ROBUST RANKING AND SELECTION WITH HEAVY-TAIL PRIORS AND ITS APPLICATIONS IN MARKET BASKET ANALYSIS

by

Hao He

Submitted in partial fulfillment of the requirements
for the degree of Master of Science

at

Dalhousie University
Halifax, Nova Scotia
April 2016

Table of Contents

List of Tables

List of Figures

Abstract

Ranking and selection is the problem of identifying the best units according to some parameters based on estimates for the parameters obtained from data. Various methods of ranking and selection have been developed, such as empirical Bayes methods ranking units based on a multi-stage Bayesian hierarchical model. Compared with the non-Bayesian methods, including local maximum likelihood and testing, the Bayesian methods have a number of advantages. However, Bayesian methods have the difficultly of choosing the prior. A common choice is to use the conjugate prior for mathematical convenience. We show that while this is often acceptable for many Bayesian analysis, it can have serious problems for ranking.

We perform a simulation study to determine the effect of choice of prior on ranking methods. We find that a heavy-tailed prior is more robust to misspecification in many ranking problems, especially when we are focused on the top ranked units. We give an example of applying the posterior mean ranking method with t-prior and normal prior and some other ranking methods in a simulated market basket data, which provide more comparison between different ranking methods. The results of the simulation study can be applied to a range of empirical Bayesian analysis.

List of Abbreviations and Symbols Used

Symbols and Abbr.	Description
ANOVA	Analysis of variance
MLE	Maximum likelihood estimator
PM	Posterior mean
PMR	Posterior mean ranking
PER	Posterior expect ranking
MBA	Market basket analysis
MSE	Mean squared error
SNP	Single-nucleotide polymorphisms
T2D	type II diabetes
i.i.d	Indenpendent and identical distributed
X	Random variable X
E[X]	Expectation of random variable X
Var(X)	Variance of random variable X
d.f	degrees of freedom

Acknowledgements

I would like to express my gratitude to my supervisor, Prof. Hong Gu, and co-supervisor, Prof. Toby Kenney whose expertise, understanding, and patience, added considerably to my graduate experience. I appreciate the assistance they provided at all levels of the research project.

I would also like to thank the other members of my committee, Prof. Chris Field, and Prof. Edward Susko for spending time reading my thesis and I am grateful their to very valuable comments on this thesis.

Finally, I would like to thank my parents for the support they have provided me through my entire life and in particular, for providing me with unfailing support and continuous encouragement throughout my years of study.

Chapter 1

Introduction

1.1 Ranking and Selection

Ranking and selection are statistical inference problems, which aim to identify the most important units from a list of candidates. Various approaches to ranking and selection have been developed over the past decades, since the problem was first introduced by Bechhofer [1] and Gupta [2]. Later work given by Bechhofer, Kiefer and Sobel [3], Gibbons, Chakraborti [4] provided a good overview of the field.

In order to simplify our discussion, a general ranking problem is described by Gibbons, Olkin and Sobel [5] as follows: Suppose there are n populations in our list, which are indexed by parameters θ_i's

$$f(x_1;\theta_1),\ f(x_2;\theta_2),\ \cdots,\ f(x_n;\theta_n) \tag{1.1}$$

Each θ_i represents the unknown parameter for the interest of ith population, and x_i is the real-valued measurement/observation sampled from the corresponding population with variance σ_i^2. Basically, the purpose is to rank these populations based on the value of θ_i's in a well-defined way (say from largest to smallest). Therefore, the true ranks of units in the list $\theta = (\theta_1, \theta_2, \cdots, \theta_n)'$ are given by

$$R_k(\theta) = \text{rank}(\theta_k) = \sum_{j=1}^{n} \text{I}(\theta_k \leq \theta_j), \tag{1.2}$$

where $\text{I}(\cdot)$ is the indicator function.We assume the populations follow normal distribution, that is $X_i \sim \mathcal{N}(\theta_i, \sigma_i^2)$ in the following discussion. Adapting the methods to other distributions is mostly straightforward.

The way we treat θ_i is critical to our methods of ranking and selection. Aitkin and Longford [6] provide two approaches in ranking the effectiveness among schools, one approach treating θ_i's as fixed and another treating them as random.

If treating the θ_i's as fixed parameters, then a natural way of proceed is using the maximum likelihood (ML) method to obtain an estimate ($\hat{\theta}_i^{MLE} = x_i$) of θ_i. The

rank of θ_i is based on the rank of $\hat{\theta}_i^{MLE}$. Another frequently used method is the testing approach, which is an application of the classical analysis of variance (ANOVA). The testing approach simply tests the evidence against some designated null hypothesis, usually $H_0 : \theta_i = 0$. The ranks from the testing approach are the ranks of p-values, smallest to largest, for the test of $H_0 : \theta_i = 0$ against $H_A : \theta_i > 0$. The first method is very useful when all of the variances σ_i^2 are relatively close (balanced case) or extremely small. Since the MLE ranking ignores the effect of sampling fluctuations, X_i's associated with fairly high standard error are over-represented among the top elements under MLE ranking. Conversely, testing approach based ranking over-represents those X_i's associated with fairly low standard error among the top ranked elements [7].

The other approach: treating θ_i as random effect, assumes θ_i independent identically distributed (i.i.d) from some prior distribution. A simple two-stage hierarchical model can be established in conjunction with Bayes or empirical Bayesian methods, with the final ranks of the θ_i's based on information from the posterior distribution. The empirical Bayesian ranking methods rely on a certain loss function.

A very straightforward Bayes and empirical Bayesian method is posterior mean ranking (PMR) [6]. This method gives the rank of θ_i as the rank of the posterior mean $\mathrm{E}[\theta_i|X_i, \sigma_i^2]$ in the list of all posterior means. Moreover, Laird and Louis [8] proposed the posterior expected ranking (PER) method, which is one of the most used methods under the view of empirical ranking. This method is based on the posterior mean of the ranks, in which it calculates the posterior distribution of θ_i, then calculates the posterior distribution of the induced ranking and then ranks according to the mean position of each unit in the ranking. Another method introduced by Berger and Deely [9] gives ranks of θ_i's by their posterior probability of θ_i being the largest conditional on data and all θ_i not being homogenous. The usual procedure for this method is: (a) a homogeneity test with null hypothesis $H_0 : \theta_1 = \theta_2 = \cdots = \theta_n$; (b) ranking the posterior probabilities $\Pr(\theta_i \text{ is the largest}|\text{data}, H_0 \text{ false})$.

A series of papers coauthored by Gupta [10, 11] have covered the methodology and application of ranking and selection problems with empirical Bayes. Laird and Louis [8], Berger and Deely [9] provided some seminal contributions to the empirical Bayesian ranking framework, which has been further developed in many ways. Lin

et al. [12] elaborates the loss functions in empirical Bayes methods, by giving a series of loss functions with which the corresponding empirical Bayes methods are able to meet different objectives in ranking (*e.g.* identifying the relatively good or relative poor units). Noma *et al.* [13] provides comparisons in microarray studies among three ranking methods based on their corresponding loss functions.

Shen and Louis [14] compare various estimators of θ_i based on their performance in estimating the parameter histogram, estimating the parameter ranks and estimating unit-specific parameters.

Henderson and Newton [7] propose another empirical Bayes ranking method, which ranks units on a statistic called "r-value" in their paper. The intention of the r-value ranking method is to choose a series of cutoffs that maximize the expected overlap between the selected unit and true top list. The cutoffs in ranking are mathematically defined by a family of threshold functions, $\mathcal{T} = \{t_\alpha : \alpha \in (0,1)\}$, where t_α is a function $t_\alpha(\sigma^2)$, indicating that the ith unit X_i is in top $\alpha\%$ if $X_i \geq t_\alpha(\sigma_i^2)$. Then the threshold function of r-value, $\mathcal{T}^* = \{t_\alpha^* : \alpha \in (0,1)\}$, satisfies that

$$\Pr\left(X_i \geq t_\alpha^*(\sigma_i^2), \theta_i \geq \theta_\alpha\right) \geq \Pr\left(X_i \geq t_\alpha(\sigma_i^2), \theta_i \geq \theta_\alpha\right) \tag{1.3}$$

for any other threshold $\mathcal{T} = \{t_\alpha : \alpha \in (0,1)\}$, where θ_α is the $(1-\alpha)$th quantile of prior, that is $\Pr(\theta_i \geq \theta_\alpha) = \alpha$. Therefore, given the r-value threshold function, we select θ_i before θ_j, if $X_i > t_\alpha^*(\sigma_i^2)$, $t_\alpha^*(\sigma_j^2) > X_j$ for some α.

