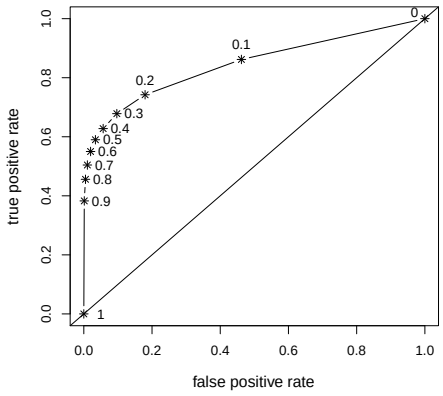


(f) V=36, Beta(8.09,8.09), AUC=0.8479

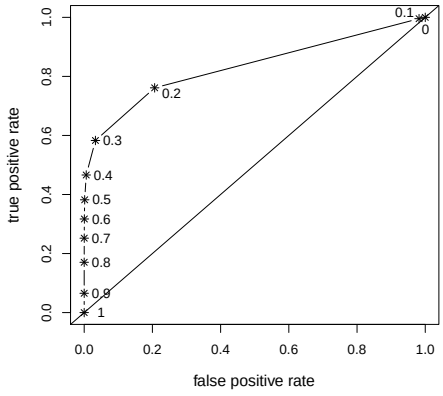
Figure 3.19: True positive rate averaged over 1000 datasets versus false positive rate averaged over 1000 datasets for $V = 36$ and six different Beta priors on π_0. The points are denoted by the corresponding cutting points, which are from 0 to 1 with increment 0.1. AUC is the area under the curve.



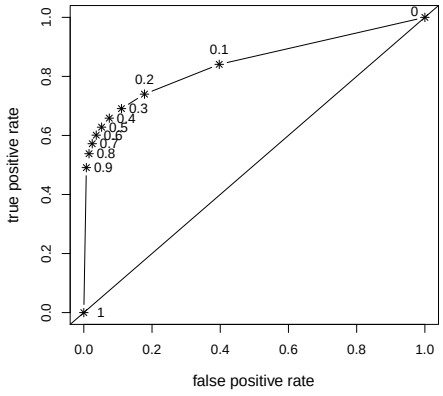
(a) V=16, Beta(11,1), AUC=0.8444



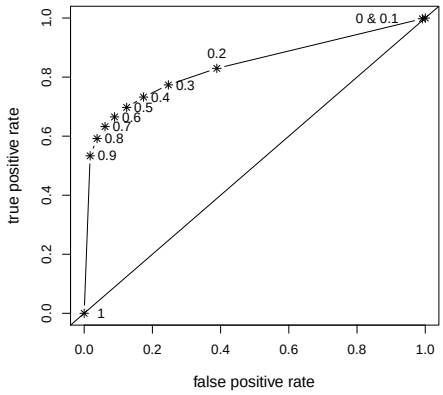
(b) V=16, Beta(8,2), AUC=0.8456



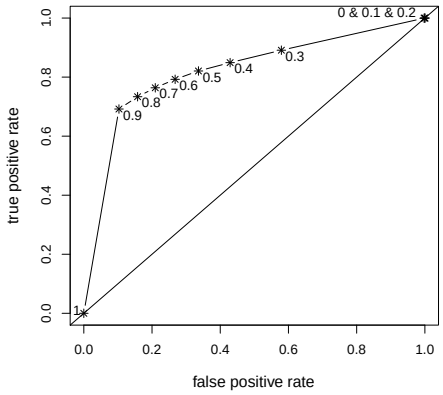
(c) V=16, Beta(0.8,0.2), AUC=0.8474



(d) V=16, Beta(80,20), AUC=0.8462

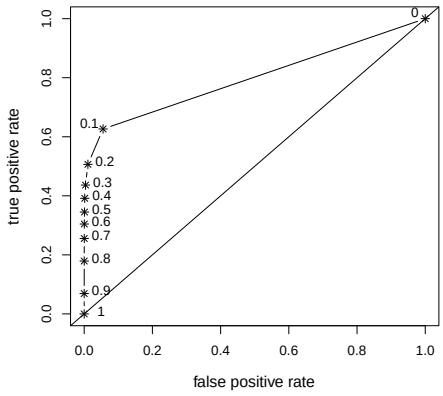


(e) V=16, Beta(9.41,4.03), AUC=0.8417

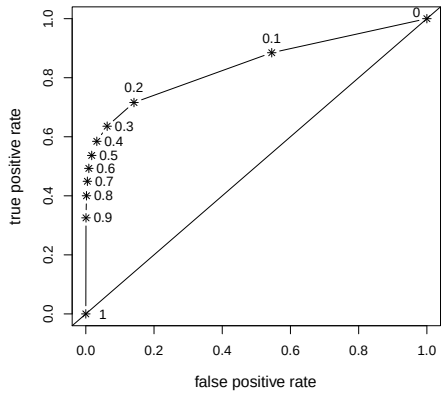


(f) V=16, Beta(8.09,8.09), AUC=0.8479

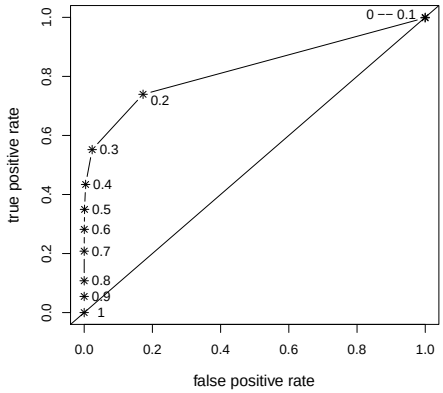
Figure 3.20: True positive rate averaged over 1000 datasets versus false positive rate averaged over 1000 datasets for $V = 16$ and six different Beta priors on π_0. The points are denoted by the corresponding cutting points, which are from 0 to 1 with increment 0.1. AUC is the area under the curve.



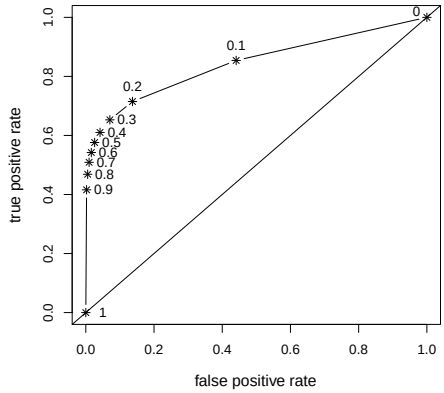
(a) V=9, Beta(11,1), AUC=0.8430



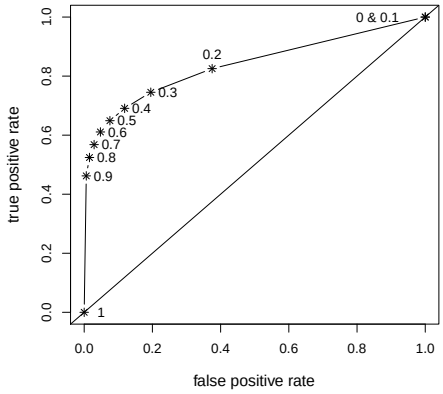
(b) V=9, Beta(8,2), AUC=0.8451



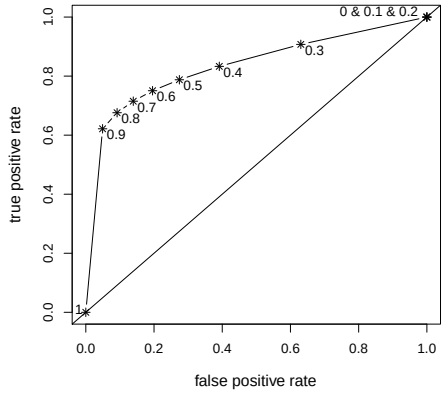
(c) V=9, Beta(0.8,0.2), AUC=0.8449



(d) V=9, Beta(80,20), AUC=0.8461

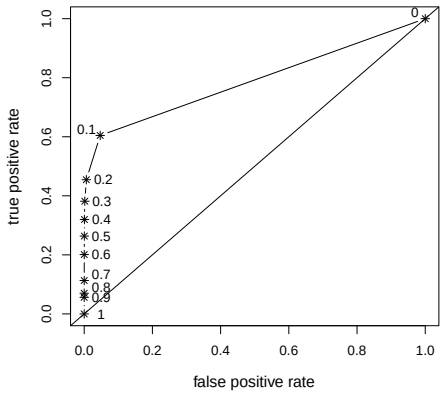


(e) V=9, Beta(9.41,4.03), AUC=0.8416

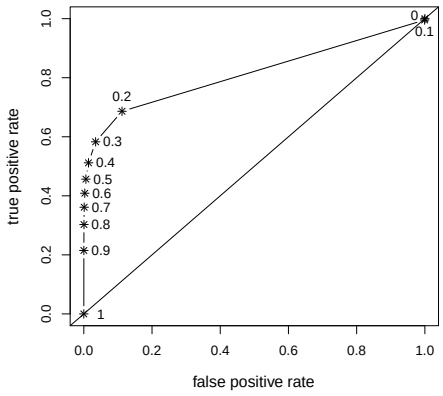


(f) V=9, Beta(8.09,8.09), AUC=0.8439

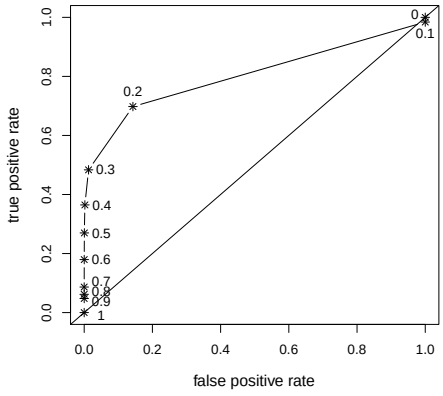
Figure 3.21: True positive rate averaged over 1000 datasets versus false positive rate averaged over 1000 datasets for $V = 9$ and six different Beta priors on π_0. The points are denoted by the corresponding cutting points, which are from 0 to 1 with increment 0.1. AUC is the area under the curve.



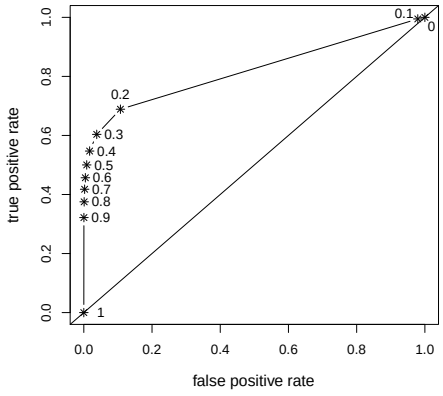
(a) V=4, Beta(11,1), AUC=0.8362



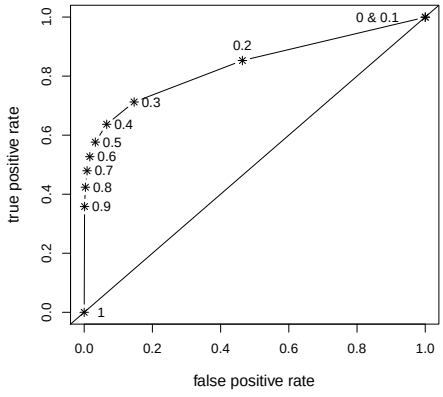
(b) V=4, Beta(8,2), AUC=0.8399



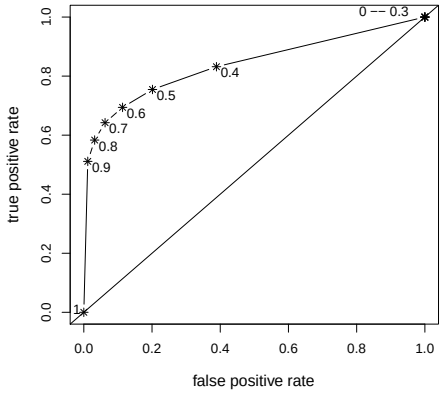
(c) V=4, Beta(0.8,0.2), AUC=0.8261



(d) V=4, Beta(80,20), AUC=0.8460



(e) V=4, Beta(9.41,4.03), AUC=0.8403



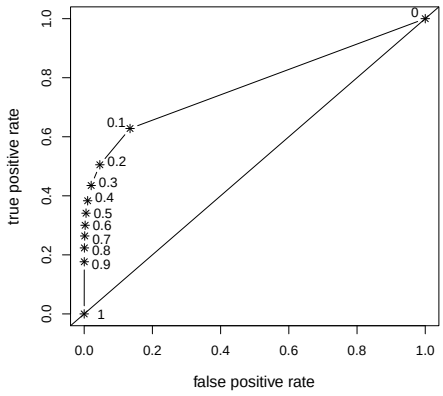
(f) V=4, Beta(8.09,8.09), AUC=0.8438

Figure 3.22: True positive rate averaged over 1000 datasets versus false positive rate averaged over 1000 datasets for $V = 4$ and six different Beta priors on π_0. The points are denoted by the corresponding cutting points, which are from 0 to 1 with increment 0.1. AUC is the area under the curve.

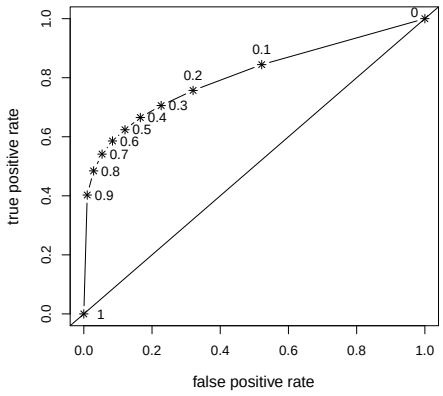
rate increases faster than the false positive rate as cutting points decreases from 0.9 to 0.3, though the true positive rate is more sensitive to the cutting points in (b) than in (d). The ROC curves are similar in (a) and (c) of Figure 3.19 – 3.22, where the false positive rate are close to 0 for large cutting points. The ROC curves in (e) and (f) of Figure 3.19 – 3.22 are similar, where the false positive rate increases faster than the true positive rate as the cutting points decrease from 0.9 to 0.2, though the true positive rate is more sensitive to the cutting points in (e) than in (f). The mean of the Beta prior in (b), (e), (f) is decreasing. Although the Beta prior with smaller mean leads to a greater TPR, it also results in a greater FPR. The variance of the Beta prior in (c), (b), (d) is decreasing. Although the Beta prior with larger variance leads to a greater TPR, it also results in a greater FPR. However, both AUC and the approximated ROC curves suggest that there is no remarkable difference using priors Beta$(11, 1)$, Beta$(0.8, 0.2)$, Beta$(8, 2)$, Beta$(80, 20)$, Beta$(9.41, 4.03)$ or Beta$(8.09, 8.09)$ on π_0 and using prior variance $V = 4, 9, 16$ or 36 for $m = 100$, $\pi_0 = 0.9$, $\sigma = 1/4$ and $V' = 16$.

Next I generate another 1000 datasets from different distribution with different parameters to study the effects of different priors on the posterior probability $P(H_i = 1 \mid r^{*2})$. First 1000 vectors $\boldsymbol{x}$ with $m = 100$ are generated from $N(0, 1)$, and 1000 random errors $\boldsymbol{\varepsilon}$ are also generated independently from $N(0, 1)$. Then 1000 vectors $\boldsymbol{H}$ are generated from Binomial(m, π_0) with $\pi_0 = 0.9$. At last, 1000 sets of deviants $\mu_1, \cdots, \mu_{m_1}$ are generated from $N(0, V')$, where $m_1 = 10$ is the number of nonzero elements in $\boldsymbol{H}$ and $V' = 3$. Note that, in this simulation, the variance of the mean shift is 3 times larger than the variance of the random errors, while the former is 4 times larger than the latter in the previous simulation. If $H_i = 0$, then $y_i = -0.5x_i + \varepsilon_i$, otherwise $y_i = -0.5x_i + \sigma\mu_i + \varepsilon_i$. Then Algorithm 3.3.1 is applied to these 1000 datasets with $n = 1000$ importance samples generated for one observation and one choice of priors. The prior of π_0 is still chosen to be Beta$(11, 1)$, Beta$(0.8, 0.2)$, Beta$(8, 2)$, Beta$(80, 20)$, Beta$(9.41, 4.03)$ and Beta$(8.09, 8.09)$. The prior variance of μ_i varies from $4, 9, 16$ to 36. By using cutting points from 0 to 1 with an increment of 0.1, a ROC curve of average TPR versus average FPR is drawn for each Beta prior and each V, where average TPR and FPR are calculated over 1000 iterations. The plots of the ROC curve for various priors on π_0 and μ are given in Figure 3.23 – 3.26 (a) to (f).

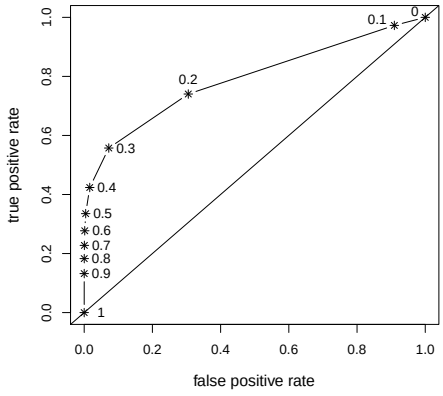
The results in Figure 3.23 – 3.26 are similar to those in Figure 3.19 – 3.22. All the



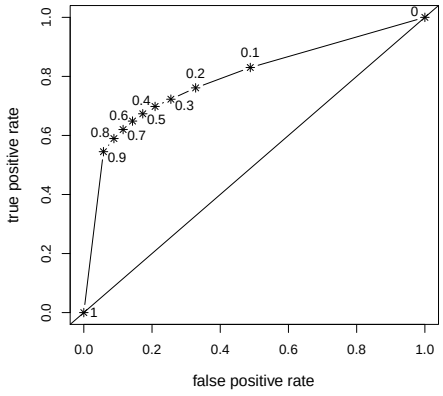
(a) V=36, Beta(11,1), AUC=0.8006



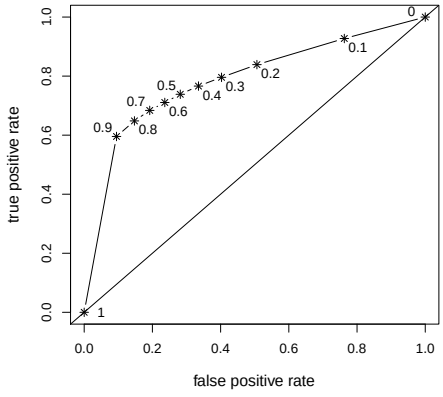
(b) V=36, Beta(8,2), AUC=0.8063



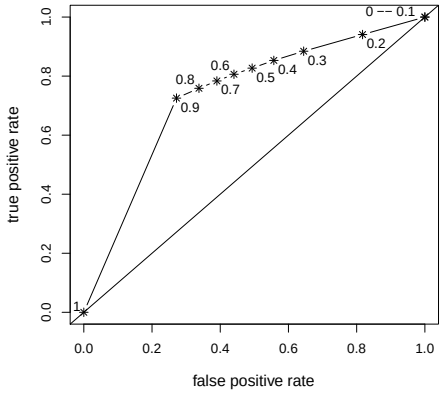
(c) V=36, Beta(0.8,0.2), AUC=0.7995



(d) V=36, Beta(80,20), AUC=0.8059

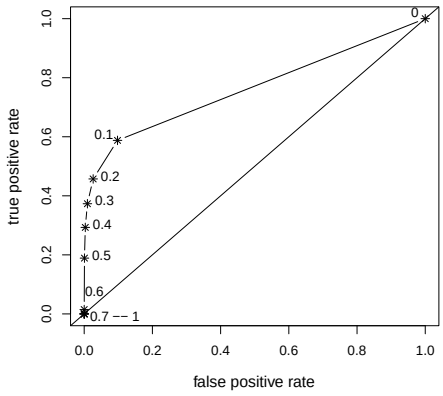


(e) V=36, Beta(9.41,4.03), AUC=0.8058

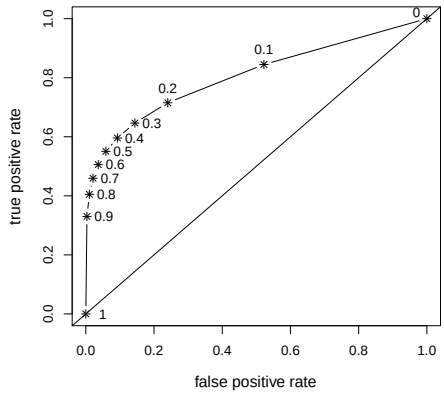


(f) V=36, Beta(8.09,8.09), AUC=0.8010

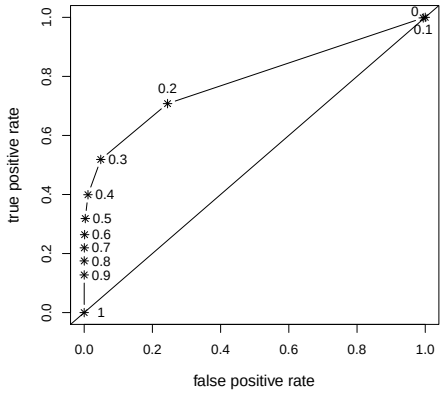
Figure 3.23: True positive rate averaged over 1000 datasets versus false positive rate averaged over 1000 datasets for $V = 36$ and six different Beta priors on π_0. The points are denoted by the corresponding cutting points, which are from 0 to 1 with increment 0.1. AUC is the area under the curve.



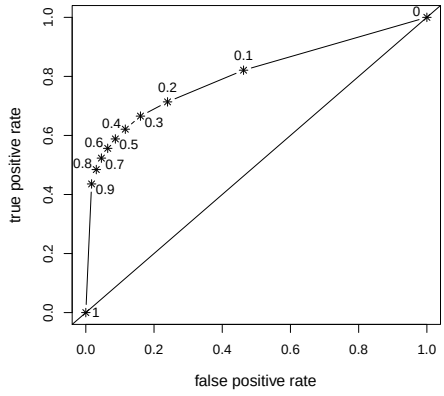
(a) V=16, Beta(11,1), AUC=0.8005



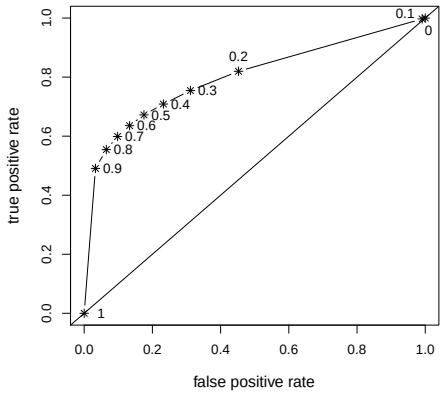
(b) V=16, Beta(8,2), AUC=0.8063



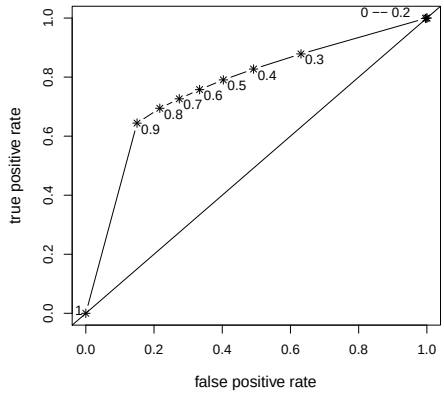
(c) V=16, Beta(0.8,0.2), AUC=0.7995



(d) V=16, Beta(80,20), AUC=0.8061

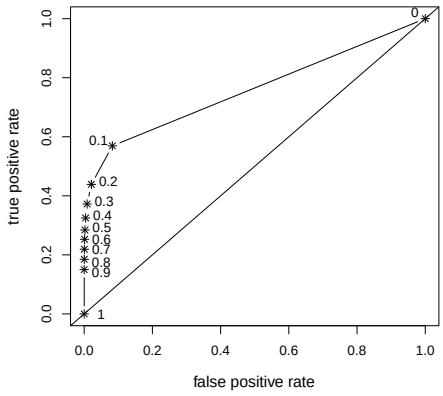


(e) V=16, Beta(9.41,4.03), AUC=0.8058

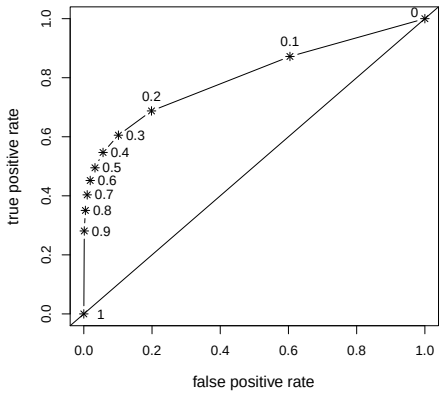


(f) V=16, Beta(8.09,8.09), AUC=0.8012

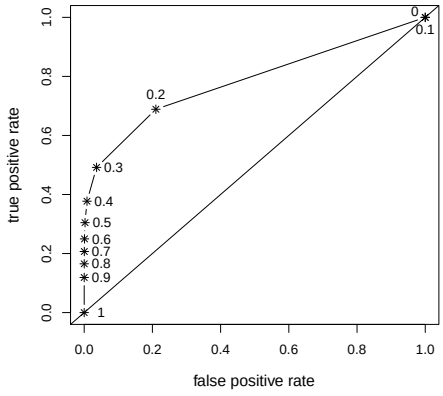
Figure 3.24: True positive rate averaged over 1000 datasets versus false positive rate averaged over 1000 datasets for $V = 16$ and six different Beta priors on π_0. The points are denoted by the corresponding cutting points, which are from 0 to 1 with increment 0.1. AUC is the area under the curve.



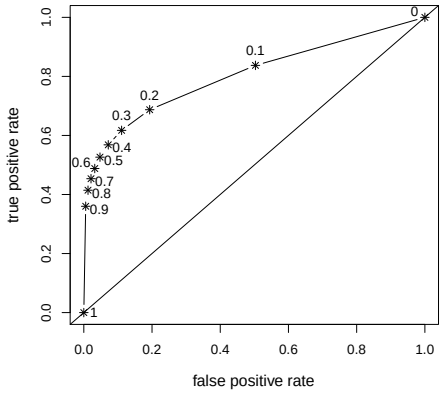
(a) V=9, Beta(11,1), AUC=0.8004



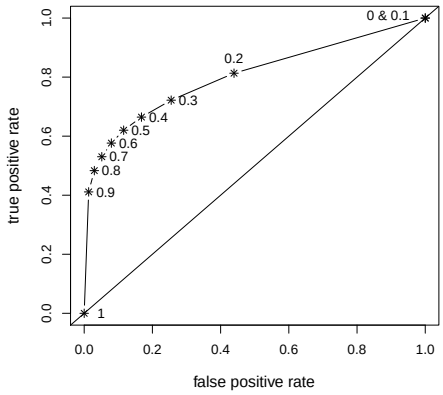
(b) V=9, Beta(8,2), AUC=0.8064



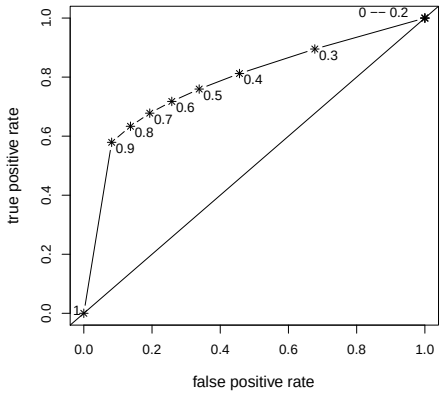
(c) V=9, Beta(0.8,0.2), AUC=0.7996



(d) V=9, Beta(80,20), AUC=0.8062

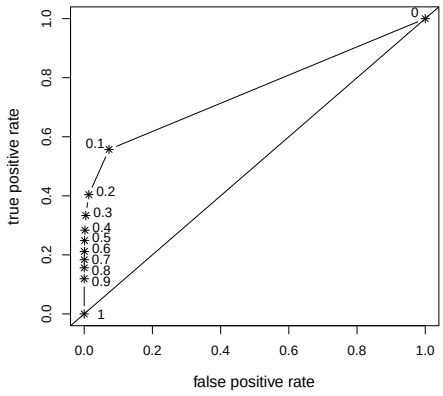


(e) V=9, Beta(9.41,4.03), AUC=0.8057

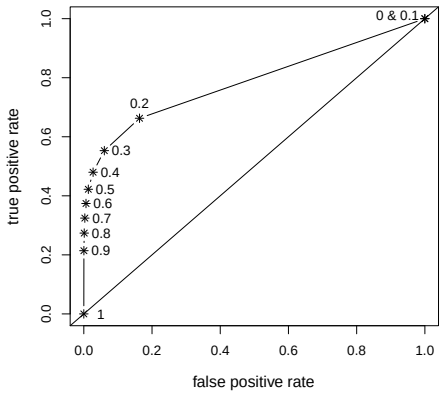


(f) V=9, Beta(8.09,8.09), AUC=0.8012

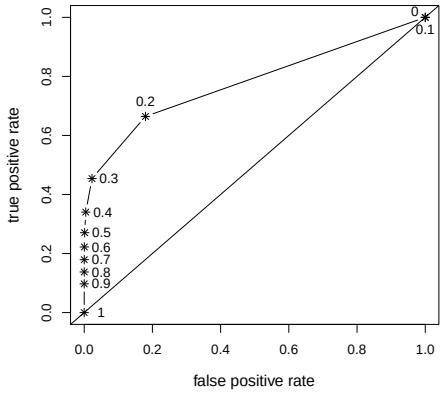
Figure 3.25: True positive rate averaged over 1000 datasets versus false positive rate averaged over 1000 datasets for $V = 9$ and six different Beta priors on π_0. The points are denoted by the corresponding cutting points, which are from 0 to 1 with increment 0.1. AUC is the area under the curve.



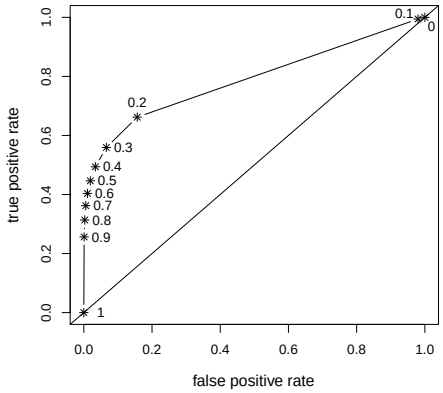
(a) V=4, Beta(11,1), AUC=0.8002



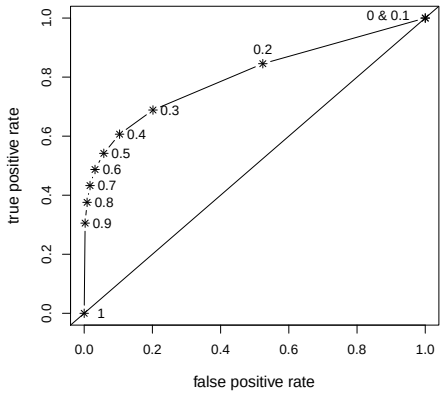
(b) V=4, Beta(8,2), AUC=0.8063



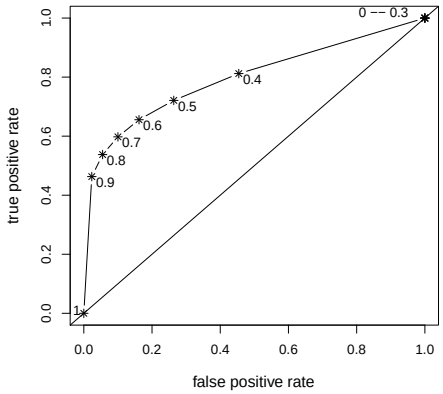
(c) V=4, Beta(0.8,0.2), AUC=0.7994



(d) V=4, Beta(80,20), AUC=0.8061



(e) V=4, Beta(9.41,4.03), AUC=0.8058



(f) V=4, Beta(8.09,8.09), AUC=0.8013

Figure 3.26: True positive rate averaged over 1000 datasets versus false positive rate averaged over 1000 datasets for $V = 4$ and six different Beta priors on π_0. The points are denoted by the corresponding cutting points, which are from 0 to 1 with increment 0.1. AUC is the area under the curve.

AUC values of the first multiple datasets are smaller than those of the second multiple datasets because the distance between the null and alternative hypotheses is smaller in the latter. The AUC values in Figure 3.23 – 3.26 are not smaller and are closer to each other (form 0.7994 to 0.8064) compared to those in Figure 3.19 – 3.22. Both AUC and the approximated ROC curves indicate that Beta$(0.8, 0.2)$ and Beta$(11, 1)$ are slightly worse than the other Beta priors. However, all AUC values and the approximated ROC curves are similar, indicating that there is no remarkable difference using priors Beta$(8, 2)$, Beta$(80, 20)$, Beta$(9.41, 4.03)$ or Beta$(8.09, 8.09)$ on π_0 and using prior variance $V = 4, 9, 16$ or 36 for datasets with $m = 100$, $\pi_0 = 0.9$, $\sigma = 1$ and $V' = 9$. Moreover, unlike the results in Figure 3.19 – 3.22, AUC in Figure 3.23 – 3.26 is not monotone in V.

All the average AUC values for the two multiple datasets are high with various simulation and prior parameters, indicating that the proposed method can identify a majority of the outliers with tolerable error. All the average AUC and the approximated ROC curves are similar for various prior parameters, indicating that the posterior is not very sensitive to the value of prior variance of the mean shift and the Beta prior of π_0.

3.5.3 Study of factorial design

The results given in Section 3.5.2 are for specific values of simulation parameters, and are therefore limited. However, the calculation of the average AUC of 1000 simulated datasets each with $m = 100$ observations, of which each with $n = 1000$ importance samples, takes about 18 hours on a PC with Intel Pentium D Dual processor of 2.8GHz and 2.79GHz, 2GB of Ram. So I perform a factorial design analysis to study the sensitivity of the proposed method to various simulation parameters, and the sensitivity of posteriors to priors for various values of simulation parameters. As indicated by the results of Section 3.5.1 and Section 3.5.2, the distribution of the explanatory variable does not seem to affect the posteriors, so I generate $\boldsymbol{x}$ from Benoulli$(1/2)$ for a value of m. Since the mean shifts of outliers in the simulated datasets are generated from $N(0, \sigma^2 V')$, the value of $\sigma^2 V'$ rather than σ affects the posterior. So I fix σ to be $1/4$, and vary V' from $4, 16$ to 36. Then a vector of indicators $\boldsymbol{H}$ is generated from Binomial(m, π_0) for each dataset. Note that in the factorial design study, the vector of indicators $\boldsymbol{H}$ is random, whereas for simplicity, it is fixed in the simulation studies of single and multiple datasets. In this design, the sample size m is chosen to be $20, 50, 100$ and 200, and the proportion of typical observations

$\pi_0 = 0.7, 0.8, 0.9$. Random errors $\boldsymbol{\varepsilon}$ are generated independently from $N(0, 1/16)$. Let m_1 be the number of nonzero elements in $\boldsymbol{H}$. Then m_1 values of μ_i are generated from $N(0, V')$ for each dataset. If $H_i = 0$, then $y_i = -0.5x_i + \varepsilon_i$, otherwise $y_i = -0.5x_i + \sigma\mu_i + \varepsilon_i$. For one simulated dataset, 1000 importance samples are taken by choosing a Beta prior Beta(a, b) on π_0 and a prior variance V of μ_i. Since Beta$(11, 1)$ seems to work similar to Beta$(0.8, 0.2)$ in Section 3.5.1 and Section 3.5.2 and it does not have the same mean or variance as any other Beta prior used in the simulation, I exclude this choice in the factorial design study in this section. The prior of π_0 is still chosen to be Beta$(0.8, 0.2)$, Beta$(8, 2)$, Beta$(80, 20)$, Beta$(9.41, 4.03)$ and Beta$(8.09, 8.09)$, where the first three have the same mean and decreasing variance, whereas Beta$(8, 2)$ and the last two have the same variance but decreasing mean. The prior variance of μ_i varies from $4, 9, 16$ to 36. Then five parameters $m, \pi_0, \sigma^2 V', V, a$ are chosen as five factors. Note that a and b are paired, so only one of them needs to be chosen as a factor. Since the simulated observations depend on σ but the importance samples are irrelevant to σ, two factors $\sigma^2 V'$ and V are distinct by scaling V' with σ. For each combination of the five factors, six datasets are generated. This factorial design is referred to as "Factorial Design 1" in the remaining part of this chapter. The factors and the levels of Factorial Design 1 are summarized in Table 3.17.

Factors	Description of Factors	Levels
m	Total number of observations	20, 50, 100, 200
π_0	Proportion of typical observations	0.7, 0.8, 0.9
$\sigma^2 V'$	Variance of $\sigma\mu_i$ of the simulated dataset	4/16, 16/16, 36/16
V	Variance of μ_i of the importance samples	4, 9, 16, 36
(a, b)	Beta Prior on π_0	$(0.8, 0.2), (8, 2), (80, 20)$ $(9.41, 4.03), (8.09, 8.09)$

Table 3.17: Factorial Design 1 on the AUC values calculated from simulated data sets with $k = 6$ datasets generated for each level of the five factors.

Then the AUC of each dataset is calculated for each combination of prior parameters by using the "R" package "verification" [65] if $m_1 > 0$. If $m_1 = 0$, then TPR is undefined. As discussed before, when a method has some skill the area under the ROC curve will exceed 0.5 [62]. So I set AUC to be 0.5 if $m_1 = 0$, where AUC $= 0.5$ means $\Pr[\{p_1|H_i = 1\} > \{p_1|H_i = 0\}] = \Pr[\{p_1|H_i = 1\} < \{p_1|H_i = 0\}]$, where $p_1 = P(H_i = 1|r_*^2)$. Hence totally $4 \times 3 \times 3 \times 4 \times 5 \times 6 = 4320$ AUC values are obtained and used for the

factorial design. The ANOVA table of Factorial Design 1 is given in Table 3.18. The means of the five main effects are given in Table 3.19 and the means of two-way interactions are given in Table 3.20 and Table 3.21. The tables of the means of interactions higher than the two-way are not given here.

From the ANOVA table in Table 3.18, we can see that the three main effects of simulation parameters, m, π_0 and $\sigma^2 V'$, but not the two main effects of priors, are significant, which are smaller than 0.005. These results indicate that the proposed methods works differently for simulated samples with different sample size or different proportion of outliers and different variance of the mean shift. Among all two-way interactions, only the p-value of $m \times \pi_0$, which are simulation parameters, is small (5.25×10^{-8}). Although the main effects of the Beta prior is not significant, the p-value of the four-way interaction among $m \times \pi_0 \times \sigma^2 V' \times a$, which is 0.00022 and equals 0.0067 for the most conservative multiple test, Bonferroni test. Thus the Beta prior may affect the posteriors. The p-values of the other two-way to five-way interactions are not small. Moreover, the grand mean in Table 3.19 is 0.81 and all the means in Table 3.19 – 3.21 are greater than 0.7, which indicates the proposed method does identify a majority of the outliers with tolerable error.

By comparing the means in Table 3.19 – 3.21, I find the means of $m = 20$ are different from those of other values of m. So I plot residual versus fitted value for the full model and denote the residuals for $m = 20$ by different symbols from the other residuals. This grouped residual vs. fitted value plot is given in Figure 3.27 (b). An un-grouped residual vs. fitted value plot is given in Figure 3.27 (a) in order to compare with that in (b). Note that the residuals are obtained from the factorial design model but not from the regression model.

The residual plots in Figure 3.27 show some serious heteroscedasticity, and we can see most extreme residuals come from datasets with $m = 20$. So I redesign the factors and exclude $m = 20$. This new factorial design is referred to as "Factorial Design 2" in the remaining part of this chapter. The factors and the levels for Factorial Design 2 are summarized in Table 3.22.

The ANOVA table of Factorial Design 2 is given in Table 3.23, the means of the main effects is given in Table 3.24. Then in order to compare Factorial Design 1 and Factorial

	Df	Sum of Square	Mean Square	F value	p-value
m	3	1.920	0.640	48.75	$<$ 2.2e-16
π_0	2	0.144	0.072	5.49	0.0042
$\sigma^2 V'$	2	17.505	8.753	666.63	$<$ 2.2e-16
V	3	0.114	0.038	2.90	0.034
a	4	0.034	0.009	0.65	0.63
$m \times \pi_0$	6	0.591	0.098	7.50	5.251e-08
$m \times \sigma^2 V'$	6	0.139	0.023	1.77	0.10
$\pi_0 \times \sigma^2 V'$	4	0.025	0.006	0.48	0.75
$m \times V$	9	0.111	0.012	0.94	0.49
$\pi_0 \times V$	6	0.046	0.008	0.59	0.74
$\sigma^2 V' \times V$	6	0.087	0.015	1.11	0.35
$m \times a$	12	0.069	0.006	0.44	0.95
$\pi_0 \times a$	8	0.214	0.027	2.04	0.039
$\sigma^2 V' \times a$	8	0.062	0.008	0.59	0.79
$V \times a$	12	0.098	0.008	0.62	0.82
$m \times \pi_0 \times \sigma^2 V'$	12	0.046	0.004	0.29	0.99
$m \times \pi_0 \times V$	18	0.191	0.011	0.81	0.69
$m \times \sigma^2 V' \times V$	18	0.236	0.013	1.00	0.46
$\pi_0 \times \sigma^2 V' \times V$	12	0.087	0.007	0.55	0.88
$m \times \pi_0 \times a$	24	0.384	0.016	1.22	0.21
$m \times \sigma^2 V' \times a$	24	0.211	0.009	0.67	0.89
$\pi_0 \times \sigma^2 V' \times a$	16	0.401	0.025	1.91	0.016
$m \times V \times a$	36	0.418	0.012	0.89	0.67
$\pi_0 \times V \times a$	24	0.289	0.012	0.92	0.58
$\sigma^2 V' \times V \times a$	24	0.157	0.007	0.50	0.98
$m \times \pi_0 \times \sigma^2 V' \times V$	36	0.402	0.011	0.85	0.72
$m \times \pi_0 \times \sigma^2 V' \times a$	48	1.192	0.025	1.89	0.00022
$m \times \pi_0 \times V \times a$	72	0.959	0.013	1.01	0.44
$m \times \sigma^2 V' \times V \times a$	72	0.872	0.012	0.92	0.66
$\pi_0 \times \sigma^2 V' \times V \times a$	48	0.468	0.010	0.74	0.90
$m \times \pi_0 \times \sigma^2 V' \times V \times a$	144	1.593	0.011	0.84	0.91
Residuals	3600	47.266	0.013		

Table 3.18: ANOVA table for Factorial Design 1 in Table 3.17.

Grand mean	0.81					
m	20	50	100	200		SE
mean	0.77	0.82	0.82	0.82		0.0049
π_0	0.7	0.8	0.9			SE
mean	0.81	0.81	0.80			0.0043
$\sigma^2 V'$	4/16	16/16	36/16			SE
mean	0.72	0.83	0.87			0.0043
V	4	9	16	36		SE
mean	0.80	0.80	0.81	0.82		0.0049
(a, b)	$(0.8, 0.2)$	$(8, 2)$	$(80, 20)$	$(9.41, 3.01)$	$(8.09, 8.09)$	SE
mean	0.80	0.81	0.81	0.81	0.81	0.0055

Table 3.19: Table of the means of main effects for Factorial Design 1 in Table 3.17.

$\pi_0 \backslash m$	20	50	100	200	SE
0.7	0.79	0.81	0.82	0.82	
0.8	0.79	0.82	0.82	0.82	
0.9	0.73	0.82	0.82	0.82	0.0085
$\sigma^2 V' \backslash m$	20	50	100	200	SE
4/16	0.70	0.73	0.73	0.73	
16/16	0.78	0.83	0.84	0.84	
36/16	0.83	0.89	0.88	0.89	0.0085
$V \backslash m$	20	50	100	200	SE
4	0.76	0.82	0.82	0.82	
9	0.76	0.81	0.81	0.82	
16	0.77	0.82	0.82	0.82	
36	0.79	0.82	0.82	0.82	0.0099
$\pi_0 \backslash \sigma^2 V'$	4/16	16/16	36/16		SE
0.7	0.72	0.83	0.88		
0.8	0.73	0.83	0.88		
0.9	0.71	0.82	0.87		0.0074
$\pi_0 \backslash V$	4	9	16	36	SE
0.7	0.81	0.80	0.81	0.82	
0.8	0.81	0.81	0.81	0.82	
0.9	0.79	0.79	0.80	0.81	0.0085

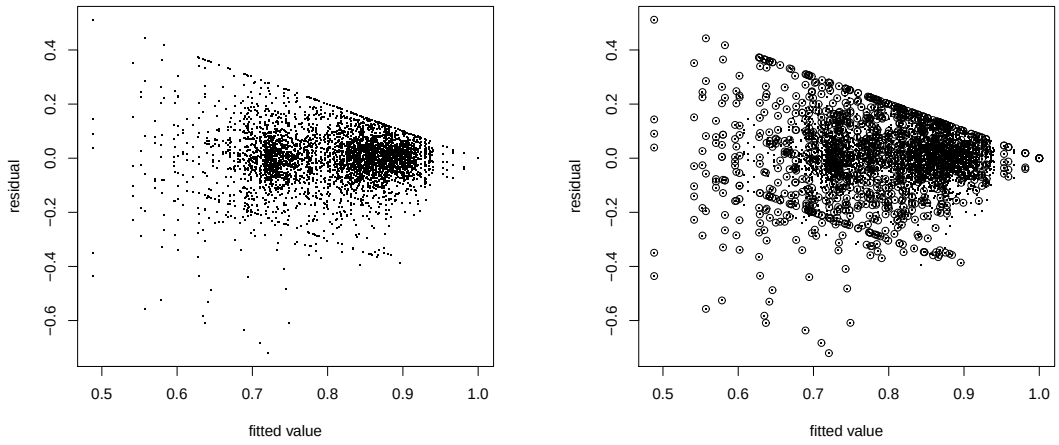
Table 3.20: Part I of the table of the means of two-way interactions for Factorial Design 1 in Table 3.17.

$\sigma^2V'\backslash V$	4	9	16	36		SE
4/16	0.73	0.71	0.72	0.73		
16/16	0.82	0.83	0.83	0.83		
36/16	0.78	0.82	0.83	0.86		0.0085
$m\backslash(a,b)$	$(0.8, 0.2)$	$(8, 2)$	$(80, 20)$	$(9.41, 3.01)$	$(8.09, 8.09)$	SE
20	0.76	0.78	0.77	0.77	0.77	
50	0.81	0.82	0.83	0.82	0.81	
100	0.82	0.82	0.82	0.82	0.82	
200	0.82	0.83	0.82	0.82	0.82	0.011
$\pi_0\backslash(a,b)$	$(0.8, 0.2)$	$(8, 2)$	$(80, 20)$	$(9.41, 3.01)$	$(8.09, 8.09)$	SE
0.7	0.80	0.81	0.81	0.81	0.82	
0.8	0.82	0.81	0.81	0.81	0.82	
0.9	0.79	0.81	0.82	0.80	0.79	0.0095
$\sigma^2V'\backslash(a,b)$	$(0.8, 0.2)$	$(8, 2)$	$(80, 20)$	$(9.41, 3.01)$	$(8.09, 8.09)$	SE
4/16	0.71	0.73	0.73	0.72	0.72	
16/16	0.83	0.83	0.83	0.83	0.82	
36/16	0.87	0.88	0.88	0.87	0.87	0.0095
$V\backslash(a,b)$	$(0.8, 0.2)$	$(8, 2)$	$(80, 20)$	$(9.41, 3.01)$	$(8.09, 8.09)$	SE
4	0.80	0.80	0.82	0.80	0.81	
9	0.80	0.80	0.81	0.80	0.80	
16	0.81	0.82	0.81	0.80	0.81	
36	0.81	0.82	0.81	0.82	0.81	0.011

Table 3.21: Part II of the table of the means of two-way interactions for Factorial Design 1 in Table 3.17.

Factors	Description of Factors	Levels
m	Total number of observations	50, 100, 200
π_0	Proportion of typical observations	0.7, 0.8, 0.9
σ^2V'	Variance of $\sigma\mu_i$ of the simulated dataset	4/16, 16/16, 36/16
V	Variance of μ_i of the importance samples	4, 9, 16, 36
(a,b)	Beta Prior on π_0	$(0.8, 0.2), (8, 2), (80, 20)$ $(9.41, 4.03), (8.09, 8.09)$

Table 3.22: Factorial Design 2 on the AUC values calculated from simulated data sets with $k = 6$ datasets generated for each level of the five factors and without $m = 20$.



(a) Residual vs. fitted value plot (b) Residual vs. fitted value plot with groups

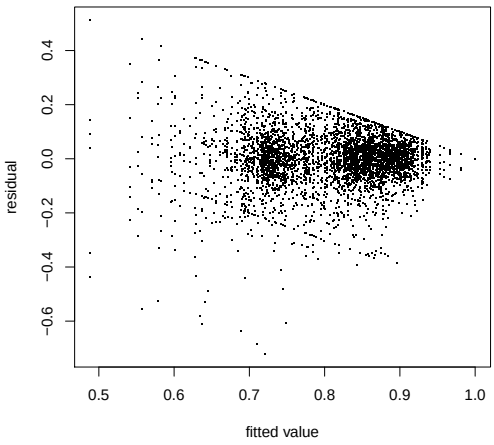
Figure 3.27: Residual vs. fitted value plots for the factorial design in Table 3.17. (a)Ungrouped residual vs. fitted value plot. (b) Grouped residual vs. fitted value plot, in which circles ($\circ$) denote the residuals for observations with $m = 20$ and dots ($\cdot$) denote the residuals for observations with $m = 50, 100, 200$.

Design 2, the residual vs. fitted value plots and the normal QQ plots of the residuals for the two models are shown in Figure 3.28.