These empirical Bayesian methods typically provide a compromise between the p-value approach which is based only on the certainty that an effect is non-zero, and not on the size of the effect [15], and the MLE approach which does not properly account for the different uncertainties in the estimates. A number of advantages of empirical Bayes methods are given by Laird and Louis [8], and Berger and Deely [9].

There have been numerous applications of ranking and selection, including identifying the most hazardous road sites by Brijs *et al.* [16]; ranking of institutions, e.g. universities, schools and hospitals by Hall and Miller [17]; ranking of animals or plants by their breeding value by Campos *et al.* [18]; ranking of risk factors, single-nucleotide polymorphisms (SNPs), for type II diabetes (T2D) [7]. The methods of ranking and selection are broadly being used in various domains of large-scale inference. In this thesis, we present an application of ranking methods to market basket analysis.

1.2 Market Basket Analysis

Market basket analysis (MBA) is a common application of a methodology known as association analysis, which is useful for discovering interesting relationships hidden in large data sets [19]. It is used to extract information useful for a variety of business-related applications such as marketing promotions, inventory management, and customer relationship management.

Table 1.1 gives a small example of two representations of a typical MBA dataset. The data consist of a list of sets of items purchased in five transactions. The data in Table 1.1 is expressed in two different ways. The left table is the original format with each row being a transaction, which contains a uniquely labeled TID and a set of items bought by a given customer, while the right table uses binary variables to re-express the transaction data with "0" being not included in the corresponding transaction, "1" being included.

TID	Items
1	{ Bread, Milk }
2	{ Bread, Diapers, Beer, Eggs }
3	{ Milk, Diapers, Beer, Cola }
4	{ Bread, Milk, Diapers, Beer }
5	{ Bread, Milk, Diapers, Cola }

$\Longrightarrow$

	Items					
TID	Bread	Milk	Eggs	Beer	Cola	Diapers
1	1	1	0	0	0	0
2	1	0	1	1	0	1
3	0	1	0	1	1	1
4	1	1	0	1	0	1
5	1	1	0	0	1	1

Table 1.1: An example of market basket transactions

Let $I = \{i_1, i_2, \cdots, i_d\}$ be the set of items in a market basket transaction data and $T = \{t_1, t_2, \cdots, t_N\}$ be the set of all transactions [19]. Our objective is to identify association rules between item sets, that is, disjoint sets of items A and B such that the probability of a transaction including both A and B is significantly different from the product of the probabilities of including A and including B which would be the probability if the purchase of A and B were independent. We use notation support count as the frequency of occurrence of any itemset A,

$$\text{count}(A) = |\{t_i | A \subseteq t_i, t_i \in T\}| \tag{1.4}$$

Then, it is common to describe an association rule as an implication expression of the form X leads to Y ($X \Rightarrow Y$), where X and Y are disjoint itemsets, *i.e.* $X \cap Y = \emptyset$. A number of measures of the strength of an association rule have been proposed in

the literature. For example, two of the most commonly used measures are support and confidence [20]

$$\text{Support}, s(A \Rightarrow B) = \frac{\text{count}(A \cup B)}{N} \tag{1.5}$$

$$\text{Confidence}, c(A \Rightarrow B) = \frac{\text{count}(A \cup B)}{\text{count}(A)} \tag{1.6}$$

Support is an estimation of the probabilities of the occurrence of corresponding association rule which determines how often a rule is applicable to a given data set, while confidence is an estimation of conditional probabilities of occurrence of corresponding association rule given itemset A, which determines how frequently items in B appear in transactions that contain A.

We will also look at the leverage measure introduced by Piatetsky-Shapiro in 1991 [21]. The Leverage measure, or Piatetsky-Shapiro (PS) measure is defined as

$$\text{leverage}(A \Rightarrow B) = s(A \Rightarrow B) - s(A)s(B) \tag{1.7}$$

We will give more detail about leverage and the reason why it is a desirable measure of our study in Section 3.2. The important point to observe about these measures is that they are statistics, so they all have standard errors associated with them. For applying the ranking methods discussed in Section 1.1, we will need to use the estimated standard errors of these statistics.

1.3 Purpose of This Study

In this thesis, I study the effect of priors in empirical Bayes ranking methods, especially using normal/normal model in posterior mean ranking. The conventional choices of priors in empirical Bayesian analysis are conjugate priors and nonparametric priors. In this thesis, I study what factors are important in choice of priors.

In the next chapter, we will focus on the heaviness of the tail of the prior distribution. We will compare posterior mean ranking using a light-tailed normal distribution, a medium-tailed gamma distribution and a heavy-tailed Student's t-distribution as priors, under the same three choices of true priors. This will provide a comparison of the loss arising from overestimating the tail weight and the loss arising from underestimating the tail weight. Additionally, there will be some discussion about the estimations of hyperparameters of priors.

In the third chapter, there is an application to simulated MBA. We compare posterior mean ranking with a suitable choice of prior to a variety of other ranking methods, including MLE approach, testing approach, posterior mean ranking and r-value ranking.

I will make some conclusions of the idea of choosing priors in Chapter 4 with several progress to be made in future work.

Chapter 2

General Choice of Prior for Posterior Mean Ranking

2.1 Introduction

Bayesian approaches are very commonly used and powerful tools in ranking and selection problems. Unlike testing approaches and local maximum likelihood approaches which only consider the information of observations/estimates, these ranking approaches rely on the posterior distribution of θ in equation (1.1) which contains information both from observations and prior information.

In a simplified general ranking problem, using the notation from equation (1.1), there are n independent populations

$$f(x_1;\theta_1,\sigma_1^2),\ f(x_2;\theta_2,\sigma_2^2),\ \cdots,\ f(x_n;\theta_n,\sigma_n^2). \tag{2.1}$$

The intention of this ranking is to give orders to θ_i's from large to small. For our treatment of the θ_i's as random effects, we assume that each θ_i has the same prior distribution $g(\theta)$, and all θ_i's are different. The conditional likelihood of observations or the sampling model is normal $\mathcal{N}(\theta_i,\sigma_i^2)$, that is

$$f(x_1;\theta_i,\sigma_i^2)=p(x_i|\theta_i,\sigma_i^2)=\frac{1}{\sqrt{2\pi}\sigma_i}e^{-\frac{(x_i-\theta_i)^2}{2\sigma_i^2}}, \tag{2.2}$$

where σ_i^2 is the variance of observation. The ranks of the true list of units $\theta=(\theta_1,\theta_2,\cdots,\theta_n)'$ are as defined in Section 1.1, given by equation (1.2), $R_k(\theta)$, for $k=1,2,\cdots,n$.

Although, for a general ranking problem, the primary objective is to give ranks to θ_i close to their true ranks $R_i(\cdot)$. For our purposes, we consider the relative values of the θ_i to also be important - if θ_i and θ_j are close, then ranking them in the wrong orders is a small error; if the difference is larger then ranking them in the wrong order should be considered a larger error. The loss function for posterior mean ranking incorporates this factor in a very natural way. We Therefore focus on the effect of prior distribution on posterior mean ranking.

For an individual variable X_i with parameter θ_i, the posterior mean estimation ($\hat{\theta}_i^{\mathrm{PM}} = \mathrm{E}[\theta_i|X_i, \sigma_i^2]$) minimizes expectation of a squared-error loss, which is

$$L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2 \tag{2.3}$$

where $\hat{\theta}$ is an estimate of θ.

For ranking, posterior mean ranking minimizes the expectation of a loss function given by the best value minus the value chosen. That is if X_i with parameter θ_i is the value we should choose, but instead we choose X_j with parameter θ_j, then the loss function in selecting the top k units is

$$\begin{aligned} L_k(\hat{R}) &= \sum_{j=1}^{n} \mathrm{I}(\hat{R}_j \leq k) \sum_{i=1}^{n} \mathrm{I}(\hat{R}_j = R_i)(\theta_i - \theta_j) \\ &= \sum_{i=1}^{n} \mathrm{I}(R_i \leq k)\theta_i - \sum_{j=1}^{n} \mathrm{I}(\hat{R}_j \leq k)\theta_j \end{aligned} \tag{2.4}$$

where $\hat{R} = (\hat{R}_1, \cdots, \hat{R}_n)$ is the ranks of all $\hat{\theta}_i$s by the ranking method and R_i is the ranking of θ_i. That is $\mathrm{E}[L(\cdot)]$ is minimized by posterior mean ranking $\hat{R}^{\mathrm{PMR}}$ defined as

$$\begin{aligned} \hat{R}_k^{\mathrm{PMR}} = \mathrm{rank}_{PM}(\theta_k) &= \sum_{j=1}^{n} \mathrm{I}(\hat{\theta}_k \leq \hat{\theta}_j) \\ &= \sum_{j=1}^{n} \mathrm{I}(\mathrm{E}[\theta_k|X_k, \sigma_k^2] \leq \mathrm{E}[\theta_j|X_j, \sigma_j^2]) \end{aligned} \tag{2.5}$$

The general form of posterior mean estimate is given by

$$\mathrm{E}[\theta_i|X_i] = \frac{\int_{\theta_i} \theta_i \; g(\theta_i) \; p(X_i|\theta_i, \sigma_i^2) \; d\theta_i}{\int_{\theta_i} g(\theta_i) \; p(X_i|\theta_i, \sigma_i^2) \; d\theta_i} \tag{2.6}$$

There are some other choices for loss functions, which will give different posterior distribution ranking. Mostly they can be converted to the posterior mean by transforming the parameter space.

The most commonly chosen prior for posterior mean ranking is the conjugate prior, which is a normal prior here, for its simplicity and generality. However, in some cases, such a choice of prior will produce a poor result of ranking, since it can favor small variance units. Especially, when we rank the top α units (θ_i's), we are more focusing on units with high observations/estimates. For example, when the

two units θ_i and θ_j have very close variances where σ_i^2 is a bit larger than σ_j^2, but very different estimates with $\hat{\theta}_i$ much larger than $\hat{\theta}_j$, the posterior mean method with normal/normal model can rank the θ_j above θ_i. Examples of this problem will be illustrated in Section 3.4.1. Thus normal priors seem to suffer from some deficiencies that can not satisfy our purpose. How to choose a general choice of prior is a critical task. The following sections will give some comparison of different types of prior: heavy-tail, medium-tail and light-tail. By comparing the result of posterior mean ranking with different priors, we are able to choose prior that is more robust to our ranking procedure.

2.2 Simulations from Different Priors

In order to achieve our goal, we will firstly draw samples from three sets of priors with different types of tail, normal distribution with light-tail, gamma distribution with medium-tail and Student's t-distribution with heavy-tail, so that we can compare the effect of using different priors in cases where the tail of the prior used is too light or too heavy.

We generate three sets of θ from different prior:

- $\theta_1^{(N)}, \theta_2^{(N)}, \cdots, \theta_n^{(N)}$ i.i.d normal distribution, $\mathcal{N}(\mu, \tau^2)$ with $\mu = 1$ and $\tau^2 = 0.1$;
- $\theta_1^{(G)}, \theta_2^{(G)}, \cdots, \theta_n^{(G)}$ i.i.d gamma distribution, $Gamma(\alpha, \beta)$ with $\alpha = 10$ and $\beta = 10$;
- $\theta_1^{(T)}, \theta_2^{(T)}, \cdots, \theta_n^{(T)}$ i.i.d Student's t-distribution, $t_\nu(\eta, \lambda)$ with d.f $\nu = 3$, location parameter (mean) $\eta = 1$ and inverse scaling parameter $\lambda = 30$.

Those settings mean that these three prior distributions have the same first two moments, *i.e.* the same mean $\mathrm{E}[\theta] = \mu = \frac{\alpha}{\beta} = \eta = 1$ and the same variance $\mathrm{Var}(\theta) = \tau^2 = \frac{\alpha}{\beta^2} = \frac{1}{\lambda}\frac{\nu}{\nu-2} = 0.1$, which simplify our analysis in the following sections. We sample observations for each θ_i from same sampling distribution $\mathcal{N}(\theta_i, \sigma_i^2)$, where $\sigma_i \sim Gamma(1, 5)$.

Then, we generate three sets of observations with corresponding prior,

- Normal prior observations: $X_i^{(N)} | \theta_i^{(N)} \sim \mathcal{N}(\theta_i^{(N)}, \sigma_i^2)$;

- Gamma prior observations: $X_i^{(G)}|\theta_i^{(G)} \sim \mathcal{N}(\theta_i^{(G)}, \sigma_i^2)$;
- Student's t prior observations: $X_i^{(T)}|\theta_i^{(T)} \sim \mathcal{N}(\theta_i^{(T)}, \sigma_i^2)$.

In this study, we generate 10 simulations with each simulation let $n = 50,000$. Since we are focusing on ranking the top $\alpha\%$, which is related to the upper tail of the posterior distribution in posterior mean ranking, we need to generate enough samples in top $\alpha\%$ ranking. Also, in the many real case of ranking, it is natural to have a large number of data, such as market basket data. Therefore, such a number of samples in each simulation satisfies our purpose.

2.3 Calculation of Posterior Means

We use the same normal prior, gamma prior and Student's t-prior distributions in simulation with different types of tail for estimating the posterior mean of these three sets of θ.

We use the same normal prior

$$g^N(\theta) = p(\theta|\mu, \tau^2) = \frac{1}{\sqrt{2\pi}\tau} e^{-\frac{(\theta-\mu)^2}{2\tau^2}} \tag{2.7}$$

with the hyperparameters $(\mu, \tau^2) = (1, 0.1)$ to estimate the posterior mean of these three sets of θ's,

Therefore, the posterior mean of these three sets of data using a normal distribution prior is given by an explicit form

$$\begin{aligned} \mathrm{E}_N[\theta_i|X_i, \sigma_i^2] &= \frac{\int_{\theta_i} \theta_i \; g^N(\theta_i) \; p(X_i|\theta_i, \sigma_i^2) \; d\theta_i}{\int_{\theta_i} g^N(\theta_i) \; p(X_i|\theta_i, \sigma_i^2) \; d\theta_i} \\ &= \frac{\tau^2}{\tau^2 + \sigma_i^2} X_i + \frac{\sigma_i^2}{\tau^2 + \sigma_i^2} \mu \end{aligned} \tag{2.8}$$

Equation (2.8) shows that the posterior mean of a normal/normal model is weighted mean of observed data and prior information. It is weighted by uncertainty of the prior and measurement error of the observed data, which are the variance prior $\hat{\tau}^2$ and the variance of observations σ_i^2. This form also explains the reason that posterior mean with normal prior prefers observations with low variance. The posterior mean of a smaller variance (σ_i^2) observation is weighted closer to the observation, which

give a larger posterior mean estimate. Therefore, the posterior mean method with normal prior give higher ranks to units with small variance.

The posterior mean when using the same gamma prior ($g^G(\theta)$) is

$$\begin{aligned} \mathrm{E}_G[\theta_i|X_i,\sigma_i^2] &= \frac{\int_{\theta_i}\theta_i\ g^G(\theta_i)\ p(X_i|\theta_i,\sigma_i^2)\ d\theta_i}{\int_{\theta_i} g^G(\theta_i)\ p(X_i|\theta_i,\sigma_i^2)\ d\theta_i} \\ &= \frac{\int_{\theta_i}\theta_i\ \theta^{\alpha-1}e^{\beta\theta_i}\ e^{-\frac{(X_i-\theta_i)^2}{2\sigma_i^2}}\ d\theta_i}{\int_{\theta_i}\theta_i^{\alpha-1}e^{\beta\theta_i}\ e^{-\frac{(X_i-\theta_i)^2}{2\sigma_i^2}}\ d\theta_i} \end{aligned} \tag{2.9}$$

with

$$g^G(\theta) = p(\theta|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{\beta\theta} \tag{2.10}$$

where (α,β) are equal to $(10,10)$ as used for the simulation.