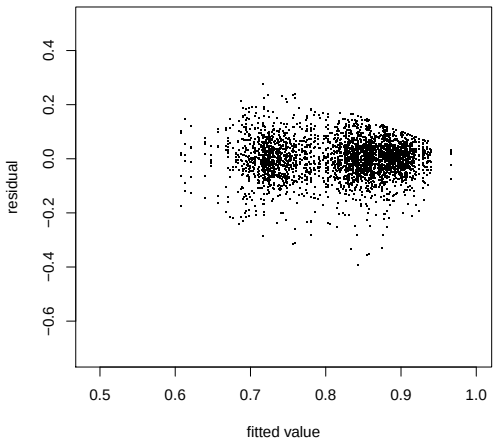
The new results in Table 3.23 indicate that there is only one significant main effect of $\sigma^2 V'$, which is the variance of outlier in the simulated dataset and is irrelevant to the hyper-parameters. Note that a boundary appears in the residual vs. fitted value plots in Figure 3.28 (a) and (b). This boundary is corresponding to the line residual + fitted value = 1 because AUC $\leq$ 1. The range of the residuals is $(-0.72,\ 0.51)$ for the full model as shown in Figure 3.28 (a), and it is reduced to the $(-0.39,\ 0.28)$ for the model excluded $m = 20$ as shown in Figure 3.28 (b). We can see that Figure 3.28 (b) is more homoscedastic than Figure 3.28 (a), and the residual in the QQ plot in Figure 3.28 (c) has a heavier tail than that in (d). However there are still some heteroscedasticity in the residual plot in Figure 3.28 (b), and the distribution of the residual in the QQ plot in Figure 3.28 (d) still has a heavier tail than the normal distribution. These indicate that the distribution of AUC may not be normal, but the analysis of variance is robust for nonnormality and heteroscedasticity more seriously violate the model than nonnormality.

	Df	Sum of Square	Mean Square	F value	p-value
m	2	0.0210	0.0105	1.75	0.17
π_0	2	0.0076	0.0038	0.64	0.53
σ^2V'	2	14.3182	7.1591	1193.27	<2e-16
V	3	0.0210	0.0070	1.17	0.32
a	4	0.0163	0.0041	0.68	0.61
$m \times \pi_0$	4	0.0080	0.0020	0.33	0.86
$m \times \sigma^2V'$	4	0.0258	0.0065	1.076	0.37
$\pi_0 \times \sigma^2V'$	4	0.0077	0.0019	0.32	0.86
$m \times V$	6	0.0088	0.0015	0.25	0.96
$\pi_0 \times V$	6	0.0249	0.0041	0.69	0.66
$\sigma^2V' \times V$	6	0.0431	0.0072	1.20	0.31
$m \times a$	8	0.0586	0.0073	1.22	0.28
$\pi_0 \times a$	8	0.0247	0.0031	0.51	0.85
$\sigma^2V' \times a$	8	0.0462	0.0058	0.96	0.46
$V \times a$	12	0.1061	0.0088	1.47	0.13
$m \times \pi_0 \times \sigma^2V'$	8	0.0270	0.0034	0.56	0.81
$m \times \pi_0 \times V$	12	0.0747	0.0062	1.038	0.41
$m \times \sigma^2V' \times V$	12	0.0988	0.0082	1.37	0.17
$\pi_0 \times \sigma^2V' \times V$	12	0.0934	0.0078	1.30	0.21
$m \times \pi_0 \times a$	16	0.0896	0.0056	0.93	0.53
$m \times \sigma^2V' \times a$	16	0.0966	0.0060	1.0068	0.45
$\pi_0 \times \sigma^2V' \times a$	16	0.0944	0.0059	0.98	0.47
$m \times V \times a$	24	0.1165	0.0049	0.81	0.73
$\pi_0 \times V \times a$	24	0.1504	0.0063	1.044	0.40
$\sigma^2V' \times V \times a$	24	0.1419	0.0059	0.99	0.48
$m \times \pi_0 \times \sigma^2V' \times V$	24	0.0796	0.0033	0.55	0.96
$m \times \pi_0 \times \sigma^2V' \times a$	32	0.1840	0.0057	0.96	0.53
$m \times \pi_0 \times V \times a$	48	0.2562	0.0053	0.89	0.69
$m \times \sigma^2V' \times V \times a$	48	0.2700	0.0056	0.94	0.60
$\pi_0 \times \sigma^2V' \times V \times a$	48	0.2141	0.0045	0.74	0.90
$m \times \pi_0 \times \sigma^2V' \times V \times a$	96	0.5661	0.0059	0.98	0.53
Residuals	2700	16.1987	0.0060		

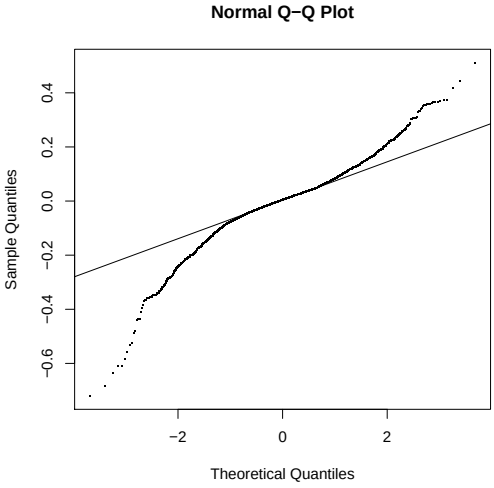
Table 3.23: ANOVA table for Factorial Design 2 in Table 3.22.



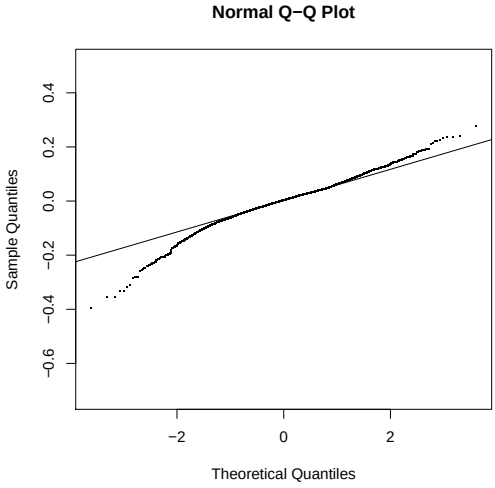
(a) Residual vs. fitted value plot for the model with $m = 20$



(b) Residual vs. fitted value plot for the model without $m = 20$



(c) Normal QQ plot of the residuals for the model with $m = 20$



(d) Normal QQ plot of the residuals for the model without $m = 20$

Figure 3.28: Residual vs. fitted value plots and normal QQ plots of Factorial Design 1 and 2 in Table 3.17 and 3.22.

Grand mean	0.82					
m	50	100	200			SE
mean	0.82	0.82	0.82			0.0033
π_0	0.7	0.8	0.9			SE
mean	0.82	0.82	0.82			0.0033
$\sigma^2 V'$	4/16	16/16	36/16			SE
mean	0.73	0.84	0.89			0.0033
V	4	9	16	36		SE
mean	0.82	0.82	0.82	0.82		0.0038
(a, b)	$(0.8, 0.2)$	$(8, 2)$	$(80, 20)$	$(9.41, 3.01)$	$(8.09, 8.09)$	SE
mean	0.82	0.82	0.82	0.82	0.82	0.0043

Table 3.24: Table of the means of the main effects for Factorial Design 2 in Table 3.22.

A more sensitive analysis needs a transformation of AUC. We can conclude that the Beta priors Beta$(8, 2)$, Beta$(80, 20)$, Beta$(9.41, 4.03)$ and Beta$(8.09, 8.09)$ and the prior variance of μ do not appreciably affect the posterior probability $P(H_i = 1|r_i^{*2})$ as long as the sample size is not too small.

3.6 Summary

Linear regression models are used in analyzing data from many fields of study. These data are usually contaminated and contain outliers. In some cases, these outliers are more interesting than the other data. The main purpose of this chapter is to identify outliers in linear regression models. The methods based on deletion residuals are powerful methods to identify single outliers. The distribution of the deletion residual of an observation under the null hypothesis is shown to be a student t distribution when this observation is the only suspicious outlier in a given data set. However, there is usually more than one outlier in a real dataset, and multiple outliers may hide the effect of each other which is called "masking" problem. Hence, simply deleting observations sequentially is not suitable and the effect of all observations needs to be taken into account simultaneously. When there is more than one outlier and the distribution of outliers is assumed to have a mean shift, I have proved the marginal distribution of the deletion residual of an observation is a doubly noncentral t distribution. Hence, the marginal distribution of the squares of deletion

residuals is a doubly noncentral F distribution.

The outlier identification problem can be viewed as a multiple testing problem. In this chapter, a multiple testing method to identify outliers was proposed. The proposed method is a Bayesian method with the test statistics being deletion residuals and the prior distributions of π_0 and μ_i being Beta and normal, respectively. An importance sampling method was used to compute the marginal posterior probability that an observation is an outlier given its own deletion residual. This posterior probability was used to measure the outlyingness of an observation. Then decisions can be made cooperating certain decision rules.

In the last section of this chapter, the proposed Bayesian method was applied to some simulated datasets. The simulation parameters were varied over a set of values and various priors were employed to study how the priors affect the posterior. First, the proposed Bayesian method was applied to two single datasets, and the ROC curve was plotted and the area under ROC curve was calculated for each dataset and each combination of parameters. Secondly, the proposed Bayesian method was applied to two sets of data, each including 1000 datasets, and the ROC curve for average TPR versus FPR was plotted and the average AUC over 1000 replicates was calculated for each set of data and different priors. In the two simulation studies, the resulting AUC values are high for various choices of priors, indicating that the proposed method can identify a majority of the outliers with tolerable error. Both ROC curves and AUC values obtained by using different priors are similar, indicating that the posterior probability is not sensitive to the chosen priors. At last, a factorial design analysis was used to compare the AUC using a wider range of simulation and prior parameters. The results of the factorial design analysis show that the priors do not affect the marginal posterior probability $P(H_i = 1 \mid r_i^{*2})$ as long as the sample size is not too small.

When calculating the posterior probabilities $P(H_i = 1 \mid r_i^{*2})$, I used Patnaik's approximation to calculate the density of the doubly noncentral F distribution. To examine the accuracy of Patnaik's approximation, I also proposed an algorithm and used it to calculate the density of the doubly noncentral F distribution, and compared the results obtained by using the two methods. For the first simulated dataset, the maximum differences of densities by using the two methods are not larger than 0.071, and the maximum differences of posteriors calculated by using the two methods are smaller than 6.22×10^{-5}. The Pat-

naik's approximation is also faster than the proposed algorithm. The results show that the doubly noncentral F density calculated by Patnaik's approximation is not too far from the true density, and the computation time can be saved by using Patnaik's approximation.

I used an importance sampling method to calculate the marginal posterior probability $P(H_i = 1 \mid r_i^{*2})$, where I chose the joint prior to be the importance function. As mentioned in Section 3.3.3, the choice may be problematic. The estimates would be poor if some importance ratios are much larger than the others. Gelman *et al.* [37] suggest to examine the distribution of sampled importance ratios to detect possible problems. I plotted the histogram of the logarithms of the importance ratios for some simulated datasets and did not find any exceedingly large ratio. Other importance functions may be considered and a comparison of different importance functions could be done in the future. For example, we can use Gibbs sampling methods to approximate the joint posterior distribution of parameters.

The proposed method measures the outlyingness of the ith observation by the marginal posterior probability that the ith observation is an outlier given its own deletion residual, and hence the information of the other observations is not included in this marginal posterior and the outlyingness of all observation are not tested simultaneously. A future work could be calculating the joint posterior probabilities $P(\boldsymbol{H} = \boldsymbol{h} \mid r_1^{*2} \cdots, r_m^{*2})$, where $\boldsymbol{h} \in \{0,1\}^m$, or to calculate the marginal posterior probability $P(H_i = 1 \mid r_1^{*2} \cdots, r_m^{*2})$, rather than the marginal posterior probability $P(H_i = 1 \mid r_i^{*2})$. Computing either $P(\boldsymbol{H} = \boldsymbol{h} \mid r_1^{*2} \cdots, r_m^{*2})$ or $P(H_i = 1 \mid r_1^{*2} \cdots, r_m^{*2})$ requires the calculation of the joint distribution of all the deletion residuals $r_1^{*2} \cdots, r_m^{*2}$, which marginally have doubly noncentral t distributions. The joint distribution describes the dependence structure of $r_1^{*2} \cdots, r_m^{*2}$, and thus the two posteriors given all the deletion residuals contain more information of data than $P(H_i = 1 \mid r_i^{*2})$. So the proposed method is expected to be improved by measuring the outlyingness of all observations simultaneously with $P(\boldsymbol{H} = \boldsymbol{h} \mid r_1^{*2} \cdots, r_m^{*2})$.

Chapter 4

Amino acid sequence similarity of viral to human proteomes (An application of the Bayesian method proposed in Chapter 3)

4.1 Introduction

An immune system protects its host against disease by identifying and killing pathogens. Meanwhile an autoimmune disease may cause an immune system to fail to kill pathogens and attack normal cells. It is proposed that some autoimmune reactions are related to similarities in proteins between a virus and a host [66]. It is possible to study the amino acid sequence similarity of viral proteomes to the human proteome since protein sequences of the human proteome and those of a number of viral proteomes are available in databanks [56]. Kanduc *et al.* [56] examined thirty proteomes for amino acid sequence similarity to the human proteome and revealed "a massive, indiscriminate, unexpected pentapeptide overlapping between viral and human proteomes", where a *pentapeptide* is a peptide having five amino acids. They also performed a linear regression analysis to determine whether a linear relationship exists between the level of viral overlaps to the human proteome and the length of viral proteomes. Their results show that the level of overlaps does have a strong linear correlation with the viral proteome length (with a coefficient of determination, $R^2 = 0.9497$). They also reported that "among the examined viruses, human T-lymphotropic virus 1, Rubella virus, and hepatitis C virus present the highest number of viral overlaps to the human proteome" [56, p.1755]. Recall from Section 1.5 the short form of these three viruses, the Kanduc *et al.* identified (KI) viruses. One interesting question that arises from this linear regression analysis is if there exist some viral proteomes sharing significantly higher or lower levels of overlaps with the human proteome than the predicted level of overlaps by the linear regression model, *i.e.* the outliers in the level of viral overlaps to the

human proteome. Are KI viruses the interesting viruses?

The Bayesian method proposed in Chapter 3 is used to identify outliers in this dataset. I received the data that was published in Kanduc *et al.* [56] from Kusalik [58]. The results confirm the report given in Kanduc *et al.* [56] in the case that only the viruses with not too large proteome size (less than $10,000$ pentapeptides) are analyzed. However, the full dataset, which are used in [56], has four viruses (Human herpesvirus 4, Human herpesvirus 6, Variola virus, and Human herpesvirus 5) with extremely large size (greater than $32,009$ pentapeptides), and these four viruses, if included, are more likely to be the outliers. Recall from Section 1.5 the short form of these four viruses, the four extremely large (FEL) viruses. Among the other 26 viruses, KI viruses still do not have greater posterior probability of being an outlier given its deletion residual than the other viruses, when all thirty viral proteomes are analyzed.

The description of the dataset examined in [56] is given in Section 4.2. The results from the analysis of the full dataset are presented in Section 4.3.1, and the results for the analysis of the dataset excluding the FEL viruses are shown in 4.3.2. The conclusion is given in the last section.

4.2 Description of the dataset

The thirty viral proteomes used for amino acid sequence similarity analysis in Kanduc *et al.* [56] were downloaded from www.ebi.ac.uk/genomes/virus.html and the human proteome was obtained from UniProtKB (www.ebi.ac.uk/integr8). After being filtered by certain rules, the human proteome analyzed in [56] consists of $36,103$ unique proteins. The 30 viral proteomes are made up of 717 proteins that is equal to $302,667$ amino acids. As described in Section 3 of Kanduc *et al.* [56], the 30 viruses were chosen by the five criteria: (1) "known to be pathogenic to human"; (2) "of significant health impact"; (3) "phylogenetically different"; (4) "proteomes established to a significant degree of completeness"; (5) the viral proteomes "span a range of proteome sizes". The proteome sizes of the 30 viruses varies from $1,613$ to $65,280$ amino acids and the total number of amino acids of all viral proteomes is equal to $302,667$. The description of the 30 viral proteomes including the numbers of their amino acids are given in Table 1 in [56], which is reproduced here as Table 4.1. This table is the same as Table 1 in [56].

Tax ID	Virus description and abbreviation	Accession	# of Proteins	# of Amino Acids
10407	Hepatitis B virus (HBV)	X51970	4	1,613
10632	JC polyomavirus (JCV)	J02226	5	1,629
10798	Human parvovirus B19	AF162273	3	2,006
12131	Human rhinovirus 14 (HRV-14)	K02121	1	2,179
12080	Human poliovirus 1 (HPV-1)	AJ132961	1	2,209
434309	Saffold virus(SAF-V)	EF165067	1	2,296
333760	Human papillomavirus type 16 (HPV16)	K02718	8	2,452
11908	Human T-cell leukemia virus 1 (HTLV-I)	U19949	6	2,589
11103	Hepatitis C virus (HCV)	AJ132997	1	3,010
11041	Rubella virus	AF188704	2	3,179
11089	Yellow fever virus (YFV)	X03700	1	3,411
307044	West Nile virus (WNV)	AY842931	1	3,433
11676	Human immunodeficiency virus 1 (HIV-1)	X01762	9	3,571
11292	Rabies virus	M31046	5	3,600
11029	Ross River virus	M20162	2	3,733
11709	Human immunodeficiency virus 2 (HIV-2)	X05291	9	3,759
162145	Human metapneumovirus (hMPV)	AF371337	9	4,163
93838	Influenza A virus (H5N1)	AF144300	10	4,467
11250	Human respiratory syncytial virus (HRSV)	AF013254	11	4,540
11216	Human parainfluenza virus 3 (HPIV3)	AB012132	6	4,842
11269	Lake Victoria marburgvirus	Z12132	7	4,846
11161	Mumps virus	AB040874	8	4,977
70149	Measles virus	AY486084	8	5,205
186538	Zaire virus	AF086833	9	5,493
63330	Hendra virus	AF017149	9	6,056
321149	SARS coronavirus (SARS-CoV)	AY864806	13	14,209
10376	Human herpesvirus 4 (HHV-4)	AY961628	69	34,911
10368	Human herpesvirus 6 (HHV-6)	X83413	112	44,720
10255	Variola virus	X69198	197	54,289
10359	Human herpesvirus 5 (HHV-5)	X17403	190	65,280
Total			717	302,667

Table 4.1: Description of the viral proteomes analyzed for similarity to human proteins [58]. The first two columns are respectively the taxonomic ID of the virus, and the description and abbreviation of the virus. The last two columns present respectively the number of proteins of the virus, and the number of the amino acids of the virus.

Each of the 30 viral proteomes was analyzed by Kanduc *et al.* [56] for pentapeptide overlapping with the human proteome since "pentapeptides are minimal structural units critically involved in biological/pathological interaction such as peptide-protein interaction and (auto)immune recognition" [56]. First, each viral proteome was decomposed into a set of pentapeptides that includes some duplicates. Secondly, the duplicates were removed to create a set of unique pentapeptides for each viral proteome. Then for each pentapeptide of one viral proteome, the human proteome was scanned for the same pentapeptide, and the number of unique pentapeptides and the number of duplicates occurring in the human proteome were recorded. Their results of pentapeptide overlapping between viral proteomes and human proteome are given in Table 4 in [56]. Column 1 and 4 of this table is reproduced here as Table 4.2. In this table, the first column is the name or the abbreviation of the virus, which is given in Table 4.1; the second column is the number of the unique pentapeptides in the viral proteome; the third column presents the total number of viral pentapeptide overlaps including duplicates in the human proteome. The last row is obtained by combining 30 viral proteomes into one viral proteome and searching for the pentapeptide overlap in the entire human proteome.

Table 4.2 shows that all the 30 viral proteomes analyzed in the study of Kanduc *et al.* [56] have high pentapeptide overlapping with the human proteome. In order to determine whether the level of viral proteome overlaps to the human proteome depends on the size of the viral proteome, Kanduc *et al.* also performed a linear regression analysis on the data in Table 4.2. I perform the same regression analysis and obtain the same results as in [56]. The scatter plot of the total number of viral pentapeptide overlaps including duplicates in the human proteome versus the unique pentapeptides in the viral proteome with the linear regression line is given in Figure 4.1. The plot in Figure 4.1 is on a log-log scale. The linear regression equation in Figure 4.1 is $y = 12.636x - 269.01$, where y is the level of pentapeptide overlaps between the viral and human proteomes and x is the viral proteome length, and the coefficient of determination, $R^2 = 0.9497$.

Then Kanduc *et al.* [56] concluded that there is a strong linear relationship between the level of viral overlaps to the human proteome and the length of the viral proteome, and "HTLV-1, Rubella virus and HCV present a number of overlaps to the human proteome above the expected number of overlaps predicted by the linear regression line" [56, p. 1761].

Virus	1	2
HBV	1,589	21,852
JCV	1,531	22,482
Human parvovirus B19	1,443	21,488
HRV-14	2,173	23,761
HPV-1	2,203	23,431
SAF-V	2,283	23,995
HPV16	2,419	28,948
HTLV-I	2,563	44,042
HCV	3,002	46,731
Rubella virus	3,154	51,859
YFV	3,400	43,245
WNV	3,424	42,670
HIV-1	3,082	35,568
Rabies virus	3,575	42,643
RRV	3,622	42,422
HIV-2	3,285	45,724
hMPV	4,120	52,915
H5N1	4,412	45,599
HRSV	4,483	46,540
HPIV3	4,807	52,934
Lake Victoria marburgvirus	4,808	67,051
Mumps virus	4,786	61,013
Measles virus	4,934	60,638
Zaire virus	4,865	56,577
Hendra virus	5,210	57,646
SARS-CoV	9,739	108,632
HHV-4	32,009	531,946
HHV-6	41,834	467,206
Variola virus	52,017	498,970
HHV-5	61,001	883,952
All	257,035	2,907,096

Table 4.2: Pentapeptide overlap between viral and human proteomes [58]. Column numbers 1-2 refer to: (1) number of unique pentapeptides in the viral proteome; (2) total number of viral pentapeptide overlap including duplicates in the human proteome.

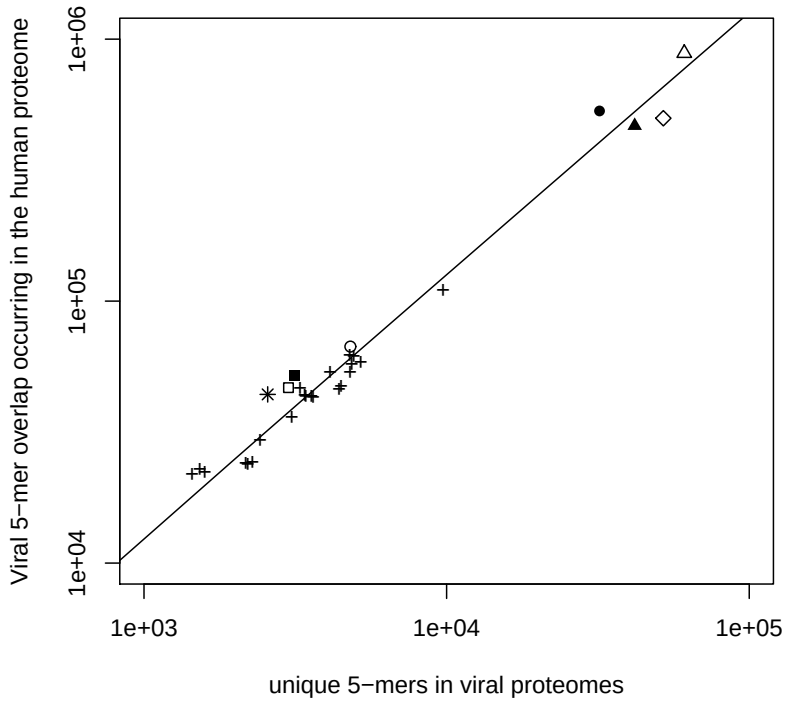


Figure 4.1: The scatter plot of the viral pentapeptide overlap including duplicates in the human proteome versus the number of unique pentapeptides in the viral proteome (see data in Table 4.2). The symbols refer to: ($*$), HTLV-1; (■), Rubella virus; (□), HCV; ($\circ$), Lake victoria marburgvirus; (●), HHV-4; (▲), HHV-6; (◊), Variola virus; (△), HHV-5; ($+$), other viral data point. The regression line (—) has an equation of $y = 12.64x - 269.01$ with a coefficient of determination, $R^2 = 0.9497$. Both x- and y-axis are log scale.

In order to examine their claim, I apply the Bayesian method proposed in Chapter 3 to the data in Table 4.2. The results of the analysis are presented in the next section.

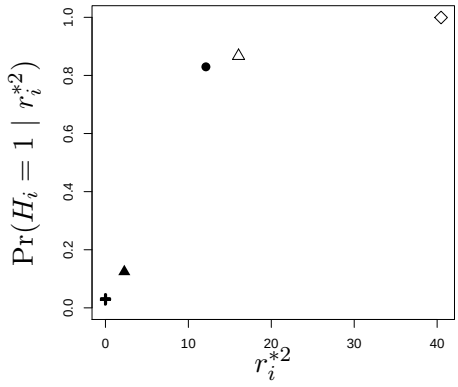
4.3 Analysis of the dataset

My goal is to identify viral proteomes sharing more pentapeptide overlapping with the human proteome than the other viral proteomes. So I need to test for outliers in the regression model of the level of overlaps and the size of viral proteomes. In this section, the Bayesian model proposed in Chapter 3 is applied to the data in Table 4.2.

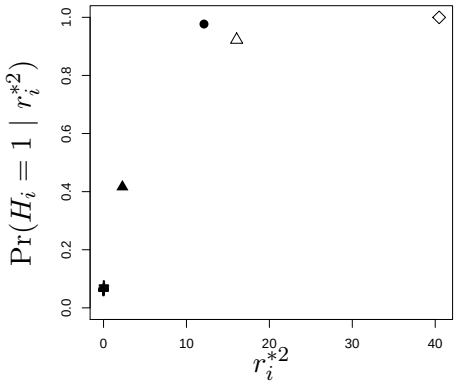
4.3.1 Analysis of the full dataset

I apply Algorithm 3.3.1 proposed in Chapter 3 to the dataset to calculate the posterior probability that an observation is an outlier given its deletion residual, in which Patnaik's approximation is used to calculate the density of the doubly noncentral F distribution. I still use various priors to study how the posterior relies on the priors. The Beta prior on π_0, the proportion of typical observations, is chosen to be Beta$(11, 1)$, Beta$(8, 2)$, Beta$(0.8, 0.2)$, Beta$(80, 20)$, Beta$(9.41, 4.03)$ and Beta$(8.09, 8.09)$. The prior distribution Beta$(11, 1)$ is used in Scott and Berger [88]. The priors Beta$(0.8, 0.2)$, Beta$(8, 2)$ and Beta$(80, 20)$ have the same mean equal to 0.8 but decreasing variances, and Beta$(9.41, 4.03)$ and Beta$(8.09, 8.09)$ have the same variance as Beta$(8, 2)$ but respectively have mean 0.7 and 0.5. I do not know the value of σ^2, but I assume it to be 1 so that $\sigma^2 V = V$. I choose $V = 4, 9, 16, 36$ to be the variances of sampled μ_i. For each observation, $n = 1000$ random samples $(\boldsymbol{H}^j_{(i)}, \boldsymbol{\mu}^j, \pi_0^j)$, $j = 1, \cdots, n$, are generated to calculate the posterior probability $P(H_i = 1 \mid r_i^{*2})$. Then I plot $P(H_i = 1 \mid r_i^{*2})$ as a function of r^{*2} for various priors on π_0 and μ_i, and the results are shown in Figure 4.2 - Figure 4.5. The prior standard deviation of μ varies from 6 to 2 in Figure 4.2 - 4.5, where (a) - (f) are for Beta$(11, 1)$, Beta$(8, 2)$, Beta$(0.8, 0.2)$, Beta$(80, 20)$, Beta$(9.41, 4.03)$ and Beta$(8.09, 8.09)$.

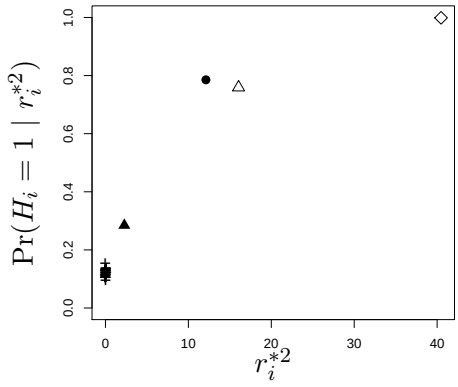
It can be seen from all 24 graphs in Figure 4.2 - 4.5 that the posteriors $P(H_i = 1 \mid r_i^{*2})$ of the FEL viruses, which have much larger proteome size (from $32,009$ to $61,000$ pentapeptides) than the other 26 viruses (from $1,589$ to $9,739$ pentapeptides), are much greater than the posteriors of the other viruses, whereas the other 26 viruses have very close value of $P(H_i = 1 \mid r_i^{*2})$. According to the posteriors, the FEL viruses are more



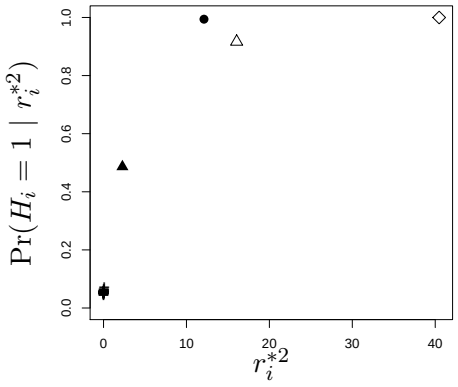
(a) Prior $V = 36$, Beta$(11, 1)$



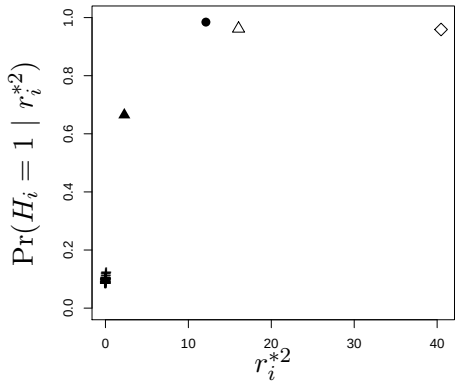
(b) Prior $V = 36$, Beta$(8, 2)$



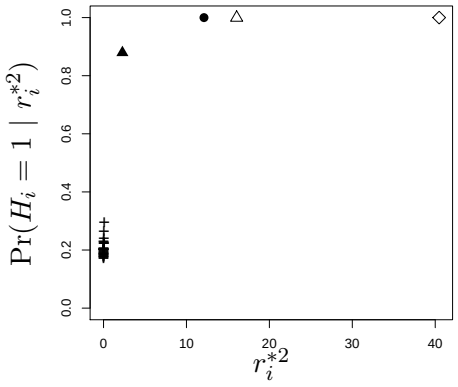
(c) Prior $V = 36$, Beta$(0.8, 0.2)$



(d) Prior $V = 36$, Beta$(80, 20)$

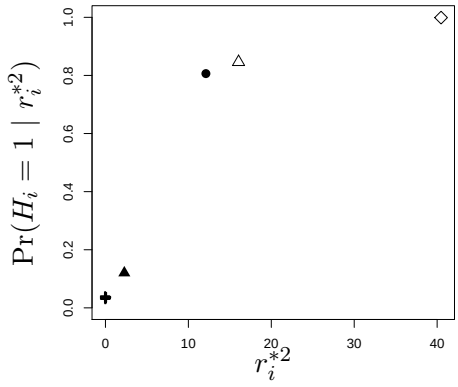


(e) Prior $V = 36$, Beta$(9.41, 4.03)$

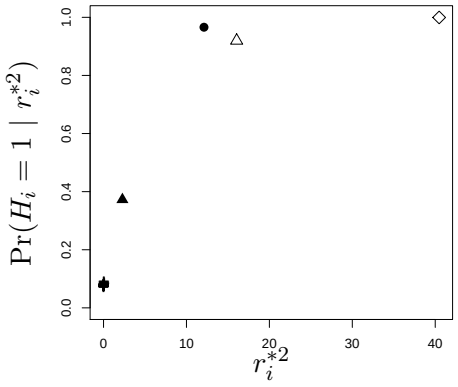


(f) Prior $V = 36$, Beta$(8.09, 8.09)$

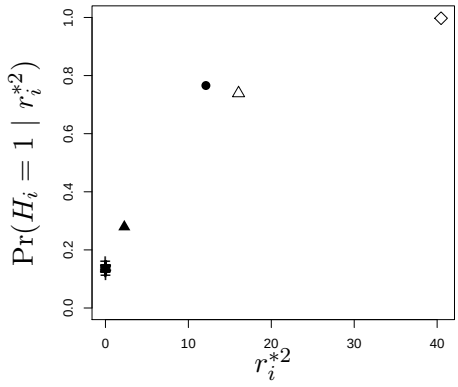
Figure 4.2: The posterior probability $P(H_i = 1 \mid r_i^{*2})$ is plotted as a function of r_i^{*2} for $V = 36$ and six different Beta priors on π_0. The symbols refer to: (•), HHV-4; (▲), HHV-6; (◊), Variola virus; (△), HHV-5; (+), other viral data point. The data under analysis is shown in Table 4.2.



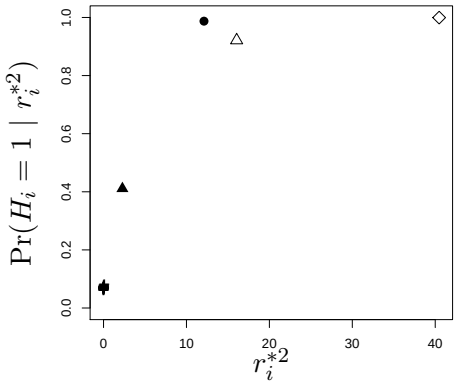
(a) Prior $V = 16$, Beta$(11, 1)$



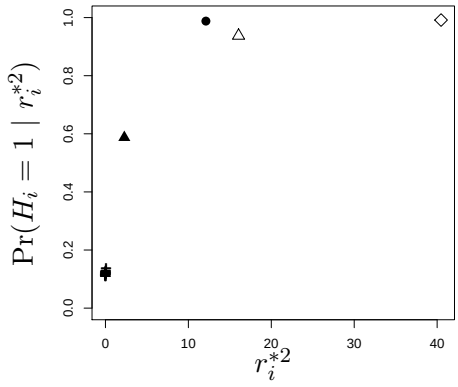
(b) Prior $V = 16$, Beta$(8, 2)$



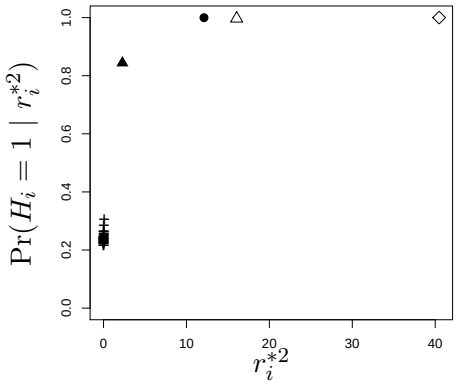
(c) Prior $V = 16$, Beta$(0.8, 0.2)$



(d) Prior $V = 16$, Beta$(80, 20)$

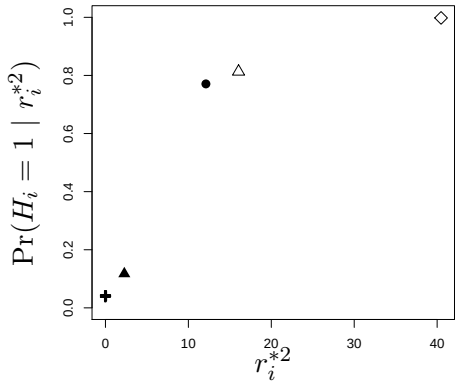


(e) Prior $V = 16$, Beta$(9.41, 4.03)$

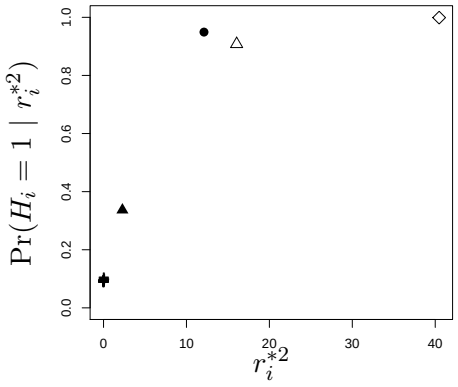


(f) Prior $V = 16$, Beta$(8.09, 8.09)$

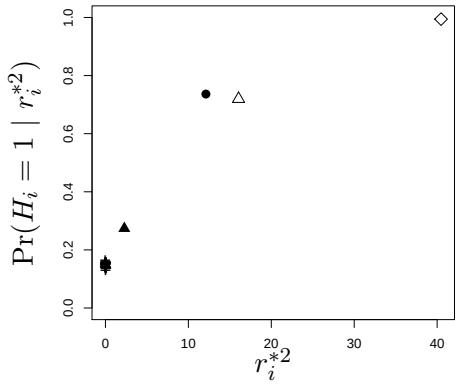
Figure 4.3: The posterior probability $P(H_i = 1 \mid r_i^{*2})$ is plotted as a function of r_i^{*2} for $V = 16$ and six different Beta priors on π_0. The symbols refer to: (•), HHV-4; (▲), HHV-6; (◊), Variola virus; (△), HHV-5; (+), other viral data point. The data under analysis is shown in Table 4.2.



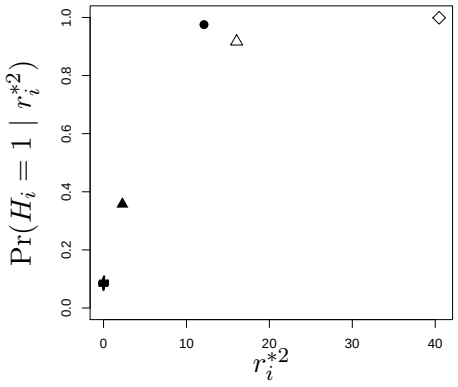
(a) Prior $V = 9$, Beta$(11, 1)$



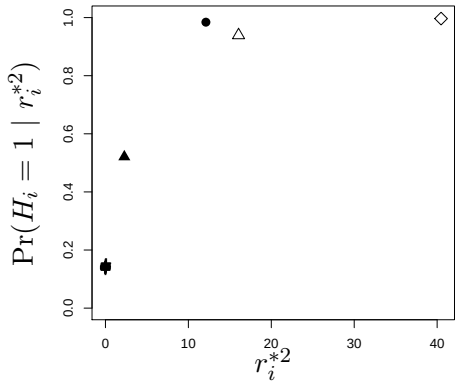
(b) Prior $V = 9$, Beta$(8, 2)$



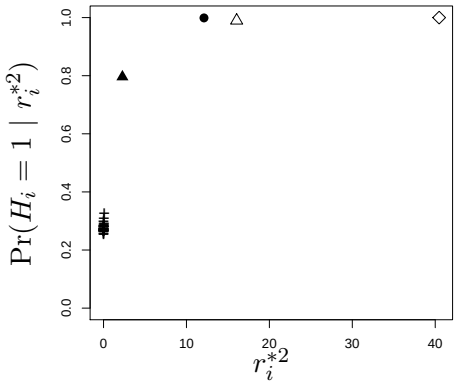
(c) Prior $V = 9$, Beta$(0.8, 0.2)$



(d) Prior $V = 9$, Beta$(80, 20)$

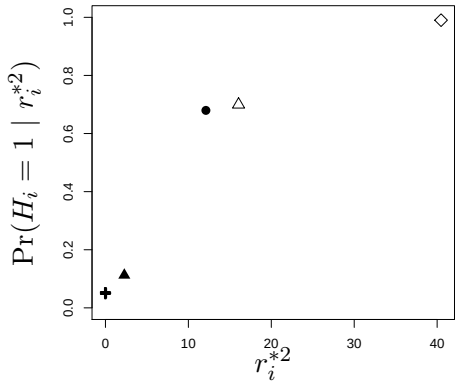


(e) Prior $V = 9$, Beta$(9.41, 4.03)$

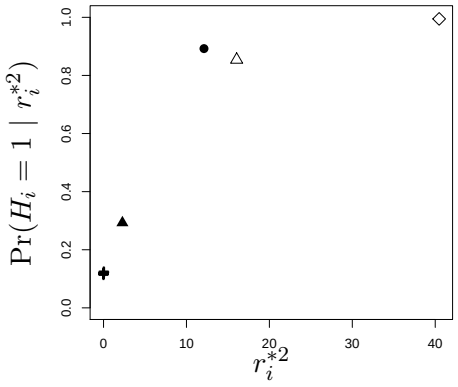


(f) Prior $V = 9$, Beta$(8.09, 8.09)$

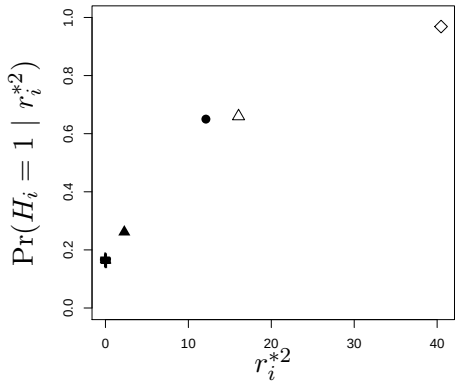
Figure 4.4: The posterior probability $P(H_i = 1 \mid r_i^{*2})$ is plotted as a function of r_i^{*2} for $V = 9$ and six different Beta priors on π_0. The symbols refer to: (•), HHV-4; (▲), HHV-6; (◊), Variola virus; (△), HHV-5; (+), other viral data point. The data under analysis is shown in Table 4.2.



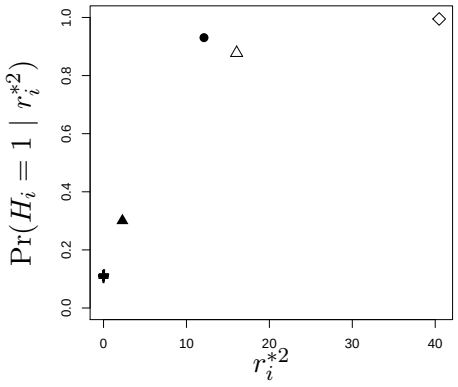
(a) Prior $V = 4$, Beta$(11, 1)$



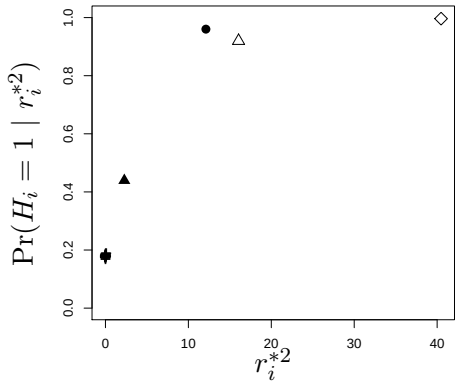
(b) Prior $V = 4$, Beta$(8, 2)$



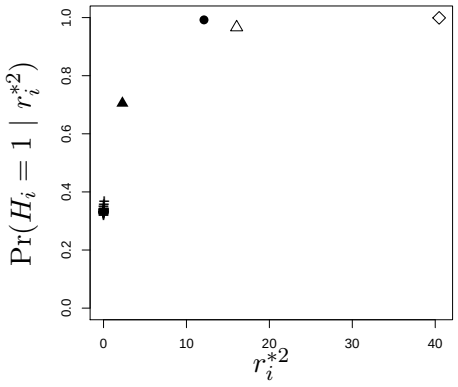
(c) Prior $V = 4$, Beta$(0.8, 0.2)$



(d) Prior $V = 4$, Beta$(80, 20)$



(e) Prior $V = 4$, Beta$(9.41, 4.03)$



(f) Prior $V = 4$, Beta$(8.09, 8.09)$

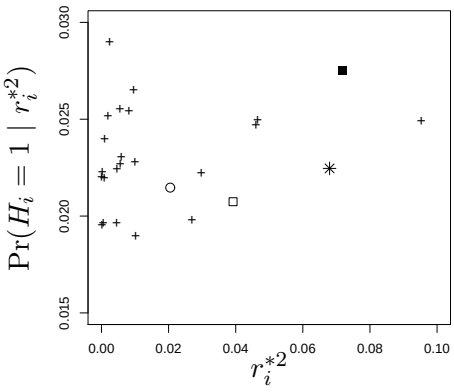
Figure 4.5: The posterior probability $P(H_i = 1 \mid r_i^{*2})$ is plotted as a function of r_i^{*2} for $V = 4$ and six different Beta priors on π_0. The symbols refer to: (•), HHV-4; (▲), HHV-6; (◊), Variola virus; (△), HHV-5; (+), other viral data point. The data under analysis is shown in Table 4.2.

likely to be outliers than the other viruses. The FEL viruses are *influential observations*, of which the explanatory variable x has much greater value than the others. In fact, even on the log-log scale of Figure 4.1, the FEL viruses are far from the other observation. Note that among the FEL viruses with large posteriors, only HHV-4 and HHV-5 are above the regression line in Figure 4.1, which means among the examined viruses, HHV-4 and HHV-5 share more pentapeptides with human proteome than the others, but HHV-6 and Variola virus represent fewer viral overlaps to the human proteome.

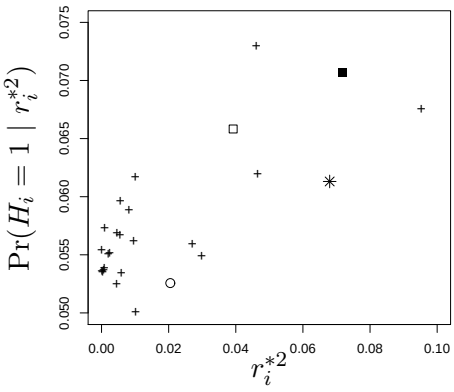
Although the FEL viruses are separated from the other data points in all graphs in Figure 4.2 – 4.5, there are still some differences among the results obtained by using different priors. The results of Beta$(11, 1)$ in Figure 4.2 – 4.5 (a) are similar to the results of Beta$(0.8, 0.2)$ in (b), though the distance between the FEL viruses and the other viruses is greater but the 26 viruses with smaller proteome size are closer for Beta$(11, 1)$. By comparing the results for Beta$(0.8, 0.2)$, Beta$(8, 2)$ and Beta$(80, 20)$ in Figure 4.2 – 4.5 (b) – (d), we can see as the variance of the Beta prior increases, the distance between the extreme observations and the others decreases but the variation among the non-extreme observations increases. In Figure 4.2 – 4.5 (b), (e), and (f), it can be observed that the extreme observations are closer to the others and the variation among the non-extreme observations becomes smaller as the mean of the Beta prior becomes larger. These results mean that the prior with smaller mean or greater variance is more sensitive to the data, and the extreme observations are more extreme by using the prior with greater mean or smaller variance.

Then I compare Figure 4.2 – 4.5 for the results of different V. As V decreases, it can be observed that the posteriors of the FEL observations decrease but the posteriors of the other 26 observations increase. The 26 non-extreme observations have smaller variation when V is smaller. However, the orders of the posteriors $P(H_i = 1 \mid r_i^{*2})$ of the FEL are almost the same for various choices of Beta priors and V, and all Beta priors and all prior variances of the mean shift could result in the same rejections of the viruses if appropriate thresholds are chosen for them. For example, if the same rejection threshold 0.1 is used for the posteriors, then the FEL viruses are identified as outliers for all 24 cases.

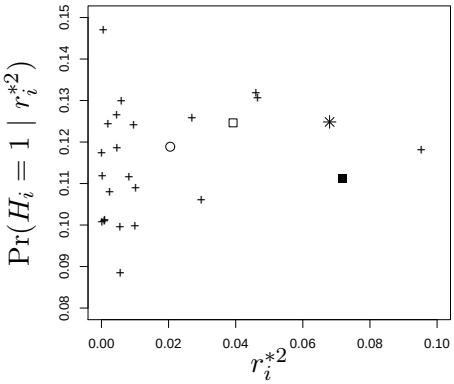
Since the 26 viruses with small proteome size are gathered together and are hard to read in Figure 4.2 – 4.5, the lower ends of all plots are magnified and shown in Figure 4.6 – Figure 4.9.



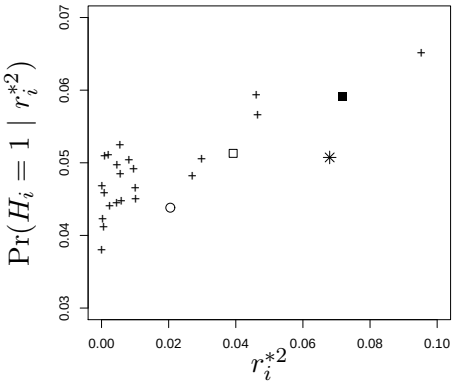
(a) Prior $V = 36$, Beta$(11, 1)$



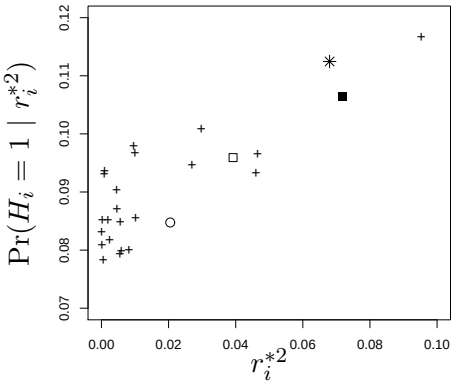
(b) Prior $V = 36$, Beta$(8, 2)$



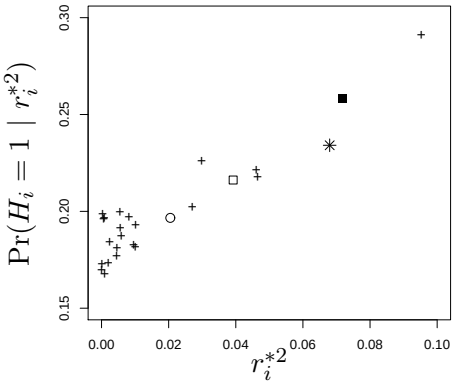
(c) Prior $V = 36$, Beta$(0.8, 0.2)$



(d) Prior $V = 36$, Beta$(80, 20)$

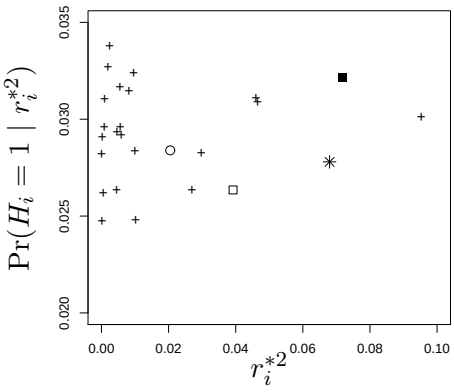


(e) Prior $V = 36$, Beta$(9.41, 4.03)$

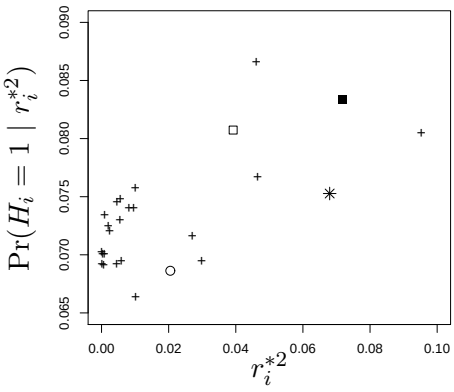


(f) Prior $V = 36$, Beta$(8.09, 8.09)$

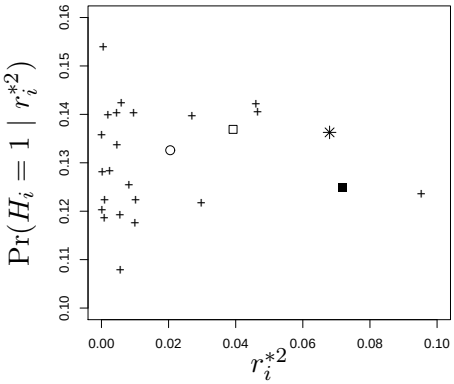
Figure 4.6: The graphs (a) - (f) magnify, respectively, the lower ends of plots (a) - (f) of Figure 4.2. The symbols refer to: ($*$), HTLV-1; (■), Rubella virus; (□), HCV; ($\circ$), Lake victoria marburgvirus; ($+$), other viral data point.