And the posterior mean when using the same Studnet's t-prior ($g^t(\theta)$) is

$$\begin{aligned} \mathrm{E}_T[\theta_i|X_i,\sigma_i^2] &= \frac{\int_{\theta_i}\theta_i\ g^T(\theta_i)\ p(X_i|\theta_i,\sigma_i^2)\ d\theta_i}{\int_{\theta} g^T(\theta_i)\ p(X_i|\theta_i,\sigma_i^2)\ d\theta_i} \\ &= \frac{\int_{\theta_i}\theta_i\ (1+\frac{\lambda(\theta_i-\mu)^2}{\nu})^{-\frac{\nu+1}{2}}\ e^{-\frac{(X_i-\theta_i)^2}{2\sigma_i^2}}\ d\theta_i}{\int_{\theta_i}(1+\frac{\lambda(\theta_i-\mu)^2}{\nu})^{-\frac{\nu+1}{2}}\ e^{-\frac{(X_i-\theta_i)^2}{2\sigma_i^2}}\ d\theta_i} \end{aligned} \tag{2.11}$$

with

$$\hat{g}^T(\theta) = p(\theta|\nu,\lambda,\eta) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})}(\frac{\lambda}{\pi\nu})^{\frac{1}{2}}(1+\frac{\lambda(\theta-\eta)^2}{\nu})^{-\frac{\nu+1}{2}} \tag{2.12}$$

where (λ,η) are the same hyperparameters of prior equal to $(1,30)$.

However, we are not able to calculate an explicit form of posterior mean for the gamma prior and the t prior as we have for the normal prior, because the integrals above do not have an analytic solution. Therefore, we can only use numerical methods to get the posterior mean for these two priors. Here we use a Gibbs sampler by taking a selected prior and likelihood to sample from its posterior distribution and calculate the mean of those posterior distribution samples.

2.4 Comparison of the Results

In this section, We show the effect of using different types of tail distributions for estimating the posterior mean.

In Section 2.3, we notice that the normal/normal model provides straightforward form of posterior mean. There are some other benefits of using normal prior besides its explicit form of posterior mean. It also gives a better estimation of θ_i than just using observation as estimates in ranking problem.

			True Prior		
			Normal	Gamma	T
Estimating Prior	Normal	MSE	0.0258	0.0256	0.0266
		std.err	0.00025	0.00027	0.00251
	Gamma	MSE	0.0263	0.0252	0.0293
		std.err	0.00025	0.00028	0.00134
	T	MSE	0.0271	0.0267	0.0218
		std.err	0.00026	0.00029	0.00032

Table 2.1: Average MSE and standard error over 10 simulations

Table 2.1 is the average of total MSE of 50,000 samples over 10 simulations and the average standard error of the total MSE over 10 simulations. Here MSE for one simulation is the mean squared error between θ_j and its posterior mean estimation $\hat{\theta}_j = \mathrm{E}[\theta_j|X_j]$, defined as

$$\mathrm{MSE} = \frac{1}{n}\sum_{j=1}^{n}(\theta_j - \hat{\theta}_j)^2 = \frac{1}{n}\sum_{j=1}^{n}(\theta_j - \mathrm{E}[\theta_j|X_j])^2 \tag{2.13}$$

The standard error in Table 2.1 is the average standard error over 10 simulations. Since the MSE given in Table 2.1 is the estimate over 10 simulations, each with 50,000 points. For each simulation, we estimated the MSE as the mean of these 50,000 squared errors. However, there is some error in this based on the variance of the estimated MSE. The standard error is an indication of how much error present to the estimated MSE, which is calculated as equation (2.14)

$$\begin{aligned} \mathrm{std.err}^2 &= \frac{Var((\theta_j - \hat{\theta}_j)^2)}{n} \\ &= \frac{\sum_{j=1}^{n}[(\theta_j - \hat{\theta}_j)^2 - \frac{1}{n}\sum_{i=1}^{n}(\theta_i - \hat{\theta}_i)^2]^2/(n-1)}{n} \end{aligned} \tag{2.14}$$

Since posterior mean estimation minimizes the expect of a squared-error loss given by equation (2.3), MSE is a very reasonable measure in evaluating the precision of the posterior mean estimator under different priors.

Obviously, using the true prior is the best choice for the overall MSE, since using the correct prior gives the smallest MSE. The effects of using prior with too heavy a tail and using a prior with too light a tail are comparable. This means that using a normal prior for convenience is likely to be reasonable unless we have good reason to believe the prior is not normal.

In general, taking normal prior is good in overall estimation. Nevertheless, the following results show the weakness of normal prior in estimating the upper tail of posterior distribution, when we are more interested in the ranking the top α units. This problem affects the performance of normal/normal model in posterior mean ranking.

Two more measurements of the performance of posterior mean ranking method with different priors are given below. Firstly, we define the MSE of top ranked $\alpha\%$ of θ_i for one simulation,

$$\text{MSE}_{\alpha\%} = \frac{1}{k}\sum_{j=i_{[1]}}^{i_{[k]}}(\theta_j - \hat{\theta}_j)^2 = \frac{1}{k}\sum_{j=i_{[1]}}^{i_{[k]}}(\theta_j - \text{E}[\theta_j|X_j])^2 \tag{2.15}$$

where $i_{[1]}, i_{[2]}, \cdots, i_{[k]}$ are indices of θ. Here it is given by ranking the true list of parameters θ_i, that is $R_{i_{[1]}}(\theta) = 1, R_{i_{[2]}}(\theta) = 2, \cdots, R_{i_{[k]}}(\theta) = n\alpha\%$. Compared to the overall MSE, the top $\alpha\%$ MSE helps us focus on the accuracy of posterior mean estimators with different choices of prior in top units (θ_i's). These are the units we are most interested in for the ranking problem. This MSE is an important indicator, because the larger the MSE of posterior mean, the greater the chance of ranking in the wrong order.

Another good measurement of the quality of ranking is the mean of the top $\alpha\%$ of ranked θ ($\bar{\theta}_\alpha$) for on simulation given by

$$\bar{\theta}_\alpha = \frac{1}{k}\sum_{j=m_{[1]}}^{m_{[k]}}\theta_j, \tag{2.16}$$

where $m_{[1]}, m_{[2]}, \cdots, m_{[k]}$ are the indices of units (θ_i's) given by a ranking method. Here it is ranked by posterior mean method, which is $\hat{R}^{\text{PMR}}_{m_{[1]}} = 1, \hat{R}^{\text{PMR}}_{m_{[2]}} = 2, \cdots, \hat{R}^{\text{PMR}}_{m_{[k]}} = n\alpha\%$. Since the posterior mean ranking minimizes the loss function of ranking $L(\cdot)$ (equation (2.4)), which is the sum of θ_i's we should select in the top minus the sum of θ_j's selected by posterior mean ranking method in the top, the $\bar{\theta}_\alpha$ is positively

correlated to the second term of loss function, so increasing $\bar{\theta}_\alpha$ will decrease the $L(\cdot)$. Therefore, the mean of top $\alpha\%$ of θ is a good measurement of performance of posterior mean ranking with different choices of prior. The larger this value is, the better the result of ranking.

		True Prior		
		Normal	Gamma	T
Estimating Prior	Normal	0.1098	0.1622	0.4044
	Gamma	0.0997	0.1341	0.1817
	T	0.1188	0.1498	0.1632

Table 2.2: Average of the MSE of top ranked 1% of θ_i ($\text{MSE}_{1\%}$) over 10 simulations

Table 2.2 shows the average top 1% MSE of 10 simulations, using equation (2.15) with $\alpha = 0.01, k = 500$. In this table, it is shown that using a normal prior when the prior is actually a heavy-tailed Student's t-distribution, the MSE increases significantly, which is much worse than using a gamma distribution as prior. Conversely, using a heavy-tail prior when the true prior has a light-tail seems to be less harmful. In the first column of Table 2.2, the MSE of estimating θ with normal prior using a t-distribution as prior is 0.1188 compared to 0.1098 using the true normal prior. In the third column of Table 2.2, the MSE of estimating t-distribution as normal distribution is 0.4044 comparing using true t-distribution prior which is 0.1632. This difference is much larger, showing that using too light a tail in the prior has much larger negative impact than using too heavy a tail.

		True Prior		
		Normal	Gamma	T
Estimating Prior	Normal	1.7672	1.9675	2.1644
	Gamma	1.7616	1.9748	2.1793
	T	1.7608	1.9719	2.1824

Table 2.3: Average of the mean of top 1% of ranked θ ($\bar{\theta}_1$) over 10 simulations

Table 2.3 shows the average mean of the top 1%, which is the top 500 values of θ ($\bar{\theta}_{1\%}$), ranked by their posterior mean under different choices of prior. It shows that taking normal/normal model in posterior mean ranking produces less satisfactory results compared to using a t-distribution as prior. There is significant drop in $\bar{\theta}_{1\%}$ using a normal prior to analyse data with a heavy-tailed prior. Also when true prior

is medium-tailed, a t-distribution prior gives better results than a normal prior.

Moreover, the following two tables give us a deeper view of the heavy-tail problem using normal/normal model in posterior mean ranking. These results are for the top 5% ranking using posterior mean.

		True Prior		
		Normal	Gamma	T
Estimating Prior	Normal	0.0713	0.0941	0.1258
	Gamma	0.0691	0.0861	0.0815
	T	0.0838	0.1024	0.0867

Table 2.4: Average of the MSE of top ranked 5% of θ_i ($\text{MSE}_{5\%}$) over 10 simulations

		True Prior		
		Normal	Gamma	T
Estimating Prior	Normal	1.5785	1.6835	1.6181
	Gamma	1.5770	1.6855	1.6167
	T	1.5780	1.6835	1.6204

Table 2.5: Average of the mean of top 5% of ranked θ ($\bar{\theta}_5$) over 10 simulations

Table 2.4 shows the result of average top 5% MSE over 10 simulations, using equation (2.15) with $\alpha = 0.05, k = 2500$. And Table 2.5 shows the average mean of top 5%, which is top 2500 of θ ($\bar{\theta}_{5\%}$), ranked by their posterior mean under different choices of priors. From Table 2.4, we see that taking prior as a gamma distribution gives the best posterior mean estimation to true value θ over these three choices of priors. As we can see in Table 2.5, in the situation that the true prior θ is a normal prior $\theta_1^{(N)}, \theta_2^{(N)}, \cdots, \theta_n^{(N)}$ or a gamma prior $\theta_1^{(G)}, \theta_2^{(G)}, \cdots, \theta_n^{(G)}$ taking t-distribution as prior gives good results for $\bar{\theta}_{5\%}$. This confirms the conclusion from Tables 2.2 and Table 2.3 that using a prior with too light a tail is far more serious than using a prior with too heavy a tail. This suggests that when selecting a parametric model for the prior distribution in a ranking problem, it is safer to be on the side of heavier tails.

2.5 Different Choices of Prior Variance

In this section, we will study the influence of using the incorrect variance of the prior distribution on the accuracy of the estimated ranking.

Table 2.6 is the average overall MSE over 10 simulated sets of data calculated the same way as equation (2.13) with their true family of prior but different variance of prior distribution. Table 2.7 is the average top 5% MSE over 10 simulated sets of data calculated the same way as equation (2.15) with their true family of prior but different variance of prior distribution. Here the posterior means are given by equation (2.8), equation (2.9) and equation (2.11) with different values for the parameters of the prior, which allows us to change the variance of prior with fixed mean of prior. That is

- Posterior mean of normal prior $\mathrm{E}_N[\theta_i^{(N)}|X_i^{(N)}]$ with hyperparameters being

$$(\mu, 0.5\tau^2),\ (\mu, 0.8\tau^2),\ (\mu, \tau^2),\ (\mu, 1.2\tau^2),\ (\mu, 1.5\tau^2) \tag{2.17}$$

- Posterior mean of gamma prior $\mathrm{E}_G[\theta_i^{(G)}|X_i^{(G)}]$ with

$$0.5^{-1}(\alpha, \beta),\ 0.8^{-1}(\alpha, \beta),\ (\alpha, \beta),\ 1.2^{-1}(\alpha, \beta),\ 1.5^{-1}(\alpha, \beta) \tag{2.18}$$

- Posterior mean of t-distribution prior $\mathrm{E}_T[\theta_i^{(T)}|X_i^{(T)}]$ with

$$(0.5^{-1}\lambda, \eta),\ (0.8^{-1}\lambda, \eta),\ (\lambda, \eta),\ (1.2^{-1}\lambda, \eta), (1.5^{-1}\lambda, \eta) \tag{2.19}$$

	Different Variance of from True Family of Prior				
	$0.5\mathrm{Var}(\theta)$	$0.8\mathrm{Var}(\theta)$	$\mathrm{Var}(\theta)$	$1.2\mathrm{Var}(\theta)$	$1.5\mathrm{Var}(\theta)$
Normal Prior	0.0281	0.0261	0.0258	0.0259	0.0266
Gamma Prior	0.0273	0.0254	0.0252	0.0253	0.0259
T Prior	0.0231	0.0222	0.0218	0.0222	0.0225

Table 2.6: Average overall MSE using different prior variances over 10 simulations

Table 2.6 reveals as expected that average overall MSE of posterior mean estimations using their true family of prior is minimized by accurately estimating the variance of the prior. And errors in either direction have similar effects on the MSE.

However, it appears to be different when we concentrate on the top $\alpha\%$ θ_i. Choosing a larger variance of prior than its true variance of prior is less harmful to our posterior mean estimators in all three priors. Indeed, overestimating the variance may even improve the MSE of the top values.

	Different Variance of True Family of Prior				
	0.5Var(θ)	0.8Var(θ)	Var(θ)	1.2Var(θ)	1.5Var(θ)
Normal Prior	0.1001	0.0794	0.0713	0.0657	0.0599
Gamma Prior	0.1181	0.0934	0.0861	0.0784	0.0727
T Prior	0.1055	0.0932	0.0867	0.0833	0.0783

Table 2.7: Average top 5% MSE ranked by true values ($\text{MSE}_{5\%}$) using different prior variance over 10 simulations

Table 2.8 and Table 2.9 present the average overall MSE and the average of top 5% MSE results over 10 simulations.

The posterior means are given by equations (2.8), using a normal prior with different variances of hyperparameters: $(\mu, 0.5\tau^2)$, $(\mu, 0.8\tau^2)$, (μ, τ^2), $(\mu, 1.2\tau^2)$, $(\mu, 1.5\tau^2)$. The last column of these two Table 2.8 and Table 2.9 are the MSE result when the variance of prior goes to infinity, which is the MSE between θ_j and its maximum likelihood estimator, which is X_j here.

	Different Variance of Normal Estimating Prior					
	0.5Var(θ)	0.8Var(θ)	Var(θ)	1.2Var(θ)	1.5Var(θ)	∞
Normal Prior	0.0278	0.0258	0.0256	0.0257	0.0264	0.0798
Gamma Prior	0.0281	0.0261	0.0257	0.0259	0.0266	0.0802
T Prior	0.0295	0.0268	0.0263	0.0262	0.0267	0.0795

Table 2.8: Average overall MSE using normal prior with different variances over 10 simulations

	Different Variance of Normal Estimating Prior					
	0.5Var(θ)	0.8Var(θ)	Var(θ)	1.2Var(θ)	1.5Var(θ)	∞
Normal Prior	0.0999	0.0794	0.0713	0.0657	0.0598	0.0799
Gamma Prior	0.1362	0.1061	0.0941	0.0855	0.0763	0.0786
T Prior	0.1945	0.1455	0.1258	0.1115	0.0965	0.0819

Table 2.9: Average top 5% MSE ranked by true values ($\text{MSE}_{5\%}$) using normal prior with different variances over 10 simulations

Table 2.8 and Table 2.9 show the same results as Table 2.6 and Table 2.7 that when using normal prior with different prior variances, the average overall MSE is minimized by accurately estimating the variance of prior, but overestimating the variance leads to decrease in MSE of the top 5% units which can be most beneficial where the true prior is a heavy-tailed t-distribution. Therefore, using overestimated

prior variance in the normal/normal model can be less harmful in estimating the top ranked θ, and therefore is better for ranking θ's.

2.6 Conclusion of Robustness

Based on the comparison results in the previous sections, we draw the following conclusions about the best choice of prior and a good estimation of prior parameters in posterior mean ranking.

As in Section 2.4, a normal/normal model of choosing normal prior generally produces good results in estimating the posterior mean of all units. Such model is simple to estimate. However, when it comes to choosing the top α units, posterior mean ranking of normal/normal model appears to be less acceptable. Moreover estimating the prior using a heavy-tail distribution such as the Student's t-distribution from the previous sections is less harmful and more robust in posterior mean ranking.

Lastly, when we are focusing on ranking the top α units, the accuracy of estimating the parameters of the prior distribution seems to be less important. In some cases, the posterior mean estimator is better when we overestimate the variance of the prior.

In the next chapter, we will apply what we have learned about the choice of prior distribution to the problem of ranking and selection in market basket analysis. We will compare a number of popular ranking methods to this problem.