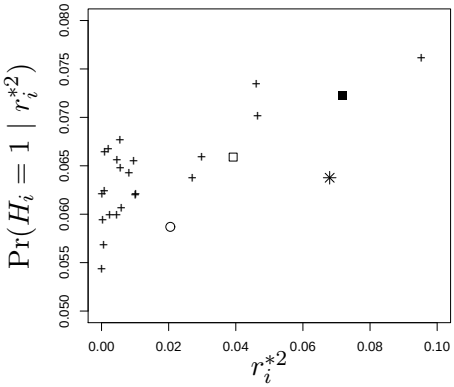
(a) Prior $V = 16$, Beta$(11, 1)$



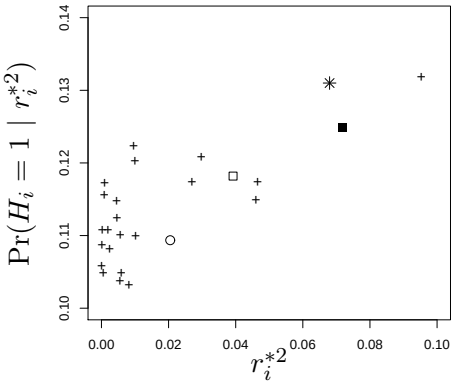
(b) Prior $V = 16$, Beta$(8, 2)$



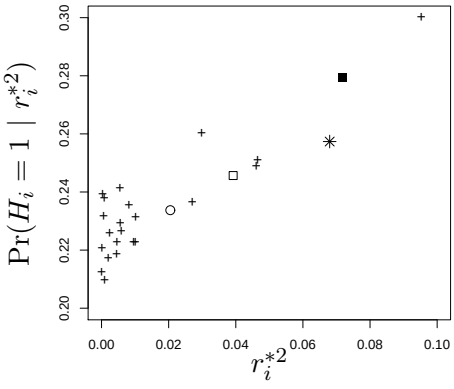
(c) Prior $V = 16$, Beta$(0.8, 0.2)$



(d) Prior $V = 16$, Beta$(80, 20)$

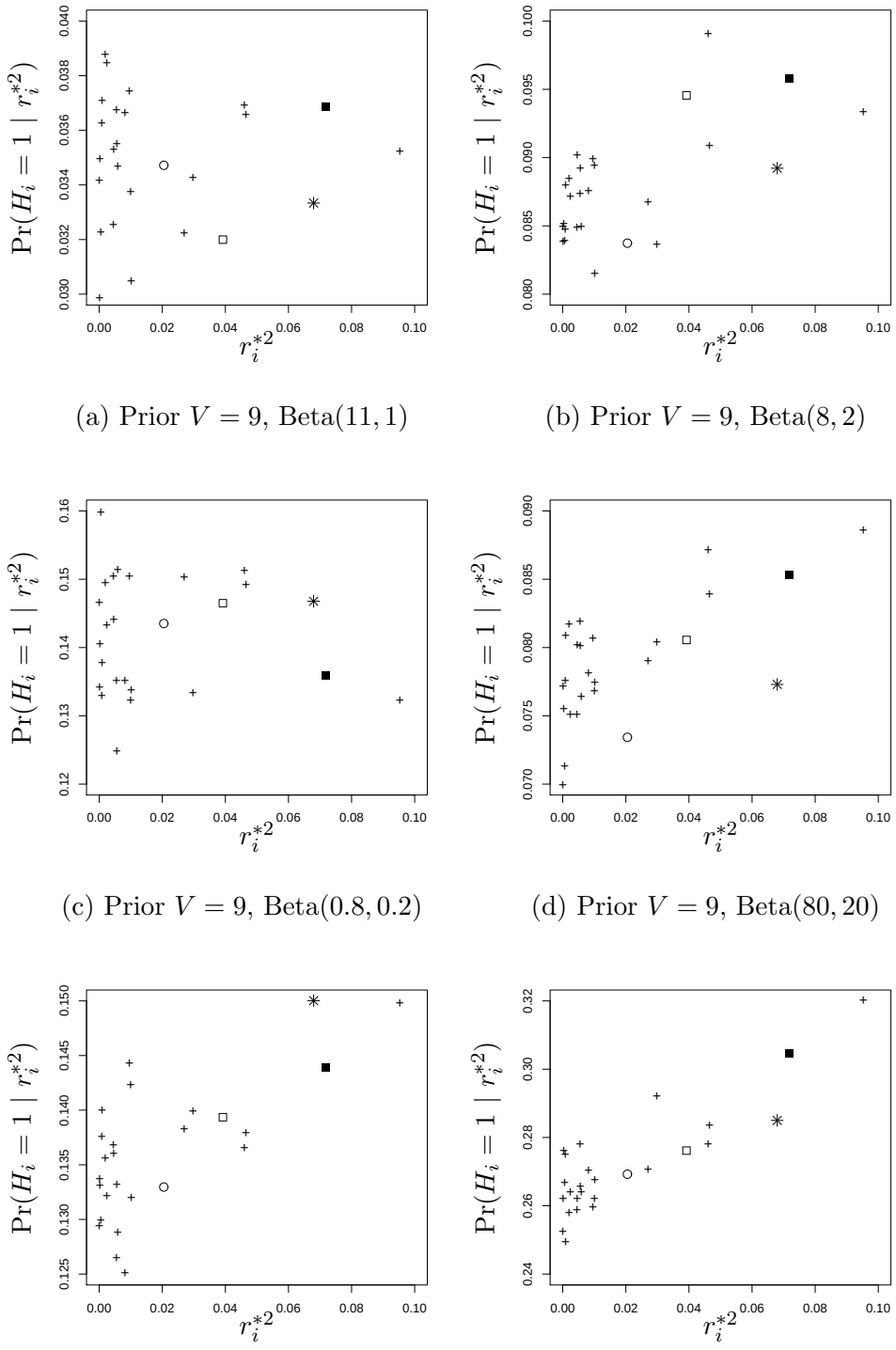


(e) Prior $V = 16$, Beta$(9.41, 4.03)$



(f) Prior $V = 16$, Beta$(8.09, 8.09)$

Figure 4.7: The graphs (a) - (f) magnify, respectively, the lower ends of plots (a) - (f) of Figure 4.3. The symbols refer to: ($*$), HTLV-1; (■), Rubella virus; (□), HCV; ($\circ$), Lake victoria marburgvirus; ($+$), other viral data point.



(a) Prior $V = 9$, Beta$(11, 1)$

(b) Prior $V = 9$, Beta$(8, 2)$

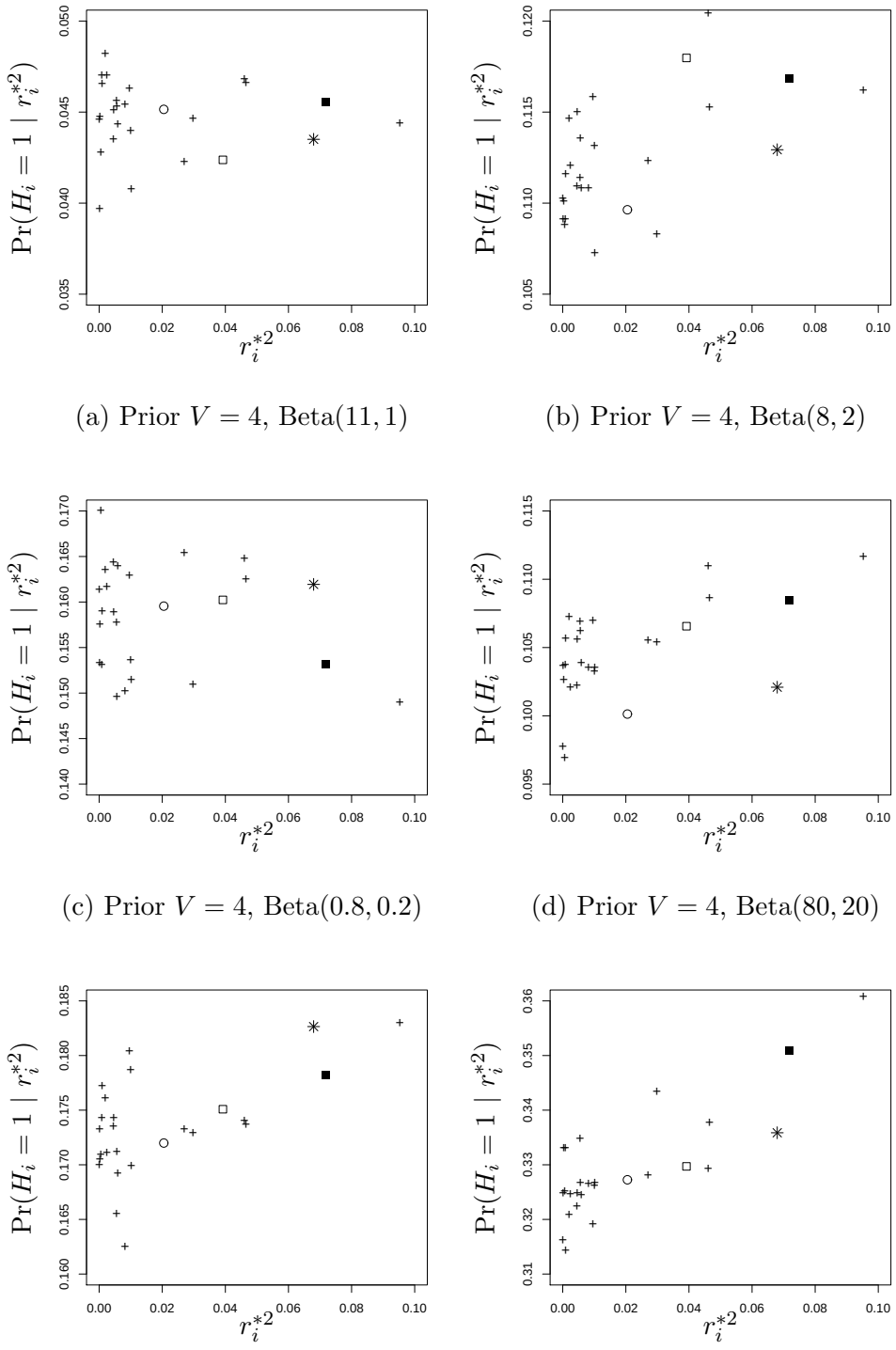
(c) Prior $V = 9$, Beta$(0.8, 0.2)$

(d) Prior $V = 9$, Beta$(80, 20)$

(e) Prior $V = 9$, Beta$(9.41, 4.03)$

(f) Prior $V = 9$, Beta$(8.09, 8.09)$

Figure 4.8: The graphs (a) - (f) magnify, respectively, the lower ends of plots (a) - (f) of Figure 4.4. The symbols refer to: ($*$), HTLV-1; (■), Rubella virus; (□), HCV; (○), Lake victoria marburgvirus; (+), other viral data point.



(a) Prior $V = 4$, Beta$(11, 1)$

(b) Prior $V = 4$, Beta$(8, 2)$

(c) Prior $V = 4$, Beta$(0.8, 0.2)$

(d) Prior $V = 4$, Beta$(80, 20)$

(e) Prior $V = 4$, Beta$(9.41, 4.03)$

(f) Prior $V = 4$, Beta$(8.09, 8.09)$

Figure 4.9: The graphs (a) - (f) magnify, respectively, the lower ends of plots (a) - (f) of Figure 4.5. The symbols refer to: ($*$), HTLV-1; (■), Rubella virus; (□), HCV; (○), Lake victoria marburgvirus; (+), other viral data point.

In Figure 4.6 – 4.9, the KI viruses do not have the three highest values of the posterior. The only exception occurs for HTLV-1 when $V = 9$ and the prior of π_0 is chosen to be Beta$(9.41, 4.03)$ (Figure 4.8, (e)). This observation suggests that the three viruses may still fail to be rejected even if a small rejection threshold is used for the posteriors. Moreover, the orders of the posteriors change dramatically for the different Beta priors, but not much for the same Beta prior and different V, indicating that when influential points exist, the posterior probability $P(H_i = 1 \mid r_i^{*2})$ for non-extreme observations is more sensitive to the prior of π_0 than to the prior of μ_i.

In Section 3.5.1, I compared results obtained by using both Patnaik's approximation and Algorithm 3.4.2 proposed in Section 3.4 for a simulated dataset with 100 observations, of which 10 are outliers. Since the number of observation and that of outliers are different from the simulated dataset, I also make the same comparison by applying both methods to the dataset in Figure 4.1. The results are included in Appendix B. There is not much difference between the results obtained by using Patnaik's approximation and those by using Algorithm 3.4.2.

4.3.2 Analysis of the reduced dataset

Since the FEL viruses are influential observations as indicated by the results given in Section 4.3.1, and they have much larger proteome size than the other viruses as shown in Figure 4.1, I remove these four viruses from the dataset in Table 4.2 and recompute the regression model for the reduced dataset. The length of viruses in the reduced dataset varies from $1,589$ to $9,739$. The scatter plot with the regression line for the reduced dataset is given in Figure 4.10, and this plot is not on a log-log scale.

The linear regression equation for the reduced dataset is $y = 10.60x + 6325.91$ and the coefficient of determination for the regression model is $R^2 = 0.9132$. The slope does not change much but the intercept increases remarkably compared to the regression equation for the full dataset. The coefficient of determination for the reduced dataset is actually smaller than that for the full dataset. This is a typical situation discussed in most textbooks where outliers increase the linear correlation between response and explanatory variables. It can be observed from Figure 4.10 that four viruses, the three KI viruses, and Lake victoria marburgvirus have larger distance to the regression lines than the other points, so they seem to be apparent outliers. In order to identify outliers in this reduced dataset, I

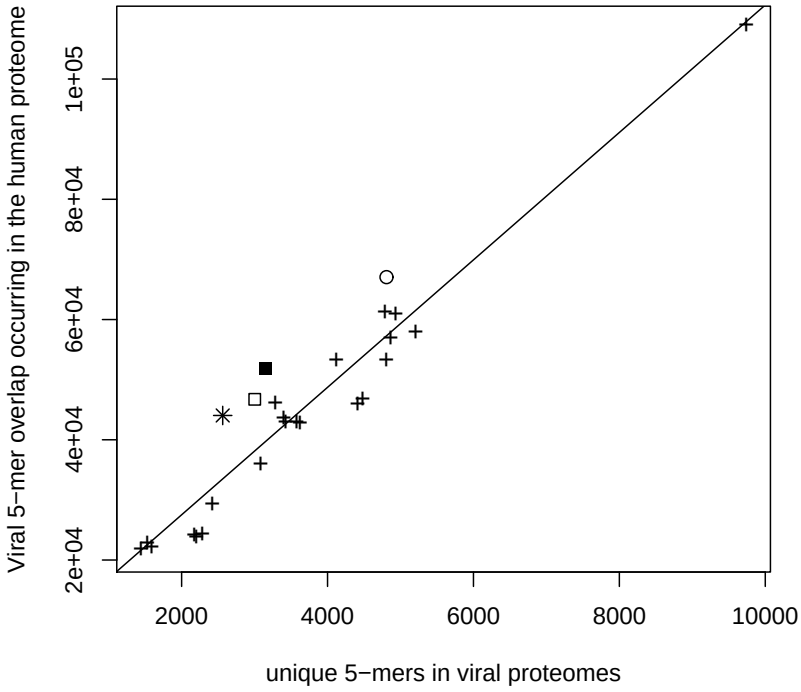


Figure 4.10: The scatter plot of the viral pentapeptide overlap including duplicates in the human proteome versus the unique pentapeptides in the viral proteome. The data under analyzing come from Table 4.2, columns 1 and 4 with HHV-4, HHV-6, Variola virus and HHV-5 deleted. The symbols refer to: ($*$), HTLV-1; (■), Rubella virus; (□), HCV; ($\circ$), Lake victoria marburgvirus; (+), other viral data point. The regression line (—) has an equation of $y = 10.60x + 6325.91$ with a coefficient of determination, $R^2 = 0.9132$.

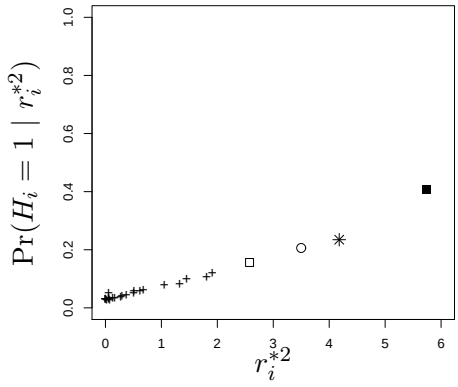
redo the computation in Section 4.3.1 by using Algorithm 3.3.1 combined with Patnaik's approximation. Figures 4.11 – 4.14 present the plots of $P(H_i = 1 \mid r_i^{*2})$ as a function of r^{*2} for various priors on π_0 and μ,. The prior standard deviation of μ varies from $6, 4, 3$ to 2 in Figure 4.11 – 4.14, and the prior on π_0 is Beta$(11, 1)$, Beta$(8, 2)$, Beta$(0.8, 0.2)$, Beta$(80, 20)$, Beta$(9.41, 4.03)$ and Beta$(8.09, 8.09)$ in the graphs (a) – (f).

In Figures 4.11 – 4.14, the four viruses with the four largest posterior probabilities $P(H_i = 1 \mid r_i^{*2})$ are HTLV-1, Rubella virus, HCV, and Lake victoria marburgvirus, of which three viruses are the KI viruses.

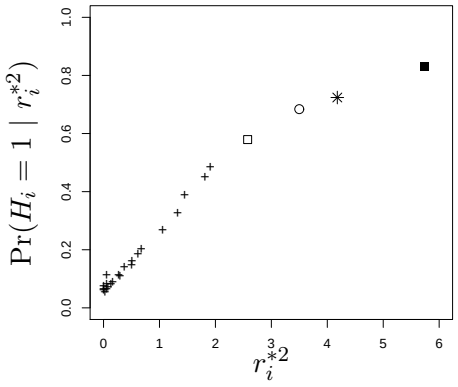
After deleting the FEL viruses, the posterior curves are more smooth and the points become dispersed. The relationship between the posterior and priors is similar to what was obtained in Section 3.5.1. The orders of the four extreme observations are almost the same in all 24 plots with different priors. By comparing the results of different Beta priors in Figures 4.11 – 4.14, we can see the values of the posterior are closer for Beta$(11, 1)$ and Beta$(0.8, 0.2)$ in graphs (a) and (c), but spread more widely for Beta$(8, 2)$ and Beta$(80, 20)$ in plots (b) and (d). It can be observed from the results for Beta$(0.8, 0.2)$, Beta$(8, 2)$ and Beta$(80, 20)$ in Figures 4.11 – 4.14 (b) – (d) that the four extreme observations are closer to the other non-extreme observations and they are also closer to each other, and the variation among the non-extreme observations becomes smaller, as the variance of the Beta prior becomes larger. In Figures 4.11 – 4.14 (b), (e), and (f), it can be observed that as the mean of the Beta prior becomes larger, the extreme observations are father from the other observations, and the variation among the non-extreme observations becomes smaller. That means, when the Beta prior becomes stronger, *i.e.*, the variance of the Beta prior becomes smaller, or its mean becomes larger, the posteriors of the extreme observations are more extreme but the posteriors of the non-extreme observations become less extreme.

Next I compare Figures 4.11 – 4.14 for the results of different V. As V decreases, the posteriors of all observations are closer, *i.e.*, the maximum posterior decreases but the minimum posterior increases. However, the posterior curves are almost the same for $V = 36$ in Figure 4.11 and $V = 16$ in Figure 4.12, indicating that V need not to be greater than 36. Moreover, the order of the four extreme values does not change for different V.

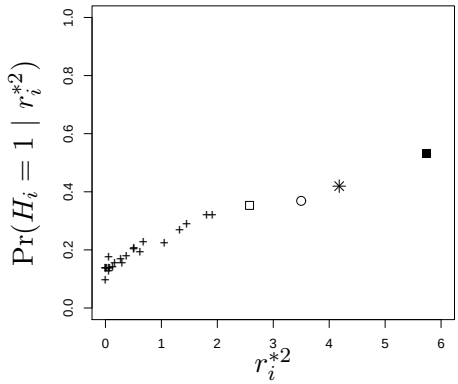
Although different priors may result in the same rejections of the viruses by choosing an appropriate rejection threshold, the farther the observations are from each other, especially



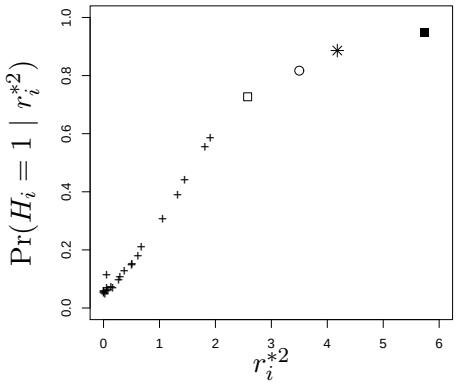
(a) Prior $V = 36$, Beta$(11, 1)$



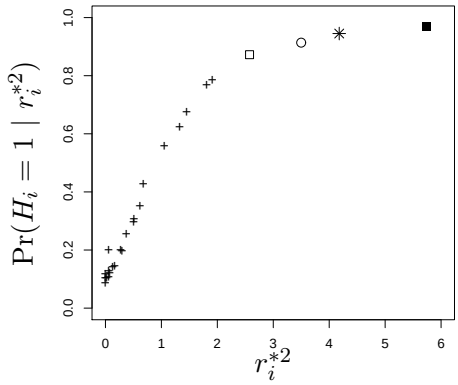
(b) Prior $V = 36$, Beta$(8, 2)$



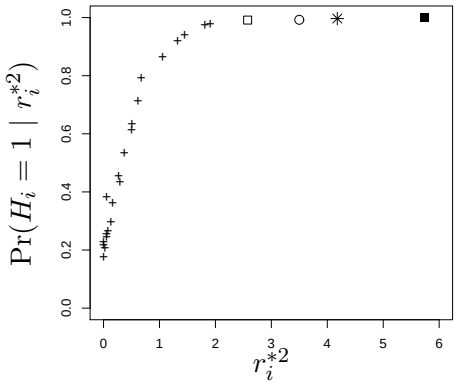
(c) Prior $V = 36$, Beta$(0.8, 0.2)$



(d) Prior $V = 36$, Beta$(80, 20)$

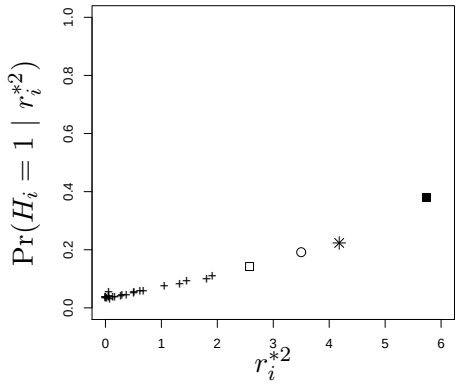


(e) Prior $V = 36$, Beta$(9.41, 4.03)$

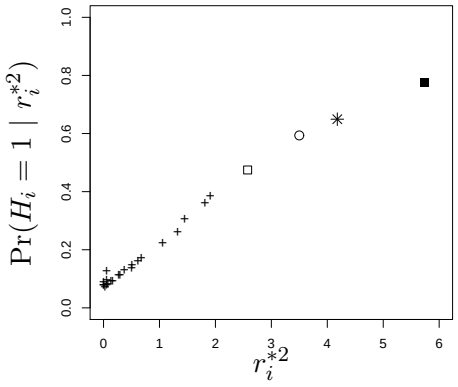


(f) Prior $V = 36$, Beta$(8.09, 8.09)$

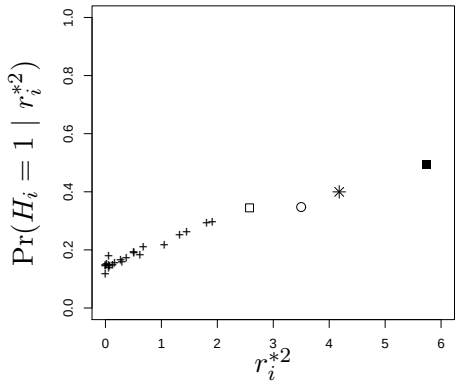
Figure 4.11: The posterior probability $P(H_i = 1 \mid r_i^{*2})$ as a function of r_i^{*2} for $V = 36$ and six different Beta priors on π_0. The data under analysis comes from Table 4.2 with HHV-4, HHV-6, Variola virus and HHV-5 deleted. The symbols refer to: $(*)$, HTLV-1; $(\blacksquare)$, Rubella virus; $(\square)$, HCV; $(\circ)$, Lake victoria marburgvirus; $(+)$, other viral data point.