Chapter 3

Application of Various Ranking Methods in Market Basket Analysis

3.1 Introduction

Of all sorts of applications of ranking and selection from large scale data, choosing the leading association rules is perhaps one of the most interesting and underdeveloped examples. In this chapter, we give an example of applying some general ranking and selection methods to rank the top associated items in market basket analysis.

First a measurement (leverage) is chosen to evaluate association patterns based on transaction data. Then I will go through some of its properties and generate a simulation of it. With the leverage measurement, I'm going to simulate a set of market basket data. Finally, I will apply some ranking and selection methods to my simulated market basket data, including empirical Bayesian approaches such as posterior mean and r-values method, and some alternative methods, and compare the results.

3.2 Objective Interestingness Measure: Leverage

Firstly, I would like to give a brief introduction of the selected objective interestingness measure, leverage, which we have mentioned in Section 1.2. Since the size of market basket data could be large, it is natural to generate hundreds and thousands of association rules. Among all these rules, identifying the top association rules of interest is a difficult job. Therefore it is important to evaluate association rules objectively using well-accepted measure.

It is more convenient to use Table 3.1 which is a 2-way contingency table for itemsets A and B to show some properties of leverage. We use notation $\overline{A}$ ($\overline{B}$) to indicate the event that itemsets A (B) don't all occur in a transaction. Each f_{ij} represents a frequency count of the corresponding event where i represents the

	B	$\overline{B}$	
A	f_{11}	f_{10}	f_{1+}
$\overline{A}$	f_{01}	f_{00}	f_{0+}
	f_{+1}	f_{+0}	N

Table 3.1: A 2-way contingency table for itemsets A and B

occurrence of A, j represents the occurrence of B. $f_{1+}(f_{+1})$, represent the support count of A(B) defined by equation (1.4). N is the total number of transactions.

The Leverage measure, Piatetsky-Shapiro (PS) measure defined by equation (1.7) is expressed as

$$\text{leverage}(A \Rightarrow B) = \frac{f_{11}}{N} - \frac{f_{1+}f_{+1}}{N^2} \tag{3.1}$$

Since leverage (PS) is symmetric, we simplify the notation of leverage as leverage(A, B). Also using probabilities to express the leverage (PS) measure, we get

$$\begin{aligned}
\text{leverage}(A, B) =& P(AB) - P(A)P(B) \\
=& P_{11} - P_{1+}P_{+1} \\
=& P_{11} - (P_{10} + P_{11})(P_{01} + P_{11}) \\
=& P_{11} - (P_{10}P_{01} + P_{10}P_{11} + P_{01}P_{11} + P_{11}P_{11}) \\
=& P_{11}(1 - P_{10} - P_{01} - P_{11}) - P_{10}P_{01} \\
=& P_{11}P_{00} - P_{01}P_{10}
\end{aligned} \tag{3.2}$$

where P_{11} is the probability that itemsets A and B both occur in a transaction, P_{00} is the probability of neither of A and B occur, P_{10} (P_{01}) is the probability only one itemset $A\overline{B}$ ($\overline{A}B$) occurs, P_{1+} (P_{+1}) is the probability of A (B) occurring.

Under the expression by probabilities, it is more explicit that leverage measures difference between the appearance of A and B together and what would be expected if A and B were independent.

In addition, there are a few more advantages that can accrue by choosing leverage measurement:

- The leverage treats the appearance of an itemset and disappearance of the identical itemset differently, that is leverage$(A, B) \neq$ leverage$(A, \overline{B})$.

- The leverage is relatively complete and simply interpretable. It is easy to make the case on business grounds that it represents the extra ratio of cases where

items A and B both are sold, compared to what would be expected if they were independent.

- The range of leverage (PS) measurement is between $-\frac{1}{4}$ to $\frac{1}{4}$, where 0 indicates items A and B are independent. Positive leverage indicates that items A and B are positively associated, and negative leverage indicates a negative association.
- Calculation of the theoretical variance of the estimator of leverage is reasonably straightforward, compared to some other measures.

Now our ranking and selection procedure is based on ranking the leverage measurement. Suppose the leverage of any two itemsets A and B is represented by $\theta_{A,B}$, we are ranking n association rules.

$$\theta_{A,B} = \text{leverage}(A, B) = PS(A, B) \tag{3.3}$$

3.2.1 Sample Distribution of Estimated Leverage

In order to learn how to estimate the leverage from the data, assume $\{f_{11}, f_{10}, f_{01}, f_{00}\}$ from the contingency table above is sampled from a multinomial distribution with parameters $(N, P_{11}, P_{01}, P_{10}, P_{00})$.

	$A \cap B$	$\overline{A} \cap B$	$A \cap \overline{B}$	$\overline{A} \cap \overline{B}$
No.	X_i	Y_i	Z_i	W_i
1	0	0	1	0
2	0	0	0	1
3	0	1	0	0
4	0	0	0	1
5	1	0	0	0

Table 3.2: Indicator variable for multinomial distribution

Denote $\{X_j,\ Y_j,\ Z_j,\ W_j\},\ j = 1, \cdots, N$, to be the indicator variables of jth trial, where 0 indicates fail and 1 indicates success. Thus our simulated data can be expressed in the form of Table 3.2.

The proportions of each case give the maximum likelihood estimates of $P_{11}, P_{01}, P_{10}, P_{00}$

$$\hat{P}_{11} = \frac{f_{11}}{N} = \frac{\sum X_i}{N}, \qquad \hat{P}_{01} = \frac{f_{01}}{N} = \frac{\sum Y_i}{N}, \qquad \hat{P}_{10} = \frac{f_{10}}{N} = \frac{\sum Z_i}{N}, \qquad \hat{P}_{00} = \frac{f_{00}}{N} = \frac{\sum W_i}{N}.$$

Therefore, the maximum likelihood estimated leverage $(\hat{\theta})$ is given by

$$\hat{\theta} = \hat{P}_{11}\hat{P}_{00} - \hat{P}_{01}\hat{P}_{10}. \tag{3.4}$$

Expected Value of the Leverage (PS) Estimator

$$
\begin{aligned}
\mathrm{E}[\hat{\theta}] &= \mathrm{E}[\hat{P}_{11}\hat{P}_{00} - \hat{P}_{01}\hat{P}_{10}] \\
&= \mathrm{E}[\hat{P}_{11}\hat{P}_{00}] - \mathrm{E}[\hat{P}_{01}\hat{P}_{10}] \\
&= \mathrm{E}\left[\frac{\sum_i X_i}{N}\frac{\sum_j W_j}{N}] - \mathrm{E}[\frac{\sum_k Y_k}{N}\frac{\sum_l Z_l}{N}\right] \\
&= \frac{1}{N^2}(\mathrm{E}[\sum_i X_i \sum_j W_j] - \mathrm{E}[\sum_k Y_k \sum_l Z_l]) \\
&= \frac{1}{N^2}(\mathrm{E}[\sum_{i\neq j} X_i W_j] - \mathrm{E}[\sum_{k\neq l} Y_k Z_l]) \\
&= \frac{1}{N^2}([N(N-1)P_{11}P_{00}] - [N(N-1)P_{01}P_{10}]) \\
&= \frac{(N-1)}{N}(P_{11}P_{00} - P_{01}P_{10}) \\
&= \frac{(N-1)}{N}\theta
\end{aligned}
\tag{3.5}
$$

Therefore our estimated leverage $\hat{\theta}$ is an asymptotically unbiased estimator of leverage θ when N is large enough.

Variance of the Leverage (PS) Estimator

We then calculate the variance of the estimated leverage $\hat{\theta}$, which is

$$
\begin{aligned}
\mathrm{Var}(\hat{P}_{11}\hat{P}_{00} - \hat{P}_{01}\hat{P}_{10}) =& \mathrm{E}[(\hat{P}_{11}\hat{P}_{00} - \hat{P}_{01}\hat{P}_{10})^2] - (\mathrm{E}[\hat{P}_{11}\hat{P}_{00} - \hat{P}_{01}\hat{P}_{10}])^2 \\
=& \mathrm{E}[(\hat{P}_{11}^2\hat{P}_{00}^2 + \hat{P}_{01}^2\hat{P}_{10}^2 - 2\hat{P}_{11}\hat{P}_{01}\hat{P}_{10}\hat{P}_{00})] \\
& - (\mathrm{E}[\hat{P}_{11}\hat{P}_{00} - \hat{P}_{01}\hat{P}_{10}])^2 \\
=& \mathrm{E}[(\hat{P}_{11}^2\hat{P}_{00}^2)] + \mathrm{E}[(\hat{P}_{01}^2\hat{P}_{10}^2)] - 2\mathrm{E}[\hat{P}_{11}\hat{P}_{00}\hat{P}_{01}\hat{P}_{10}] \\
& - (\mathrm{E}[\hat{P}_{11}\hat{P}_{00} - \hat{P}_{01}\hat{P}_{10}])^2
\end{aligned}
\tag{3.6}
$$

We can divide this formula into small pieces and calculate each part of it.