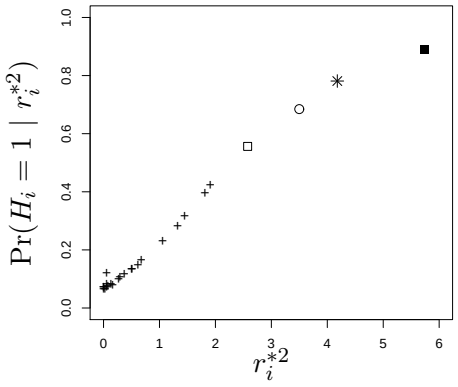
(a) Prior $V = 16$, Beta$(11, 1)$



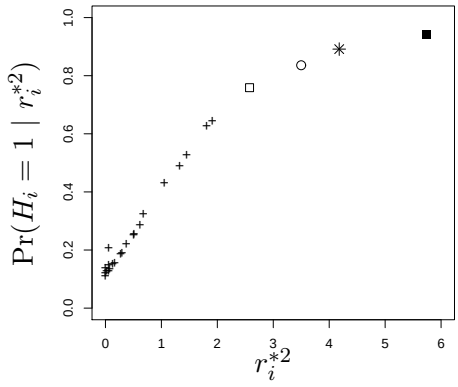
(b) Prior $V = 16$, Beta$(8, 2)$



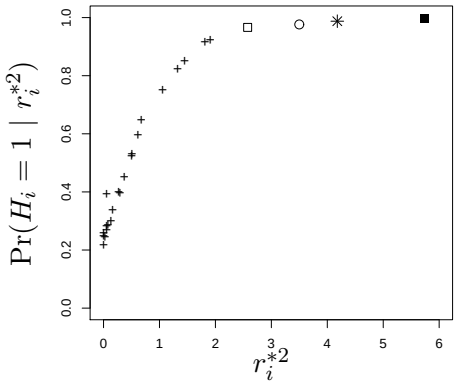
(c) Prior $V = 16$, Beta$(0.8, 0.2)$



(d) Prior $V = 16$, Beta$(80, 20)$

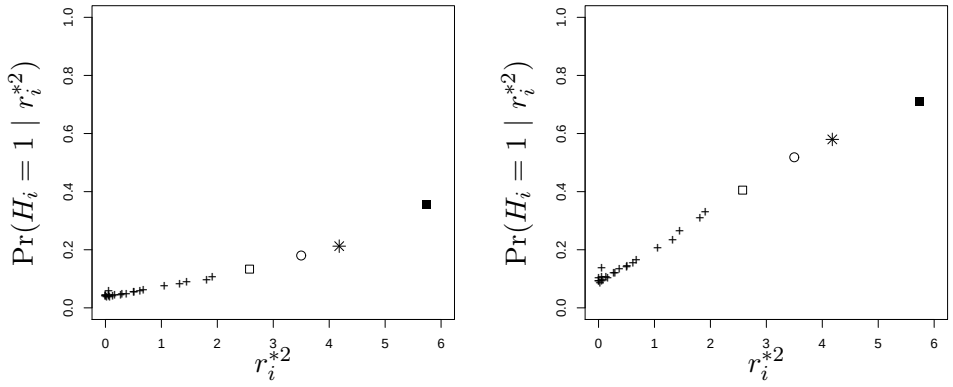


(e) Prior $V = 16$, Beta$(9.41, 4.03)$



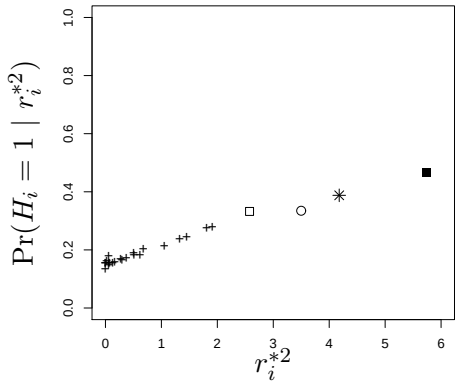
(f) Prior $V = 16$, Beta$(8.09, 8.09)$

Figure 4.12: The posterior probability $P(H_i = 1 \mid r_i^{*2})$ as a function of r_i^{*2} for $V = 16$ and six different Beta priors on π_0. The data under analysis comes from Table 4.2 with HHV-4, HHV-6, Variola virus and HHV-5 deleted. The symbols refer to: $(*)$, HTLV-1; $(\blacksquare)$, Rubella virus; $(\square)$, HCV; $(\circ)$, Lake victoria marburgvirus; $(+)$, other viral data point.

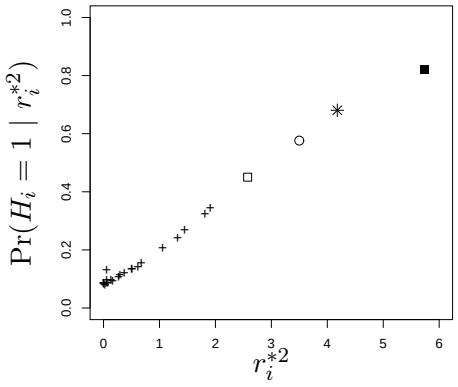


(a) Prior $V = 9$, Beta$(11, 1)$

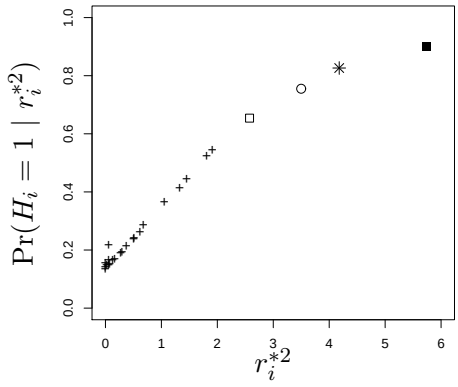
(b) Prior $V = 9$, Beta$(8, 2)$



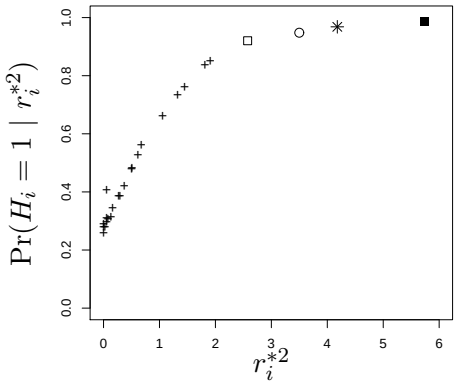
(c) Prior $V = 9$, Beta$(0.8, 0.2)$



(d) Prior $V = 9$, Beta$(80, 20)$

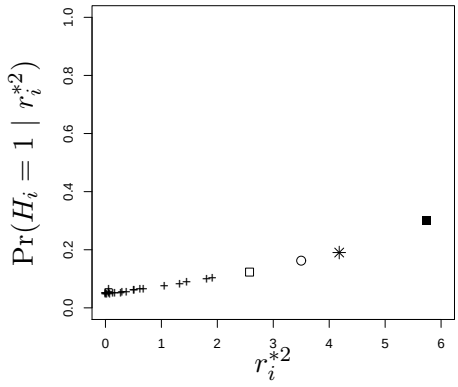


(e) Prior $V = 9$, Beta$(9.41, 4.03)$

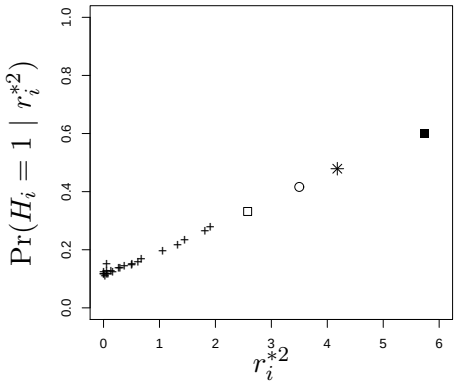


(f) Prior $V = 9$, Beta$(8.09, 8.09)$

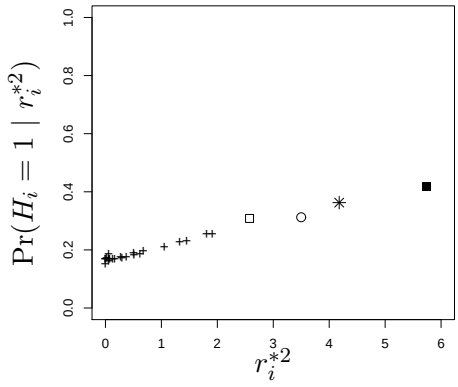
Figure 4.13: The posterior probability $P(H_i = 1 \mid r_i^{*2})$ as a function of r_i^{*2} for $V = 9$ and six different Beta priors on π_0. The data under analysis comes from Table 4.2 with HHV-4, HHV-6, Variola virus and HHV-5 deleted. The symbols refer to: $(*)$, HTLV-1; $(\blacksquare)$, Rubella virus; $(\square)$, HCV; $(\circ)$, Lake victoria marburgvirus; $(+)$, other viral data point.



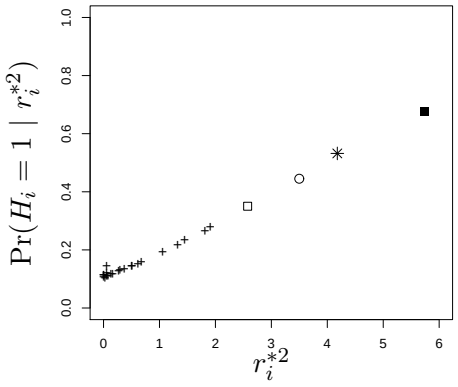
(a) Prior $V = 4$, Beta$(11, 1)$



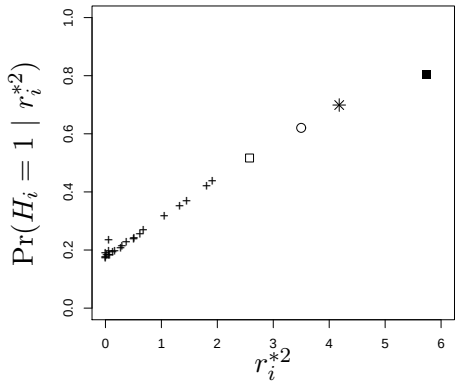
(b) Prior $V = 4$, Beta$(8, 2)$



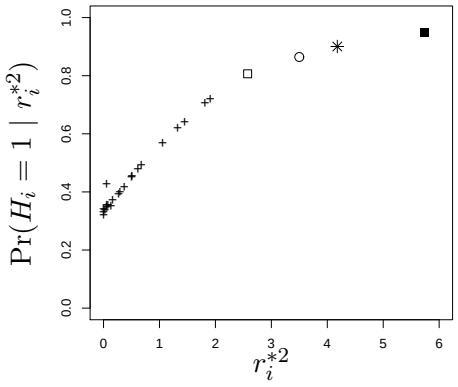
(c) Prior $V = 4$, Beta$(0.8, 0.2)$



(d) Prior $V = 4$, Beta$(80, 20)$



(e) Prior $V = 4$, Beta$(9.41, 4.03)$



(f) Prior $V = 4$, Beta$(8.09, 8.09)$

Figure 4.14: The posterior probability $P(H_i = 1 \mid r_i^{*2})$ as a function of r_i^{*2} for $V = 4$ and six different Beta priors on π_0. The data under analysis comes from Table 4.2 with HHV-4, HHV-6, Variola virus and HHV-5 deleted. The symbols refer to: $(*)$, HTLV-1; $(\blacksquare)$, Rubella virus; $(\square)$, HCV; $(\circ)$, Lake victoria marburgvirus; $(+)$, other viral data point.

for the observations with high posteriors, the better the decision that can be made, because the observations with very close posterior values may have to be rejected at the same time. In this sense, Beta$(8, 2)$ with $V = 16$ or 36 may be better than the other priors.

Although the results after deleting the FEL viruses confirm the results reported by Kanduc *et al.* [56], the KI viruses do not have the fifth to seventh largest posterior when the FEL viruses are included in the analysis, which means the KI viruses cannot be identified without excluding the four extreme observations. In other words, some outliers may be "masked" by extreme outliers by using this method. The proposed method may be improved by calculating the joint posterior probabilities $P(\boldsymbol{H} = \boldsymbol{h} \mid r_1^{*2} \cdots, r_m^{*2})$, where $\boldsymbol{h} \in \{0, 1\}^m$ instead of marginal posterior $P(H_i = 1 \mid r_i^{*2})$.

4.4 Conclusion

The sequence similarity analysis for viral proteomes to the human proteome is important for studying diseases, especially autoimmune diseases. Kanduc *et al.* [56] examined thirty proteomes for amino acid sequence similarity to the human proteome, and they found all the viral proteomes share a high number of pentapeptide overlaps to the human proteome. They also carried out a linear regression analysis to the level of overlap and the size of viral proteomes, and concluded that there is a strong linear relationship between the level of overlap and the size of viral proteomes. They also reported that three viruses, HTV-1, Rubella virus, and HCV, present the relatively highest number of viral overlaps to the human proteome.

With the purpose of identifying outliers in the dataset given in [56] and determining whether the three viruses reported by Kanduc *et al.* [56] are outliers, I applied the Bayesian method proposed in Chapter 3 to this dataset. The results show that the four viruses with extremely large size are more likely to be the outliers, and among the other 26 viruses, the KI viruses cannot be rejected without other viruses being rejected. Then I removed the FEL viruses and used the proposed Bayesian method to compute the posterior probability $P(H_i = 1 \mid r_i^{*2})$ for the reduced dataset. Among the 26 viruses in the reduced dataset, the KI viruses and Lake victoria marburgvirus have the four largest posterior probabilities of being outliers. The KI viruses and Lake victoria marburgvirus seem to present higher sequence similarity to the human proteome than the other viruses in the reduced dataset.

The results for the reduced dataset are different than for the full dataset. The reason may be that the proposed method measures the outlyingness of the ith observation by the marginal posterior $P(H_i = 1 \mid r_i^{*2})$, but the ith observation being an outlier may depend on the outlyingness of the other observations and the deletion residuals are actually dependent. Therefore, the method could be improved, if the joint distribution of all deletion residuals can be worked out, and then the joint posterior probabilities $P(\boldsymbol{H} = \boldsymbol{h} \mid r_1^{*2} \cdots, r_m^{*2})$ can be approximated by a MCMC method and used to measure the outlyingness of all observations simultaneously in the future.

In this chapter, I am interested in the distribution of the posterior probability $P(H_i = 1 \mid r_i^{*2})$ and the sensitivity of the posterior to various priors of π_0 and μ, rather than deciding which observations should be rejected. Decisions can be made by combining the calculated posterior probabilities $P(H_i = 1 \mid r_i^{*2})$ with a proper decision rule. For example, one may use the decision rules combining Bayesian FDR and FNR proposed by Muller *et al.* [64] or rules with information about the cost of two types of errors.

Chapter 5

Conclusions and future work

5.1 Conclusions

Research on multiple hypothesis testing started in the 1940's. Many methods were developed to test more than one hypothesis simultaneously. In recent years, the fast development of computer technology introduced many large and complex datasets. For example, the analysis of DNA microarray data requires testing thousands or millions of gene expression levels simultaneously. Compared to the large number of genes, the size of available samples is small. The methods introduced to test a small number of hypotheses were shown too conservative to analyze such datasets. This motivates the development of new multiple hypothesis testing techniques. Benjamini and Hochberg [9] proposed a procedure **BH** that controls FDR and then determines the rejection region as a function of FDR. Storey [89] introduced another procedure **FSL** that fixes significance level and then estimates the corresponding FDR. Black [18] considered an adaptive FDR-controlling procedure **AFDR** that incorporates an estimate of the proportion of true null hypotheses into **BH.** One objective of this thesis is to analyze these multiple hypothesis testing methods and clarify the relationship among them.

In Chapter 2, a simulation study similar to that in Storey [89] was performed to compare the FDR-controlling method of Benjamini and Hochberg with the fixed rejection region method of Storey. This simulation study used a wider range of simulation parameters, including the sample size, the proportion of true null hypotheses and the mean of the alternative distribution, than that used by Storey. The simulation results revealed that contrary to Storey's claim, **FSL** does not necessarily have more power than **BH**. In my simulation study, **BH** performed better than **FSL** when the relative number of alternatives was small. Moreover, I have proved that **BH** rejects at least the same number of hypotheses as **FSL**, and hence has at least the same power as **FSL**, when the estimate of the proportion

of true nulls is equal to one. The simulation results showed that **BH** can reject more hypotheses than **FSL**, and therefore may be more powerful than **FSL** when the number of true alternative hypotheses is relatively small. In fact, in a fair comparison, my simulations for the majority of the parameter values showed that **BH** had superior power.

BH has been shown in several papers to be conservative when the proportion of true null hypotheses is small, because the actual FDR of **BH** is proportional to the target FDR level by a factor equal to the proportion of true nulls. Black showed the FDR-controlling method of Benjamini and Hochberg could be less conservative by incorporating the estimate of the proportion of true nulls that is used by Storey. In Chapter 2, I proved that the **AFDR** procedure introduced by Black is at least as powerful as **FSL**. Then I implemented a simulation study similar to that in Black [18] but including a wider range of simulation parameters. The simulation results confirmed my result and also showed that **AFDR** can have substantially greater power than **FSL** when the proportion of true nulls is large.

At the end of Chapter 2 I presented a simulation study comparing the power of **BH** and **FSL** when their actual FDR values are set to be the same. Such a fair comparison of the two procedures has not been found in previous works. The simulation results showed that the difference in power between **BH** and **FSL** is near zero when the distance between the null and alternative distributions is large, but the former gains more power as the distance decreases. When the distance between the null and alternative distributions is small, the greater the proportion of true null hypotheses, the larger the difference in power between **BH** and **FSL**. The identification of differentially expressed genes in DNA microarray data is a motivating problem of this chapter. A feature of DNA microarray data is that most genes are not expected to be differentially expressed, *i.e.* the proportion of true nulls is close to 1. Furthermore, there may be genes that are differentially expressed, but whose distributions are close to those of the reference genes. The simulation results of the fair comparison implied that in this situation **BH** performs better than **FSL** as long as one incorporates a good estimate of the proportion of true nulls. The fair comparison results were presented for finite sample sizes (no more than 5000) since it is known that **BH** is asymptotically equivalent to **FSL,** under some mild assumptions. The simulation results for various sample sizes indicated that the power of **BH** decreases to that of **FSL** as the sample size increases to infinity.

The focus of Chapter 2 has been on the case where the multiple hypothesis tests are independent. This simplified assumption is usually invalid for real data, including microarray data. Multiple hypothesis testing can also be applied to regression diagnostics, where the p-values are dependent. In regression models, data usually contain multiple outliers and the number of outliers is unknown. Another objective of this thesis is to develop a new method to identify multiple outliers in linear regression models. In fact, the problem of identifying multiple outliers in regression analysis can be viewed as a problem of multiple hypothesis testing. Each observation can be assigned a null hypothesis that this observation does follow the assumed distribution, and an alternative hypothesis that it is an outlier and follows a distribution different from the null distribution. A powerful method to identify single outliers is based on deletion residuals. The null distribution of the deletion residual of an observation is shown to be a student t distribution when this observation is the only outlier in a given data set.

In Chapter 3, I assumed that the random error of an atypical observation follows a normal distribution with a mean shift, whereas that of a typical observation follows a normal distribution with mean equal to zero. Then I have proved a new result that for the proposed model, the marginal distribution of the deletion residual of an observation under both null and alternative hypotheses are doubly noncentral t distributions, when there exists more than one outlier. The non-central parameters of the doubly noncentral t distributions depend on the number of outliers in the dataset. Consequently, the marginal distributions of the square of the deletion residual are the doubly noncentral F distributions under both null and alternative hypotheses. Then I proposed a Bayes method assuming that the proportion of typical observations follows a Beta prior distribution, and the mean shift of outliers has a normal prior distribution. The outlyingness of an observation can be measured by the marginal posterior probability that the ith observation is an outlier given its own deletion residual. In this chapter, I have also proposed an importance sampling method to calculate this marginal posterior probability. This algorithm involves the computation of the density of the doubly noncentral F distribution, which is achieved by using Patnaik's approximation. In order to examine the accuracy of Patnaik's approximation to the density of the doubly noncentral F distribution, I have proposed an algorithm to compute this density. Both methods were applied to various noncentral parameters and quantiles with the sample size fixed to 100. The largest difference between

the density calculated by Patnaik's approximation and that by the proposed algorithm is 0.0014. I also incorporated both Patnaik's approximation and the proposed algorithm with the proposed importance sampling method, to calculate the required marginal posterior probability, for a simulated dataset. The maximum difference of densities computed by the two methods is no larger than 0.071, and the maximum difference of the posteriors is smaller than 6.22×10^{-5}. The shortest CPU time used, by incorporating the proposed algorithm to calculate the posteriors of the given dataset with certain priors, is 6261.31 seconds, whereas that by incorporating Patnaik's approximation is 72.39 seconds. The results showed that using Patnaik's approximation to calculate the doubly noncentral F density can save massive computation time without losing much accuracy.

In the last section of Chapter 3, the proposed Bayesian method was applied to some simulated datasets. Various simulation parameters, including the total number of observations, the proportion of outliers and the variance of the mean shift, were used for the simulation study. In order to study the sensitivity of the posteriors to the priors, I selected several Beta priors of the proportion of typical observations and various variances of the normal prior of the mean shift. Then the Bayes samples, described in the proposed importance sampling, were generated from various prior distributions for each simulated dataset. I firstly implemented the proposed procedure to two single datasets. Each dataset was generated with the same sample size and the same proportion of outliers but different variance of the mean shift. Although the explanatory variable is not random in a linear regression model, it was assumed to follow a Bernoulli distribution for one dataset and the standard normal distribution for the other. For each combination of priors, the marginal posterior probability $P(H_i = 1 \mid r_i^{*2})$ was plotted as a function of the deletion residual r_i^{*2}, the ROC curves were also plotted, and the AUC values were calculated. Although the plots of the posteriors are different for various priors, the AUC values are close for selected priors and the smallest AUC is 0.7678. The high AUC values indicate that the proposed method can identify the majority of outliers with tolerable error. The similar AUC values for various choices of priors indicate that the posterior probability is not very sensitive to the chosen priors. However, the results for certain datasets may not be true for another dataset. Therefore, the proposed Bayesian method was secondly applied to two sets of data, each including 1000 iterations. The simulation and the prior parameters were the same as those used for the two single datasets. The ROC curves of the average TPR versus

FPR for selected thresholds were plotted and the average AUC over 1000 replicates were calculated for different priors. The average AUC values are close for selected priors and the smallest AUC is 0.7994. The results also showed that the proposed Bayesian multiple outlier identification method performs well, and there is not much effect of the priors on the posterior. It took a long time to calculate the average AUC for 1000 iterations for each combination of simulation parameters. Hence I thirdly employed a factorial design analysis to compare AUC by choosing various simulation and priors parameters as factors. This simulation study included a much wider range of simulation and priors parameters. All the AUC values for the simulated datasets are greater than 0.7 and the grand mean is equal to 0.8, indicating the proposed method can identify a majority of the outliers. When a small value of 20 for the factor sample size was included in the analysis, the resulting ANOVA table indicated a significant four-way interaction among $m \times \pi_0 \times \sigma^2 V' \times a$, where $\sigma^2 V'$ is the prior variance of the mean shift and a is the hyper-parameter of the Beta prior of π_0. This result showed that the priors may affect the posterior. There is also a significant main effect of m in the ANOVA table, and both the table of means and the residual plot suggest a large difference between 20 and the other values of the sample size. Hence I removed the data with $m = 20$ and performed the factorial design analysis again, and there is no longer any significant effect involving the factors of the prior parameters. These results showed that the priors do not affect the marginal posterior probability $P(H_i = 1|r_i^{*2})$ as long as the sample size is not too small.

In Chapter 4, the Bayesian method proposed in Chapter 3 was applied to a dataset given in Kanduc *et al.* [56]. Kanduc *et al.* compared the amino acid sequences of thirty proteomes to those of the human proteome, and they found all the viral proteomes share a high number of pentapeptide overlaps to the human proteome. They also performed a linear regression analysis to the level of overlap and the size of viral proteomes and concluded that three viruses, human T-lymphotropic virus 1, Rubella virus, and hepatitis C virus, present the highest relative number of viral overlaps to the human proteome. With the purpose of identifying outliers in the dataset given in [56] and determining whether the three viruses reported by Kanduc *et al.* are outliers, I implemented the Bayesian approach to this dataset. To examine how sensitive the posterior distribution is to the prior distributions, various prior distributions of the proportion of typical observations and various prior variances of the mean shift were used when generating Bayes samples.