The first part of this formula is

$$\begin{aligned}
\mathrm{E}[(\hat{P}_{11}^2\hat{P}_{00}^2)] =& \mathrm{E}\left[\sum_{ijkl}\frac{X_iX_jW_kW_l}{N^4}\right] \\
=& \mathrm{E}\Bigg[\frac{1}{N^4}\{\sum_{i=j}X_i(\sum_{k\neq l\neq i}W_kW_l+\sum_{k=l\neq i}W_k) \\
&+\sum_{i\neq j}X_iX_j(\sum_{k\neq l\neq i\neq j}W_kW_l+\sum_{k=l\neq i,j}W_k)\}\Bigg] \\
=& \frac{1}{N^4}\{NP_{11}[(N-1)(N-2)P_4^2+(N-1)P_{00}] \\
&+N(N-1)P_1^2[(N-2)(N-3)P_4^2+(N-2)P_{00}]\} \\
=& \frac{N-1}{N^3}P_{11}P_{00}[1+(N-2)P_{11}+(N-2)P_{00}+(N-2)(N-3)P_{11}P_{00})]
\end{aligned} \tag{3.7}$$

Then the third part of this formula is

$$\begin{aligned}
\mathrm{E}[\hat{P}_{11}\hat{P}_{00}\hat{P}_{01}\hat{P}_{10}] &= \frac{1}{N^4}\sum_{\substack{i,j,k,l\\ \text{distinct}}}X_iY_jZ_kW_l \\
&= \frac{(N-1)(N-2)(N-3)}{N^3}P_{11}P_{01}P_{10}P_{00}
\end{aligned} \tag{3.8}$$

The last part of it is

$$\begin{aligned}
(\mathrm{E}[\hat{P}_{11}\hat{P}_{00}-\hat{P}_{01}\hat{P}_{10}])^2 &= \frac{(N-1)^2}{N^2}(P_{11}P_{00}-P_{01}P_{10})^2 \\
&= \frac{N-1}{N^3}[N(N-1)(P_{11}^2P_{00}^2+P_{01}^2P_{10}^2-2P_{11}P_{01}P_{10}P_{00})]
\end{aligned} \tag{3.9}$$

So the variance can be written as

$$\begin{aligned}
\mathrm{Var}(\hat{P}_{11}\hat{P}_{00}-\hat{P}_{01}\hat{P}_{10}) =& \frac{N-1}{N^3}[P_{11}P_{00}(1+(N-2)P_{11}+(N-2)P_{00} \\
&+(N-2)(N-3)P_{11}P_{00})+P_{01}P_{10}(1+(N-2)P_{01} \\
&+(N-2)P_{10}+(N-2)(N-3)P_{01}P_{10}) \\
&-2(N-2)(N-3)P_{11}P_{01}P_{10}P_{00} \\
&-N(N-1)(P_{11}^2P_{00}^2+P_{01}^2P_{10}^2-2P_{11}P_{01}P_{10}P_{00})] \\
=& \frac{N-1}{N^3}[P_{11}P_{00}+P_{01}P_{10} \\
&+(N-2)(P_{11}^2P_{00}+P_{11}P_{00}^2+P_{01}^2P_{10}+P_{01}P_{10}^2) \\
&-(4N-6)P_{11}^2P_{00}^2-(4N-6)P_{01}^2P_{10}^2 \\
&+2(4N-6)P_{11}P_{01}P_{10}P_{00}]
\end{aligned} \tag{3.10}$$

And since

$$(N-2)(P_{11}^2P_{00}+P_{11}P_{00}^2+P_{01}^2P_{10}+P_{01}P_{10}^2) =(N-2)[P_{11}P_{00}(1-P_{01}-P_{10}) + P_{01}P_{10}(1-P_{11}-P_{00})] \tag{3.11}$$

So the variance can be simplified as

$$\sigma^2(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) =\frac{N-1}{N^3}[(P_{11}P_{00}+P_{01}P_{10}) + (N-2)((P_{11}P_{00}+P_{01}P_{10}) - P_{11}P_{01}P_{10} - P_{11}P_{01}P_{00} - P_{11}P_{10}P_{00} - P_{01}P_{10}P_{00}) - (4N-6)(P_{11}P_{00}-P_{01}P_{10})^2] \tag{3.12}$$

Therefore the variance of the estimator (σ^2) is related to P_{11}, P_{00}, P_{01} and P_{10} which is related to θ.

3.2.2 Simulation Result of Leverage

Assume a true leverage θ is 0.007, with $P_{11} = 0.01, P_{01} = 0.05, P_{10} = 0.04, P_{00} = 0.90$. We generate 10,000 estimated leverages calculated from the sample proportion of a multinomial distribution population, each with $N = 100,000$ and $P_{11} = 0.01, P_{01} = 0.05, P_{10} = 0.04, P_{00} = 0.90$. The following two graphs show some properties of $\hat{\theta}$.

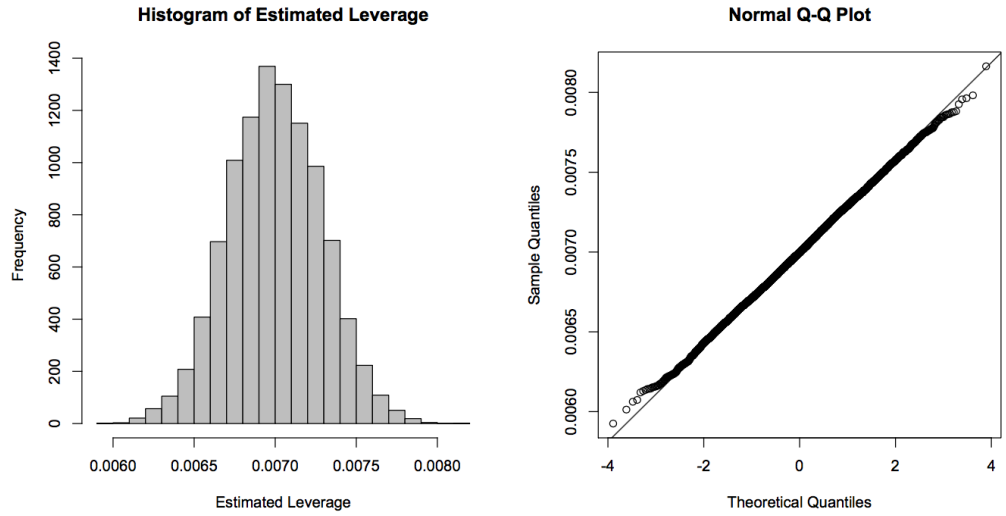

Figure 3.1: Histogram plot and Q-Q plot of $\hat{\theta}$

The left panel of Figure 3.1 is the histogram plot of estimated leverage, while the right panel is the Q-Q plot of sample quantile against theoretical normal quantile with a theoretical line.

These two figures show that the distribution of estimated leverage $\hat{\theta}$ is well approximated by normal distribution with mean θ and variance σ^2. Even though these two parameters are correlated, the correlation between mean and variance sampling distribution shouldn't harm our simulation in ranking problem. Therefore, we assume these two parameter are independent, in conclusion, $\hat{\theta}|\theta, \sigma^2 \sim \mathcal{N}(\theta, \sigma^2)$.

3.3 Simulated Market Basket Data

To simplify our simulation, rather than simulate a full market basket analysis, we will simulate market basket analysis for independent pairs of itemsets. That is, we will generate contingency tables like Table 3.1 for independent pairs of itemsets. This will give a similar marginal distribution for the leverages, but will ignore the relationship between different pairs of itemsets due to the overlap. Since we have not yet adapted our method to account for this relationship, simulating in this way allows us to focus on the issues investigated in Chapter 2.

Firstly, we simulate the probabilities of any itemset occurring in a transaction, which is the probabilities of a customer buying item, following a beta distribution with parameters $(1, 50)$. Here, the reason for choosing a beta distributions is that it represents a typical situation where most items are fairly rare, and a few are more common. Thus

The probability that itemset A occurs in a transaction is P_{1+}.