The marginal posterior probability $P(H_i = 1 \mid r_i^{*2})$ was plotted as a function of the deletion residual r_i^{*2} for various choices of priors. All plots of the marginal posterior indicated that the four viruses with extremely large sizes, which do not include the three viruses reported by Kanduc *et al.*, are more likely to be the outliers. The magnified lower ends of the plots of the marginal posterior are distinct for various priors, and among the other 26 viruses, the three reported viruses still do not have a larger size than the others. The results showed that the three reported viruses cannot be rejected without other viruses being rejected, even if the four extreme viruses are rejected .

Then I removed the four viruses with extremely large size and used the proposed Bayesian method to compute the posterior probability $P(H_i = 1 \mid r_i^{*2})$ for the reduced dataset. The prior distributions were also varied among the same range as for the full dataset. The results for the reduced dataset confirmed the claim of Kanduc *et al.* [56]. Among the 26 viruses in the reduced dataset, three viruses and another virus have the four largest posterior probabilities of being outliers. The three viruses reported by Kanduc *et al.* [56] and Lake victoria marburgvirus seem to present higher sequence similarity to the human proteome than the other viruses.

5.2 Future Work

It is shown in Chapter 2 that modifying the **BH** to incorporate an estimate of the proportion of true null hypotheses as proposed by Black gives a procedure **AFDR** with superior power than **FSL**. However, to implement this, an estimate of π_0 is needed. My future investigations will examine the effects of using alternative methods for estimating π_0, such as those proposed by Benjamini and Hochberg [10], Efron et al [30], Storey and Tibshirani [93], and Bickis [15].

In Chapter 3, an importance sampling method was used to calculate the marginal posterior probability $P(H_i = 1 \mid r_i^{*2})$, where the importance function was chosen to be the joint prior density. The performance of the importance sampling would be poor if some importance ratios are much larger than the others. I plotted the histogram of the logarithms of the importance ratios for some simulated datasets and the concerned problems did not occur. The distributions of sampled importance ratios for all simulated datasets need to be examined in the future. My future work is to use other importance functions and compare

the results of using different importance functions. For example, we can use Gibbs sampling methods to approximate the joint posterior distribution of parameters.

In Chapter 3 and 4, no decision rule was developed. The calculated marginal posterior probabilities need to be combined with a proper decision rule to decide which observations are going to be rejected. Different decision rules can be considered, for instance, the decision rules combining Bayesian FDR and FNR proposed by Muller *et al.* [64].

Atkinson [2] recently proposed a forward search to identify multiple outliers and this method is introduced in detail in the books [3] and [6]. The "R" package "forward" is available to fulfill the forward search of Atkinson [67]. I also plan to compare the proposed Bayesian multiple outlier identification procedure with the forward search of Atkinson.

When the Bayesian multiple outlier identification procedure proposed in Chapter 3 is applied to the dataset given in Kanduc *et al.* [56], the results for the reduced dataset are different from those for the full dataset. This indicates that the proposed procedure has some limitations. Addressing them leads to my future work.

First the outlyingness of the ith observation is measured by its marginal posterior and the correlation between deletion residuals is ignored. Indeed, the ith observation being an outlier may also depend on the outlyingness of the other observations, and the deletion residuals are actually dependent. The joint posterior probabilities $P(\boldsymbol{H} = \boldsymbol{h} \mid r_1^{*2} \cdots, r_m^{*2})$ can provide a measurement of outlyingness for all observations simultaneously. These joint posterior probabilities can be approximated by MCMC methods, for example, the Gibbs sampling method [75]. The posterior distribution of the indicators $\boldsymbol{H}$ can be estimated from a large number of random samples generated from the joint posterior distribution of $\boldsymbol{H}$, $\boldsymbol{\mu}$ and π_0 which is proportional to $f_{r_1^*,\cdots r_m^*}(r_1, \cdots, r_m \mid \boldsymbol{H}, \boldsymbol{\mu}, \pi_0) p_{\boldsymbol{\mu}}(u \mid \boldsymbol{H}) p_{\boldsymbol{H}}(\omega \mid \pi_0) p_{\pi_0}(\pi)$. Hence the joint density of m deletion residuals needs to be computed. The joint distribution of all the deletion residuals is the joint distribution of m doubly non-central t random variables. The explicit expression for this joint distribution may not have a simple expression, but MCMC methods, for example, the Gibbs sampling method, may be used to approximate it. This is part of the future work.

Secondly, in the proposed Bayesian approach, the posterior distribution of $\boldsymbol{H}$ is conditional on the deletion residuals. This approach has the advantage that the distribution of r_i^* is independent of the regression coefficient $\boldsymbol{\beta}$ and noise variance σ^2. Therefore, when computing the posterior probabilities, the dimension of the integral can be reduced for

the proposed method. However, it also introduces a complex dependence structure among deletion residuals. Although the deletion residuals are ancillary for the model parameters, they are not ancillary for the outlier parameters ($\boldsymbol{H}$ and $\boldsymbol{\mu}$), since the distribution of the fitted values $\hat{\boldsymbol{\beta}}_{(i)}$ also depends on those parameters. I intend to examine the loss of information from basing the inference on the deletion residuals in my future research.

Thirdly, in the proposed Bayes model, the prior for the mean shift μ_i has a point mass at 0 when $H_i = 0$, which is subject to the problem identified in Lindley's Paradox [61]. Lindley's Paradox states that the posterior probability of $H_i = 0$ can be arbitrarily large if the prior variance of μ_i is chosen to be sufficiently large. Therefore, the posterior distribution of $\boldsymbol{H}$ is very sensitive to the prior variance of μ_i. This limitation can be conquered by assuming that the prior variance of mean shifts is case-specific and follows a inverse gamma distribution with parameters $\nu/2$ and $\nu s_0/2$, and then the prior distribution of μ_i is a heavy-tailed t distribution with small degrees of freedom ν and scale parameter s_0. Such a prior is appropriate since most of μ_i are expected to be very close to 0 (but may be different from 0), while a few of them are extraordinarily large. The proposed Bayes model can be modified to allow a normal and Inverse-Gamma distribution on the regression coefficients and the variance, respectively. Since the observations given corresponding values of explanatory variables are assumed to be independent, the joint distribution of m observations is equal to the product of m marginal distributions. Then the joint posterior distribution of $\boldsymbol{\mu}$, $\boldsymbol{\beta}$ and σ^2 can be sampled, by using Gibbs sampling method, from the distribution proportional to the joint distribution of data and parameters $\prod_{i=1}^{m} \left[\phi_{y_i}(y \mid \boldsymbol{\beta}, \mu_i, \sigma^2)\, t_{\mu_i}(u \mid \nu, s_0)\right] p_{\boldsymbol{\beta}}(B) p_{\sigma^2}(\tau)$, where $\phi_{y_i}(\cdot \mid \mu_i, \sigma^2)$ is the normal density function of y_i with the mean $\boldsymbol{x}_i^T \boldsymbol{\beta} + \mu_i$ and the variance σ^2, $t_{\mu_i}(\cdot \mid \nu, s_0)$ is the scaled t density function of μ_i with ν degrees of freedom and scale parameter s_0, and $p_{\boldsymbol{\beta}}$ and p_{σ^2} are the prior densities of $\boldsymbol{\beta}$ and σ^2. For computational reasons, this model can also be expressed by $\prod_{i=1}^{n} \left[\phi_{y_i}(y \mid \boldsymbol{\beta}, \mu_i, \sigma^2)\, \phi_{\mu_i}(u \mid s_i)\, \mathrm{IG}_{s_i}(s \mid \nu/2, \nu s_0/2)\right] p_{\boldsymbol{\beta}}(B) p_{\sigma^2}(\tau)$, where s_i is the prior variance of μ_i and $\mathrm{IG}_{s_i}(\cdot \mid \nu/2, \nu s_0/2)$ denotes the Inverse-Gamma density function of s_i with parameters $\nu/2$ and $\nu s_0/2$. Then all the conditional distributions can be sampled directly with standard methods in Gibbs sampling framework. The outlyingness of an observation can be measured by the posterior distribution of s_i or by that of μ_i.

Bibliography

[1] Atkinson, A. C. (1985). Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis. Oxford Statistical Science Series, Oxford University Press, New York.

[2] Atkinson, A. C. (1994). Fast very robust methods for the detection of multiple outliers. Journal of the American Statistical Association. Vol. 89, No. 428, Theory and Method, 1329-1339.

[3] Atkinson, A. C. and Riani, M. (2000). Robust diagnostic regression analysis. Spring-Verlag, New York.

[4] Atkinson, A. C. and Riani, M. (2002). Tests in the fan plot for robust, diagnostic transformations in regression. Chemometrics and Intelligent Laboratory Systems, Vol. 60, 87-100.

[5] Atkinson, A. C. and Riani, M. (2006). Distribution theory and simulations for tests of outliers in regression. Journal of Computational and Graphical Statistics, Vol. 15, No. 2, 460-476.

[6] Atkinson, A. C., Riani, M. and Cerioli, A. (2004). Exploring multivariate data with the forward search. Spring-Verlag, New York

[7] Barnett, V. and Lewis, T. (1994), Outliers in Statistical Data (3rd edition). John Wiley & Sons.

[8] Bayarri, M. J. and Berger, J. O. (2004). Multiple testing (the problem and some solutions). Conference presentation. International Society for Bayesian Analysis.

[9] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, B 57, 289-300. Mathematical Reviews (MathSciNet): MR1325392

[10] Benjamini, Y. and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. Journal of Educational and Behavioral Statistics, Vol.25, 60-83.

[11] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. The Annals of Statistics, Vol.29, No.4, 1165-1188. Mathematical Reviews (MathSciNet): MR1869245

[12] Berger, J. O. (1985). Statistical Decision Theory and Bayesian Analysis (2nd edition). Springer, New York.

[13] Berger, J. O. (2006). The Case for Objective Bayesian Analysis. Bayesian Analysis, Vol. 1, No.3, 385-402. Mathematical Reviews (MathSciNet): MR2221271

[14] Beger, J. O. and Pericchi, L. R. (2001). Objective Bayesian Methods for Model Selection: Introduction and Comparison. Lecture Notes-Monograph Series, Vol. 38, Model Selection (2001), 135-207.

[15] Bickis, M., (2004). Coping with multiplicity by exploiting the empirical distribution of P-values. Contributed paper, 6th World Congress of the Bernoulli Society & 67th Annual Meeting of the Institute of Mathematical Statistics Barcelona, Spain, July 29.

[16] Bickis, M. and Krewski, D., (1989). Statistical issues in the analysis of the long-term carcinogenicity bioassay in small rodents: an empirical evaluation of statistical decision rules. Fundamental and Applied Toxicology, Vol. 12, Issue 2, 202-221.

[17] Bickis, M., Bleuer, S. and Krewski, D. (1996). On the estimation of the proportion of positives in a sequence of screening experiments. The Canadian Journal of Statistics, Vol.24, No.1, 1-15.

[18] Black, M. A. (2004). A note on the adaptive control of false discovery rates. Journal of the Royal Statistical Society, B 66, 297-304.

[19] Butler R. W. and Paolella M. S. 2002. Calculating the density and distribution function for the singly and doubly noncentral F. Statistics and Computing, Vol. 12, Issue 1, 9-16.

[20] Chattamvelli, R. (1995). On the doubly noncentral F ditribution. Computational Statistics and Data Analysis, Vol. 20, 481-489.

[21] Chen, X. (1999). Advanced statistical inference. University of Science and Technology of China Publishing House, Hefei, China.

[22] Chi, Z. (2007). On the performance of FDR control: constraints and a partial solution. Ann. Statist. 35, 4, 1409–1431. Mathematical Reviews (MathSciNet): MR2351091

[23] Cohen, A. and Sackrowitz, H. B. (2005). Decision theory results for one-sided multiple comparison procedures. Ann. Statist. 33, 126-144.

[24] Cohen, A. and Sackrowitz, H. B. (2005). Characterization of Bayes procedures for multiple endpoints problems and inadmissibility of the step-up procedures. The Annals of Statistics, Vol. 33, 145-158.

[25] Cook, R. D. and Weisberg, S. (1982). Residuals and influence in regression. Chapman & Hall, London.

[26] Draper, N. R. and Smith, H. (1981). Applied regression analysis. 2nd edition. Wiley, New York.

[27] Dudoit, S. and van der laan, M. J. (2008). Multiple testing procedures with applications to genomics. Springer, New York.

[28] Efron, B. (2004). Large -scale simultaneous hypothesis testing: the choice of a null hypothesis. Journal of the American Statistical Association, Vol.99, 96-104. Mathematical Reviews (MathSciNet): MR2054289

[29] Efron, B. and Tibshirani, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. Genetic Epidemiology 23, 70-86.

[30] Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. (2001). Empirical Bayes Analysis of a microarray experiment. Journal of the American Statistical Association, Vol. 96, No. 456, 1151-1160. Mathematical Reviews (MathSciNet): MR1946571

[31] Einot, I. and Gabriel, K. R. (1975). A study of the powers of several methods of multiple comparison. Journal of the American Statistical Association, Vol. 70, No.351, 574-583.

[32] Feller, W. (1968). An introduction to probability theory and its applications (3rd edition, Vol. 1). John Wiley & Sons, Inc., New York.

[33] Ferreira, J. A. and Zwinderman, A. H. (2006). On the Benjamini-Hochberg method. The Annals of Statistics, Vol. 34, No.4, 1827-1849.

[34] Finner, H. and Roters, M. (2001). On the false discovery rate and expected type I errors. Biometri. J. 43, 985–1005. MR1878272

[35] Fisher, R. A. (1970). Statistical methods for research workers (14th edition). Edinburgh, Oliver and Boyd.

[36] Fishman, G. S. (1996). Monte Carlo: concepts, algorithms and applications. Springer, New York.

[37] Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004). Bayesian Data Analysis (2nd edition). Chapman & Hall/CRC Texts in Statistical Science.

[38] Genovese, C. R. and Wasserman, L. (2002a). Operating characteristics and extensions of the false discovery rate procedure. Journal of the Royal Statistical Society, B 64, 499-517. Mathematical Reviews (MathSciNet): MR1924303

[39] Genovese, C. R. and Wasserman, L. (2002b). Bayesian and frequentist multiple testing. Technical Report. Department of Statistics, Carnegie-Mellon University.

[40] Genovese, C. and Wasserman, L. (2004a). A stochastic process approach to false discovery control. The Annals of Statistics. Vol.32, No.3, 1035–1061. Mathematical Reviews (MathSciNet): MR2065197

[41] Genovese, C. and Wasserman, L. (2004b). False discovery control for random fields. Journal of the American Statistical Association, Vol. 99, No. 468, 1002-1014.

[42] Genovese, C. and Wasserman, L. (2006). Exceedance control of the false discovery proportion. Journal of the American Statistical Association, Vol. 101, No. 476, 1408-1417.

[43] Gordon, A., Glazko, G., Qiu, X. and Yakovlev, A. (2007). Control of the mean number of false discoveries, Bonferroni and stability of multiple testing. The annals of applied statistics, Vol.1, No.1, 179-190.

[44] Hadi, A. S. (1992). Identifying multiple outliers in multivariate data. Journal of the Royal Statistical Society, Series. B, Statistic Methodology, Vol. 54, 761-771.

[45] Hadi, A. S. and Simonoff, J. S. (1993). Procedures for the identification of multiple outliers in linear models. Journal of the American Statistical Association, Vol. 88, No. 424, 1264-1272.

[46] Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, Vol. 143, 29-36.

[47] Hanley, J. A. and McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology, Vol. 148, 839-843.

[48] Hawkins, D. M. (1980). Identification of outliers. Chapman & Hall, London.

[49] Hawkins, D. M. (1983). Discussion of paper by Beckman and Cook. Technometric, Vol. 25, No. 2, 155-156.

[50] Hochberg, Y. A (1988). sharper Bonferroni procedure for multiple tests of significance. Biometrika Vol. 75, 800-802.

[51] Hochberg, Y. and Tamhane, A.C. (1987). Multiple Comparison Procedures. Wiley, New York.

[52] Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, Vol. 6, 65-70.

[53] Ji, Y., Lu, Y. and Mills, G. B. (2008). Bayesian models based on test statistics for multiple hypothesis testing problems. Bioinformatics, Vol. 24, No, 7, 943-949.

[54] Johnson, V. E. (2005). Bayes factors based on test statistics. Journal of the Royal Statistical Society, Series. B, Statistic Methodology, Vol. 67, 689-701.

[55] Johnson, N. L. and Kotz, S. (1970). Distributions in Statistics: Continuous Univariate Distributions-2. Wiley, New York.

[56] Kanduc, D., Stufano, A., Lucchese, G and Kusalik, A. (2008). Peptides, Vol. 29, 1755-1766.

[57] Knudsen, S. (2002). A biologist's guide to analysis of DNA microarray data. Wiley, New York.

[58] Kusalik, A. (2009). Personal communication.

[59] Lehmann, E. L. (1986). Testing Statistical Hypotheses (2nd edition). Wiley, New York.

[60] Lehmann, E. L. and Romano, J. P. (2005). Generalizations of the familywise error rate. The Annals of Statistics, Vol.33, No.3, 1138-1154.

[61] Lindley, D. V. (1957). A Statistical Paradox. Biometrika, Vol. 44, No. 1-2, 187–192.

[62] Mason, S. J. and Graham, N. E. (2002). Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. Quarterly Journal of the Royal Meteorological Society, Vol. 128, 2145-2166.

[63] Miller, R. G. (1981). Simultaneous Statistical Inference (2nd edition). Wiley, New York.

[64] Muller, P., Parmigiani, G., and Rice, K. (2006). FDR and Bayesian multiple comparisons rules. Johns Hopkins University, Dept. of Biostatistics Working Papers. Working Paper 115. http://www.bepress.com/jhubiostat/paper115

[65] NCAR Research Application Program (2008). Verification: Forecast verification utilities. R package version 1.29.

[66] Oldstone, M. B. (1998). Molecular mimicry and immune-mediated diseases. The FASEB Journal, Vol. 12, 1255-1265.

[67] Originally written for S-Plus by: Kjell Konis and Marco Riani Ported to R by Luca Scrucca <luca@stat.unipg.it> (2009). forward: Forward search. R package version 1.0.2.

[68] Patnaik, P.B. (1949). The non-central χ^2- and F-distribution and their applications. Biometrika Trust, Vol. 36, No. 1/2, 202-232.

[69] Pepe, M. S. (2003). The statistical evaluation of medical tests for classfication and prediction. Oxford University Press, New York.

[70] Prescott, P. (1975). An approximate test for outliers in linear models. Technometrics, Vol. 17, No. 1, 129-132.

[71] R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

[72] Rawlings, J. O., Pantula, S. G. and Dickey, D. A. (1998). Applied regression analysis: a research tool. 2nd edition. Springer-Verlag, New York.

[73] Ravishanker, N. and Dey, D. K. (2002). A first course in linear model theory. Chapman & Hall/CRC.

[74] Rao, C. R. (1973). Linear statistical inference and its applications. 2nd edition. Wiley, New York.

[75] Robert, C. P. and Casella, G. (2004). Monte Carlo Statistical Methods. Springer, New York.

[76] Romano, J. P. and Shaikh, A. M. (2006). Step-up procedures for control of generalizations of familywise error rate. The Annals of Statistics, Vol. 34, No.4, 1850-1873.

[77] Romano, J. P. and Wolf, M. (2007). Control of generalized error rates in multiple testing. The Annals of Statistics, Vol. 35, No.4, 1378-1408.

[78] Rousseeuw, P. J. (1984). Least median of squares regression. Journal of the American Statistical Association, Vol. 79, No. 388, 871-880.

[79] Ryan, T. A. (1959). Multiple comparisons in psychological research. Psychological Bulletin, Vol. 56, 26-47.

[80] Ryan, T. A. (1960). Significance tests for multiple comparison of proportions, variance and other statistics. Psychological Bulletin, Vol. 57, 318-328.

[81] Sarkar, S. K. (2002). Some results on false discovery rate in stepwise multiple testing procedures. The Annals of Statistics, Vol. 30, No. 1, 239-257.

[82] Sarkar, S. K. (2006). False discovery and false non-discovery rates in single-step multiple testing procedures. The Annals of Statistics, Vol. 34, No.1, 394-415.

[83] Sarkar, S. K. (2008). Generalizing Simes' test and Hochberg's step-up procedures. The Annals of Statistics, Vol. 36, No.1, 337-363.

[84] Sarkar, S. K., and Liu, F. (2008). A note on estimating the false discovery rate. WWW document <http://astro.temple.edu/~sanat/reports/alternate/>. Draft[2](Estimation of FDR_Sarkar and Liu_Submitted).pdf

[85] Sidák, Z. (1968). On multivariate normal probabilities of rectangles: their dependence on correlations. The Annals of Mathematical Statistics, Vol. 39, 1425-1434.

[86] Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. Biometrika, Vol. 73, 751-754.

[87] Shaffer, J. P. (1995). Multiple hypothesis testing: a review. Annual Review of Psychology, Vol. 46, 561-584.

[88] Scott, J. G. and Berger, J. O. (2006). An exploration of aspects of Bayesian multiple testing. Journal of Statistical Planning and Inference, Vol.136, Issue 7, 2144-2162.

[89] Storey, J. D. (2002). A direct approach to false discovery rates. Journal of the Royal Statistical Society, B 64, 479-498. Mathematical Reviews (MathSciNet): MR1924302

[90] Storey, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-values. The Annals of Statistics, Vol.31, No.6, 2013-2035. Mathematical Reviews (MathSciNet): MR2036398

[91] Storey, J. D. (2007). The optimal discovery procedure: a new approach to simultaneous significance testing. Journal of the Royal Statistical Society, Series. B, Statistic Methodology, 69, 1, 1–22. MR2323757

[92] Storey, J. D. and Tibshirani, R. (2001). Estimating false discovery rate under dependence, with applications to DNA microarrays. Technical Report 2001-28, Department of Statistics, Stanford University.

[93] Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. The Proceedings of the National Academy of Sciences, Vol. 100, No. 16, 9440-9445.

[94] Storey, J. D., Taylor, J. E. and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneously conservative consistency of false discovery rates: a unified approach. Journal of the Royal Statistical Society, Series. B, Statistic Methodology, Vol. 66, 187-206. Mathematical Reviews (MathSciNet): MR2035766

[95] Tietjen, G. L., Moore, R. H. and Beckman, R. J. (1973). Testing for a single outlier in simple linear regression. Technometrics, Vol. 15, No. 4, 717-721.

[96] Tiku, M. L. (1966). A note on approximation to the non-central F distribition. Biometrika Trust, Vol. 53, No. 3/4, 606-610.

[97] Tsai, C., Hsueh, H. and Chen, J. J. (2003). Estimation of false discovery rates in multiple testing: application to gene microarray. Biometrics, Vol. 59, 1071-1081.

[98] Tusher, V., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences, Vol. 98, No. 9, 5116-5121.

[99] Wouters, L., Göhlmann, H. W., Bijnens, L, Kass, S. U., Molenberghs, G., Lewi, P. J. (2003). Graphical Exploration of Gene Expression Data: A Comparative Study of Three Multivariate Methods, Biometrics Vol. 59, 1131–1139.

[100] Wu, W. (2008). On false discovery control under dependence. The Annals of Statistics, Vol. 36, No. 1, 364-380.

[101] Yin, Y., Soteros, C. E. and Bickis, M. G. (2009). A Clarifying comparison of methods for cotrolling the false discovery rate. Journal of Statistical Planning and Inference, Vol. 139, Issue 7, 2126-2137

Appendix A

Proof

The proof of Lemma 3.4.2 is given below.

Proof. [of Lemma 3.4.2]

(1).

For a finite h, the ratio $r_h(l)$ becomes less than 1 in a finite number of terms because

$$r_h(l) < 1 \tag{A.1}$$

$$\Leftrightarrow \frac{\frac{\upsilon_1}{2}+\frac{\upsilon_2}{2}+l+h}{\frac{\upsilon_1}{2}+l}\frac{\frac{1}{2}\zeta\upsilon_1 x}{\upsilon_2+\upsilon_1 x}\frac{1}{l+1} < 1 \tag{A.2}$$

$$\Leftrightarrow l^2+(\frac{\upsilon_1}{2}+1-\frac{\frac{1}{2}\zeta\upsilon_1 x}{\upsilon_2+\upsilon_1 x})l+\frac{\upsilon_1}{2}-\frac{\frac{1}{2}\zeta\upsilon_1 x}{\upsilon_2+\upsilon_1 x}(\frac{\upsilon_1}{2}+\frac{\upsilon_2}{2}+h)>0, \tag{A.3}$$

and setting the left hand side of (A.3) equal to 0 gives the equation for the roots of a convex parabola. Hence there exists some l satisfying the inequality (A.1).

(2).

For a finite l, the ratio $t_l(h)$ becomes less than 1 in a finite number of terms because

$$t_l(h) < 1 \tag{A.4}$$

$$\Leftrightarrow \frac{\frac{\upsilon_1}{2}+\frac{\upsilon_2}{2}+l+h}{\frac{\upsilon_2}{2}+h}\frac{1}{h+1}\frac{\frac{1}{2}\eta\upsilon_2}{\upsilon_2+\upsilon_1 x} < 1 \tag{A.5}$$

$$\Leftrightarrow h^2+(\frac{\upsilon_2}{2}+1-\frac{\frac{1}{2}\eta\upsilon_2}{\upsilon_2+\upsilon_1 x})h+\frac{\upsilon_2}{2}-\frac{\frac{1}{2}\eta\upsilon_2}{\upsilon_2+\upsilon_1 x}(\frac{\upsilon_1}{2}+\frac{\upsilon_2}{2}+l)>0, \tag{A.6}$$

then there exists some h satisfying the inequality (A.6), and hence satisfying (A.4). For any finite l, if $t_l(h) \leq 1$, then $W_{l,h}$ decreases in h; if $t_l(h) > 1$, then $W_{l,h}$ increases in h until a maximum and then starts decreasing in l once $r_h(l)$ becomes less than one.

(3).

I show next that there exist h and l such that both $r_h(l) < 1$ and $t_l(h) < 1$, which is equivalent to solving

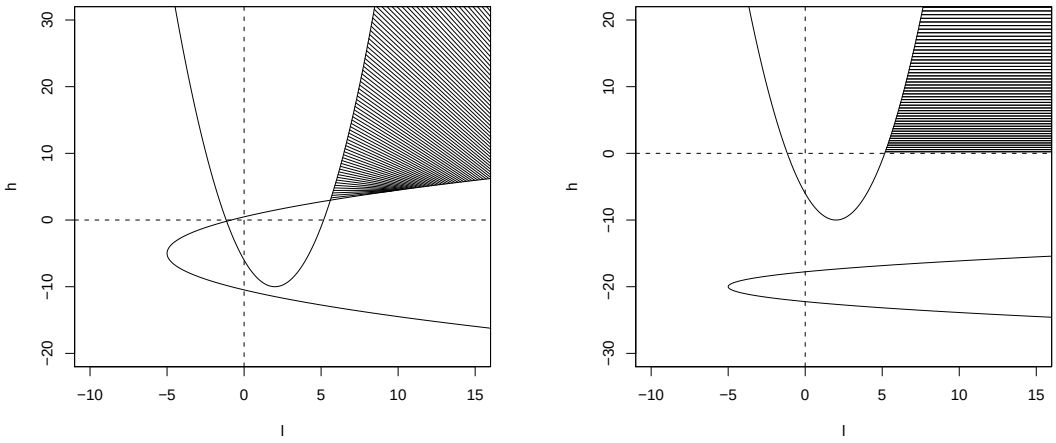
$$\begin{cases} l^2+(\frac{\upsilon_1}{2}+1-\frac{\frac{1}{2}\zeta\upsilon_1 x}{\upsilon_2+\upsilon_1 x})l+\frac{\upsilon_1}{2}-\frac{\frac{1}{2}\zeta\upsilon_1 x}{\upsilon_2+\upsilon_1 x}(\frac{\upsilon_1}{2}+\frac{\upsilon_2}{2}+h)>0 \\ h^2+(\frac{\upsilon_2}{2}+1-\frac{\frac{1}{2}\eta\upsilon_2}{\upsilon_2+\upsilon_1 x})h+\frac{\upsilon_2}{2}-\frac{\frac{1}{2}\eta\upsilon_2}{\upsilon_2+\upsilon_1 x}(\frac{\upsilon_1}{2}+\frac{\upsilon_2}{2}+l)>0 \end{cases} \tag{A.7}$$

$$\Leftrightarrow \begin{cases} \left[l+\frac{\upsilon_1}{4}+\frac{1}{2}-\frac{\frac{1}{4}\zeta\upsilon_1 x}{\upsilon_2+\upsilon_1 x}\right]^2 > \frac{\frac{1}{2}\zeta\upsilon_1 x}{\upsilon_2+\upsilon_1 x}(h+\frac{\upsilon_1}{2}+\frac{\upsilon_2}{2})-\frac{\upsilon_1}{2}+\left(\frac{\upsilon_1}{4}+\frac{1}{2}-\frac{\frac{1}{4}\zeta\upsilon_1 x}{\upsilon_2+\upsilon_1 x}\right)^2 \\ \left[h+\frac{\upsilon_2}{4}+\frac{1}{2}-\frac{\frac{1}{4}\eta\upsilon_2}{\upsilon_2+\upsilon_1 x}\right]^2 > \frac{\frac{1}{2}\eta\upsilon_2}{\upsilon_2+\upsilon_1 x}(l+\frac{\upsilon_1}{2}+\frac{\upsilon_2}{2})-\frac{\upsilon_2}{2}+\left(\frac{\upsilon 2}{4}+\frac{1}{2}-\frac{\frac{1}{4}\eta\upsilon_2}{\upsilon_2+\upsilon_1 x}\right)^2 \end{cases} \tag{A.8}$$

$$\Leftrightarrow \begin{cases} \left[l+\frac{\upsilon_1}{4}+\frac{1}{2}-\frac{\frac{1}{4}\zeta\upsilon_1 x}{\upsilon_2+\upsilon_1 x}\right]^2 > \frac{\frac{1}{2}\zeta\upsilon_1 x}{\upsilon_2+\upsilon_1 x}\left[h+\frac{\upsilon_1+\upsilon_2}{2}+\frac{\upsilon_2+\upsilon_1 x}{\frac{1}{2}\zeta\upsilon_1 x}\left(\frac{\upsilon_1}{4}+\frac{1}{2}-\frac{\frac{1}{4}\zeta\upsilon_1 x}{\upsilon_2+\upsilon_1 x}\right)^2\right] \\ \left[h+\frac{\upsilon_2}{4}+\frac{1}{2}-\frac{\frac{1}{4}\eta\upsilon_2}{\upsilon_2+\upsilon_1 x}\right]^2 > \frac{\frac{1}{2}\eta\upsilon_2}{\upsilon_2+\upsilon_1 x}\left[h+\frac{\upsilon_1+\upsilon_2}{2}+\frac{\upsilon_2+\upsilon_1 x}{\frac{1}{2}\eta\upsilon_2}\left(\frac{\upsilon 2}{4}+\frac{1}{2}-\frac{\frac{1}{4}\eta\upsilon_2}{\upsilon_2+\upsilon_1 x}\right)^2\right] \end{cases}. \tag{A.9}$$

If both inequalities in (A.9) are replaced by equalities, then two equalities denote two parabolas, of which one opens upwards and the other opens to the right. The y-coordinate

of the vertex of the first parabola is negative and the sign of its x-coordinate is indefinite, while the x-coordinate of the vertex of the second parabola is negative and the sign of its y-coordinate is indefinite. Then there are intersections between the two areas denoted by the two inequalities (A.9), no matter whether there is an intersection or not of the two parabolas. Figure A.1 gives two examples of the two cases, in which the shaded areas denotes the solutions to (A.9). The case that there are two intersections of the two parabolas is similar to that in Figure A.1 (a).



(a) Two parabolas have an intersection. (b) Two parabolas have no intersection.

Figure A.1: Two examples for inequality system A.9, of which the solution sets are indicated by shaded area.

■

Appendix B

Tables

In Section 4.3.1, the Bayesian procedure proposed in Chapter 3 was applied to the dataset given in Kanduc *et al.* [56]. Both Patnaik's approximation and Algorithm 3.4.2 were used to calculate the density of doubly noncentral F distribution for this dataset. The tables given below present the difference between the results obtained by using two methods.

$V = 36$			
Beta prior	max difference of densities	max difference of posteriors	max error bound
Beta$(11, 1)$	0.0035	0.00080	5.16×10^{-8}
Beta$(8, 2)$	0.0035	0.00031	5.59×10^{-8}
Beta$(0.8, 0.2)$	0.0035	0.00048	7.86×10^{-8}
Beta$(80, 20)$	0.0035	0.00022	7.08×10^{-8}
Beta$(9.41, 4.03)$	0.0035	0.00027	6.13×10^{-8}
Beta$(8.09, 8.09)$	0.0035	2.78×10^{-5}	$\mathbf{8.63\times10^{-8}}$

Table B.1: Comparison between the results calculated by Patnaik's approximation and by Algorithm 3.4.2 for $V = 36$ and various Beta priors. The data under analyzing are shown in Table 4.2, columns 1 and 4. The error bound E in Algorithm 3.4.2 is chosen to be 10^{-10}. The column named "max difference of densities" present the maximum difference between the doubly noncentral F densities calculated by Patnaik's approximation and by Algorithm 3.4.2; the column named "max difference of posteriors " show the maximum difference between the posterior probabilities computed by Patnaik's approximation and those by Algorithm 3.4.2; the last column present the maximum computation error on a computer with Intel Pentium D Dual processor of 2.8GHz and 2.79GHz, 2GB of Ram when using Algorithm 3.4.2.

$V = 16$			
Beta prior	max difference of densities	max difference of posteriors	max error bound
$\text{Beta}(11, 1)$	0.0035	0.00088	2.51×10^{-8}
$\text{Beta}(8, 2)$	0.0035	0.00045	2.79×10^{-8}
$\text{Beta}(0.8, 0.2)$	0.0035	0.00047	3.99×10^{-8}
$\text{Beta}(80, 20)$	0.0035	0.00056	3.32×10^{-8}
$\text{Beta}(9.41, 4.03)$	0.0035	0.00033	3.01×10^{-8}
$\text{Beta}(8.09, 8.09)$	0.0035	0.00016	4.04×10^{-8}

Table B.2: Comparison between the results calculated by Patnaik's approximation and by Algorithm 3.4.2 for $V = 16$ and various Beta priors. The data under analyzing are shown in Table 4.2, columns 1 and 4. The error bound E in Algorithm 3.4.2 is chosen to be 10^{-10}. The column named "max difference of densities" present the maximum difference between the doubly noncentral F densities calculated by Patnaik's approximation and by Algorithm 3.4.2; the column named "max difference of posteriors " show the maximum difference between the posterior probabilities computed by Patnaik's approximation and those by Algorithm 3.4.2; the last column present the maximum computation error on a computer with Intel Pentium D Dual processor of 2.8GHz and 2.79GHz, 2GB of Ram when using Algorithm 3.4.2.

$V = 9$			
Beta prior	max difference of densities	max difference of posteriors	max error bound
$\text{Beta}(11, 1)$	0.0035	0.00085	1.52×10^{-8}
$\text{Beta}(8, 2)$	0.0035	0.00063	1.76×10^{-8}
$\text{Beta}(0.8, 0.2)$	0.0035	0.00029	2.47×10^{-8}
$\text{Beta}(80, 20)$	0.0035	0.00079	1.97×10^{-8}
$\text{Beta}(9.41, 4.03)$	0.0035	0.00050	1.91×10^{-8}
$\text{Beta}(8.09, 8.09)$	0.0035	0.00041	2.41×10^{-8}

Table B.3: Comparison between the results calculated by Patnaik's approximation and by Algorithm 3.4.2 for $V = 9$ and various Beta priors. The data under analyzing are shown in Table 4.2, columns 1 and 4. The error bound E in Algorithm 3.4.2 is chosen to be 10^{-10}. The column named "max difference of densities" present the maximum difference between the doubly noncentral F densities calculated by Patnaik's approximation and by Algorithm 3.4.2; the column named "max difference of posteriors " show the maximum difference between the posterior probabilities computed by Patnaik's approximation and those by Algorithm 3.4.2; the last column present the maximum computation error on a computer with Intel Pentium D Dual processor of 2.8GHz and 2.79GHz, 2GB of Ram when using Algorithm 3.4.2.

$V = 4$			
Beta prior	max difference of densities	max difference of posteriors	max error bound
Beta$(11, 1)$	0.0035	0.00075	8.2×10^{-9}
Beta$(8, 2)$	0.0035	0.00093	9.5×10^{-9}
Beta$(0.8, 0.2)$	0.0035	8.44×10^{-5}	1.28×10^{-8}
Beta$(80, 20)$	0.0035	**0.0011**	1×10^{-8}
Beta$(9.41, 4.03)$	0.0035	0.00091	1.07×10^{-8}
Beta$(8.09, 8.09)$	0.0035	0.00087	1.18×10^{-8}

Table B.4: Comparison between the results calculated by Patnaik's approximation and by Algorithm 3.4.2 for $V = 4$ and various Beta priors. The data under analyzing are shown in Table 4.2, columns 1 and 4. The error bound E in Algorithm 3.4.2 is chosen to be 10^{-10}. The column named "max difference of densities" present the maximum difference between the doubly noncentral F densities calculated by Patnaik's approximation and by Algorithm 3.4.2; the column named "max difference of posteriors " show the maximum difference between the posterior probabilities computed by Patnaik's approximation and those by Algorithm 3.4.2; the last column present the maximum computation error on a computer with Intel Pentium D Dual processor of 2.8GHz and 2.79GHz, 2GB of Ram when using Algorithm 3.4.2.

Appendix C

Code

Appendix B include the program code written in "R" for the two main algorithms of this thesis.

C.1 Code for Algorithm 3.3.1

The code given below with notes starting by "#".

```
# The main function to compute the marginal posterior probability,
# P(Hi=1|r*^2) is "posterior".
# This is the function to generate binary array for input uniform random variables x
# with probability=prob.
equalbinary=function(x,prob)
{
    if (x<=prob)
    {
        index=0
    }
    else
    {
        index=1
    }
    return(index)
}
# approximate the density of the double noncentral F distribution,
# by Patnaik's approximation.
# x = the quantile.
# dgf1 = the numerator degrees of freedom.
# dgf2 = the denominator degrees of freedom.
# ncp1 = the numerator noncentrality parameter.
# ncp2 = the denominator noncentrality parameter.
approxdoublef=function(x,dgf1,dgf2,ncp1,ncp2)
{
    if(ncp2==0)
    {
        density=df(x,dgf1,dgf2,ncp1)
    }
    else
    {
        if(ncp1==0)
        {
            density=1/(x^2)*df(1/x,dgf1,dgf2,ncp2)
        }
        else
```

```
        {
            scaler=1/(1+ncp2/dgf2)
            newdf2=(dgf2+ncp2)^2/(dgf2+2*ncp2)
            density=(1/scaler)*df(x/scaler, dgf1, newdf2, ncp1, log = FALSE)
        }
    }
    return(density)
}
# This is the function computing the marginal posterior probability P(Hi=1|r*^2).
# x1 is the input vector of values of the explanatory variable.
# sig is the standard deviation of the prior distribution of mu,
#which is deviation of the mean of outliers.
# a,b are the parameters of beta prior for pi0.
# n is the number of generated Bayes samples.
posterior<-function(x1,y,sig,a,b,n)
{
    m=length(x1)
    p=2
    x0=rep(1,m)
    X=cbind(x0,x1)
    lr_model=lm(y~x1)
    deletion_r=rstudent(lr_model)
    deletion_f=deletion_r^2
    #calculate hat matrix.
    invX=solve(t(X)%*%X)
    hatmatrix=X%*%invX%*%t(X)
    # use R function hatvalues() to calculate the diagonal elements of the hatmatrix.
    diaghat=hatvalues(lr_model)
    # generate n random probabilities of a point not being an outlier from,
    # Beta distribution Beta (a,b).
    rpi0=rbeta(n,a,b)
    # generate n vectors of indices of alternative.
    rH=matrix(0,nrow=n,ncol=m-1)
    rmu=matrix(0,nrow=n,ncol=m-1)
    for (i in 1:n)
    {
        uniform2=runif(m-1)
        rH[i,]=sapply(uniform2,equalbinary,prob=rpi0[i])
        rmu[i,]=rH[i,]
        # generate k=#{the elements of the i-th row of rH = 1} mu,
        # from normal(0,sd=sig),
        # where {the elements of rH = 1} represents alternatives.
        nonzeromu=rnorm(sum(rH[i,]),sd=sig)
        # replace all 1's in the i-th row of rH by generated mu's.
        rmu[i,which(rmu[i,]==1)]=nonzeromu
    }
    # the noncentral parameter of the numerator under the null.
    nlambdasq=matrix(0,0,nrow=n,ncol=m)
```

```
    # the noncentral parameter of the numerator under the alternative.
    nlambdasq_a=matrix(0,0,nrow=n,ncol=m)
    # the noncentral parameter of the denominator.
    ndelta=matrix(0,0,nrow=n,ncol=m)
    # function h1 in (3.60).
    nullden=matrix(0,0,nrow=n,ncol=m)
    # function h2 in (3.60).
    alterden=matrix(0,0,nrow=n,ncol=m)
    p_i=numeric(m)
    I=diag(1,(m-1),(m-1))
    for (i in 1:m)
    {
        x_i=X[-i,]
        nume=0
        denom=0
        # calculate inverse of (t(x_i)%*%x_i).
        invX_i=invX+invX%*%X[i,]%*%t(X[i,])%*%invX/(1-diaghat[i])
        # calculate the hat matrix after deleting the ith observation.
        P=x_i%*%invX_i%*%t(x_i)
        IP=I-P
        for (k in 1:n)
        {
            mui=rnorm(1,sd=sig)
            # use the formula 3.41 given in Chapter 3 of my thesis to calculate
            # the numerator noncentral parameter of the doubly noncentral
            # F distribution under the null hypothesis
            indexH=which(rH[k,]==1)
            newhat=hatmatrix[,-i]
            offdiaghat=newhat[i,indexH]
            lambda=-sum(offdiaghat*rmu[k,indexH])/sqrt(1-diaghat[i])
            # use the formula 3.40 to calculate the denominator,
            # noncentral parameter of the doubly noncentral F distribution.
            delta=t(rmu[k,])%*%IP%*%rmu[k,]
            lambdasq=lambda^2
            nlambdasq[k,i]=lambdasq
            ndelta[k,i]=delta
            # use the formula 3.42 to calculate the numerator noncentral
            # parameter of the doubly noncentral F distribution under the althernative
            nlambdasq_a[k,i]=(lamda+mui*sqrt(1-diaghat[i]))^2
            # use the function "approxdoublef" to calculate the density of,
            # the doubly noncentral F distribution.
            nullden[k,i]=approxdoublef(deletion_f[i],1,m-p-1,lamdasq,delta)
            alterden[k,i]=approxdoublef(deletion_f[i],1,m-p-1,nlamdasq_a[k,i],delta)
            denom=denom+nullden[k,i]*rpi0[k]
            nume=nume+alterden[k,i]*(1-rpi0[k])
        }
    p_i[i]=1/(1+nume/denom)
}
```

```
p1_i=1-p_i
return(list(p1=p1_i,nlamdasq,nlamdasq_a,ndelta,nullden,alterden,rpi0,rH))
}
```

C.2 Code for Algorithm 3.4.2

The code given below with notes starting by "#".

```
# The main function to compute the density of doubly noncentral F,
# distribution is "doublef".
# This is the function computing the terms of the infinite sums,
# of the density of double noncentral F.
# x = the quantile.
# df1 = the numerator degrees of freedom.
# df2 = the denominator degrees of freedom.
# ncp1 = the numerator noncentrality parameter.
# ncp2 = the denominator noncentrality parameter.
# l = the index associated with ncp1.
# h = the index associated with ncp2.
termdoublef<-function(x,df1,df2,ncp1,ncp2,l,h)
{
    term=dpois(l, lambda=ncp1/2)*dpois(h, lambda=ncp2/2)*(df1*x/(df2+df1*x))^l*
    (df2/(df2+df1*x))^h*beta(df1/2,df2/2)/beta((df1/2+l),(df2/2+h))
    return(term)
}
# This is the function computing the ratio of two adjacent terms of the infinite sums,
# (S(l,h+1)/S(l,h)) when l is fixed.
termratioh<-function(x,df1,df2,ncp1,ncp2,l,h)
{
    termra=ncp2*df2/(2*(df2+df1*x)*(h+1))*(1+(df1/2+l)/(df2/2+h))
    return(termra)
}
# This isthe function computing the ratio of two adjacent terms of the infinite sums,
# (S(l+1,h)/S(l,h)) when h is fixed.
termratiol<-function(x,df1,df2,ncp1,ncp2,l,h)
{
    termra=ncp1*df1*x/(2*(df2+df1*x)*(l+1))*(1+(df2/2+h)/(df1/2+l))
    return(termra)
}
# This is the function computing theratio of two adjacent terms of the infinite sums,
# (S(l+1,k+1)/S(l,k)).
termratio<-function(x,df1,df2,ncp1,ncp2,l,h)
{
    termra=ncp1*ncp2*df1*df2*x/(4*(df2+df1*x)^2*(h+1)*(l+1))*
    (1+(df1/2+l)/(df2/2+h))*(2+(df2/2+h)/(df1/2+l))
    return(termra)
}
# This is the function to computing the density of the double noncentral F distribution,
```

```
# which equals infinite sum of index l and h.
# when the noncentral parameter in the numeritor=0,
# the function df in R can calculate the density of noncentral F.
# when the noncentral parameter in the denominator=0,
# density of the density of noncentral F=1/f^2*df(1/f).
# f = the quantile.
# dgf1 = the numerator degrees of freedom.
# dgf2 = the denominator degrees of freedom.
# ncpa1 = the numerator noncentrality parameter.
# ncpa2 = the denominator noncentrality parameter.
doublef<-function(f,dgf1,dgf2,ncpa1,ncpa2,errorbound)
{
    if(ncpa2==0)
    {
        dens=df(f,dgf1,dgf2,ncpa1)
        return(list(density=dens,sigterm=0,bound=0))
    }
    else
        {
            if(ncpa1==0)
            {
                dens=1/f^2*df(1/f,dgf1,dgf2,ncpa2)
                return(list(density=dens,sigterm=0,bound=0))
            }
            else
            {
                # start index l, the first sum in the density,
                # of double noncentral F, from 0.
                ind1=0
                ind2=0
                term=termdoublef(x=f,df1=dgf1,df2=dgf2,
                        ncp1=ncpa1,ncp2=ncpa2,l=ind1,h=ind2)
                termsum=term
                ratioh=termratioh(x=f,df1=dgf1,df2=dgf2,
                        ncp1=ncpa1,ncp2=ncpa2,l=ind1,h=ind2)
                allind2=numeric()
                rowerror=numeric()
                while (ratioh>1)
                {
                    ind2=ind2+1
                    term=termdoublef(x=f,df1=dgf1,df2=dgf2,
                            ncp1=ncpa1,ncp2=ncpa2,l=ind1,h=ind2)
                    ratioh=termratioh(x=f,df1=dgf1,df2=dgf2,
                            ncp1=ncpa1,ncp2=ncpa2,l=ind1,h=ind2)
                    termsum=termsum+term
                }
                modh=ind2
                termmodh=term
```

```
ratiohmodh=ratioh
while (term/(1-ratioh)>errorbound)
{
    ind2=ind2+1
    term=termdoublef(x=f,df1=dgf1,df2=dgf2,ncp1=ncpa1,
            ncp2=ncpa2,l=ind1,h=ind2)
    ratioh=termratioh(x=f,df1=dgf1,df2=dgf2,ncp1=ncpa1,
            ncp2=ncpa2,l=ind1,h=ind2)
    termsum=termsum+term
}
rowerror[1]=term/(1-ratioh)
allind2[1]=modh
term=termmodh
ratioh=ratiohmodh
ratiol=termratiol(x=f,df1=dgf1,df2=dgf2,ncp1=ncpa1,
        ncp2=ncpa2,l=ind1,h=modh)
ratiol_2=termratiol(x=f,df1=dgf1,df2=dgf2,ncp1=ncpa1,
        ncp2=ncpa2,l=ind1,h=modh+1)
ratio=ratioh*ratiol_2
s=1
while (ratiol_2>1|ratio>1|(term/(1-ratio)*((1-ratioh)^(-1)+
        (1-ratiol)^(-1)))>errorbound)
{
    ind1=ind1+1
    s=s+1
    ind2=0
    term=termdoublef(x=f,df1=dgf1,df2=dgf2,ncp1=ncpa1,
            ncp2=ncpa2,l=ind1,h=ind2)
    ratioh=termratioh(x=f,df1=dgf1,df2=dgf2,ncp1=ncpa1,
            ncp2=ncpa2,l=ind1,h=ind2)
    termsum=termsum+term
    while(ratioh>1)
    {
        ind2=ind2+1
        term=termdoublef(x=f,df1=dgf1,df2=dgf2,ncp1=ncpa1,
                ncp2=ncpa2,l=ind1,h=ind2)
        ratioh=termratioh(x=f,df1=dgf1,df2=dgf2,ncp1=ncpa1,
                ncp2=ncpa2,l=ind1,h=ind2)
        termsum=termsum+term
    }
    modh=ind2
    termmodh=term
    ratiohmodh=ratioh
    while (term/(1-ratioh)>errorbound)
    {
        ind2=ind2+1
        term=termdoublef(x=f,df1=dgf1,df2=dgf2,ncp1=ncpa1,
                ncp2=ncpa2,l=ind1,h=ind2)
```

```
                    ratioh=termratioh(x=f,df1=dgf1,df2=dgf2,ncp1=ncpa1,
                            ncp2=ncpa2,l=ind1,h=ind2)
                termsum=termsum+term
                }
                allind2[s]=modh
                rowerror[s]=term/(1-ratioh)
                term=termmodh
                ratioh=ratiohmodh
                ratiol=termratiol(x=f,df1=dgf1,df2=dgf2,ncp1=ncpa1,
                        ncp2=ncpa2,l=ind1,h=modh)
                ratiol_2=termratiol(x=f,df1=dgf1,df2=dgf2,ncp1=ncpa1,
                            ncp2=ncpa2,l=ind1,h=modh+1)
                ratio=ratioh*ratiol_2
            }
            lasterror=term/(1-ratio)*((1-ratioh)^(-1)+(1-ratiol)^(-1))
            lastind1=ind1
            dens=df(f,dgf1,dgf2)*termsum
            return(list(density=dens,sigterm=lastind1+allind2[s]+1,
                    bound=sum(rowerror)+sum(colerror)+lasterror))
        }
    }
}
```

Index