The probability that itemset B occurs in a transaction is P_{+1}.

P_{1+} and P_{+1} are independently distributed as Beta(1,50).

The prior distribution of leverage (PS) θ is

$$\theta \sim \begin{cases} 0 & \text{with probability } 0.8 \\ \mathcal{N}(0, \tau^2) & \text{with probability } 0.2 \end{cases} \tag{3.13}$$

where $\tau = \min\{0.2P_{1+}P_{+1},\ 0.0002\}$. The simulated leverage values are chosen to follow a typical situation, where most pairs of itemsets have no association. We also assume that all the different association rules are independent. Such an assumption

is not very realistic, but it is a useful simplification. And it shouldn't do any harm for our simulation. The choice of using truncated variance is designed to avoid negative simulated values of P_{ij} or the values greater than 1, with high probability.

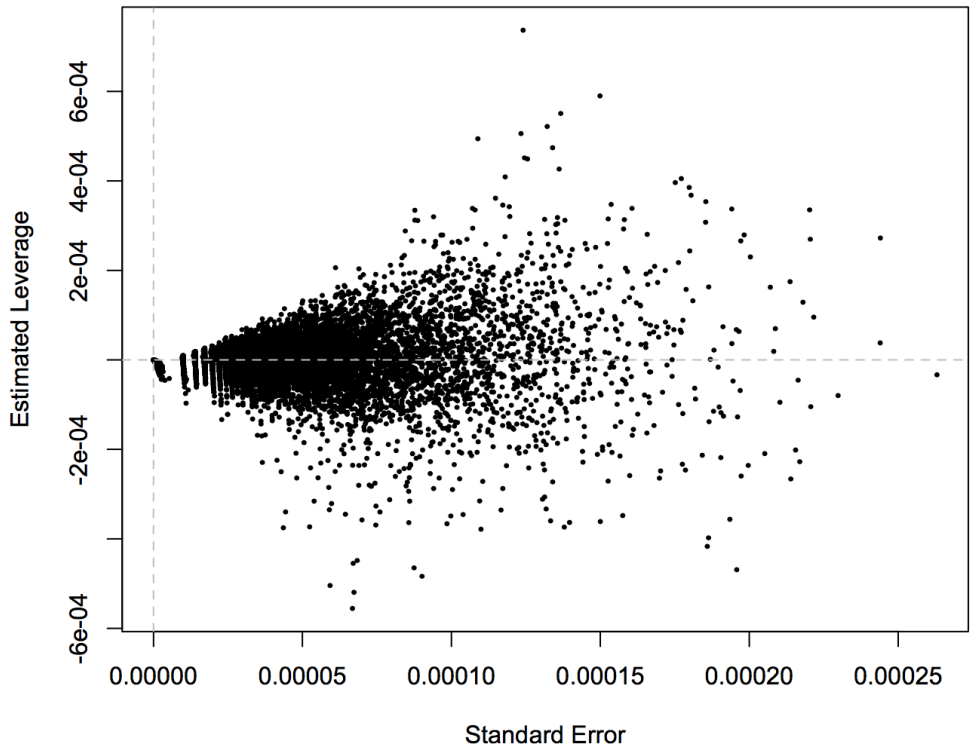

Figure 3.2: A market basket simulated data sample: 10,000 association rules measured by leverage from 100,000 transactions.

Then we can use θ, P_{1+} and P_{+1} to get P_{11}, P_{01}, P_{10} and P_{00} with the following equations:

$$P_{11} = P_{1+}P_{+1} + \theta$$
$$P_{10} = P_{1+} - P_{11}$$
$$P_{01} = P_{+1} - P_{11}$$
$$P_{00} = 1 - P_{11} - P_{01} - P_{10}$$

Since it is easy to generate a large number of association rules from a real market basket transaction data, we want to make this simulation study to be similar to a real market basket data example. Therefore, we simulate 10,000 contingency tables to calculate the corresponding true leverage of each association rules

(θ_i, for $i = 1, 2, \cdots, 10,000$). For each contingency table and corresponding true leverage, we generate an estimated leverage from the multinomial distribution with parameters $(N, P_{11}, P_{01}, P_{10}, P_{00})$ with $N = 100,000$ using equation (3.4). And we calculate the variance of estimated leverage (σ^2). Note that, since in the real market basket data P_{11}, P_{00}, P_{01} and P_{10} are usually unknown, we use estimated proportions $(\hat{P}_{11}, \hat{P}_{01}, \hat{P}_{10}, \hat{P}_{00})$ to calculate the variance.

In this simulated market basket data, we obtain 10,000 leverage (θ_i), corresponding estimated leverage ($\hat{\theta}_i$), and the variance of this estimation σ_i^2, for $i = 1, 2, \cdots, 10,000$.

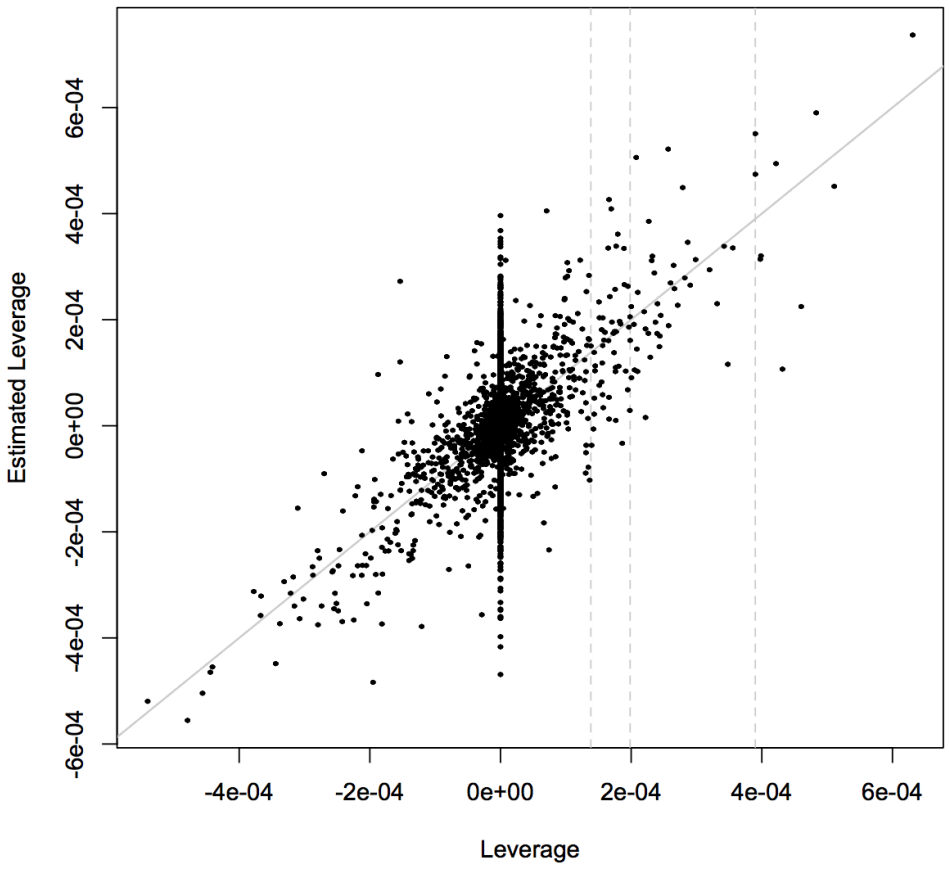

Figure 3.3: A market basket simulated data sample: 10,000 association rules measured by leverage from 100,000 transactions. The gray dash lines are the top 0.1%, top 1% and top 5% of true leverage θ_i.

Figure 3.2 and Figure 3.3 show the results of the simulated market basket data. Figure 3.2 is a scatterplot of estimates and estimated standard error similar to what would be available in a real data example. Note that the standard errors are estimated

from estimated probabilities, instead of true simulation probabilities. In a real data case, it is natural that we can only get estimates and standard errors from observations. Figure 3.2 also shows the fact that there is some relationship between estimates and standard error, since they are both related to the true values of leverage.

Figure 3.3 shows that although the large θ_i tend to have large estimated $\hat{\theta}_i$, there is still some variation in the estimated leverage. Also there are a large number of θ_i equal to zero which makes the prior distribution of θ more heavy-tailed which, as explained previously, can lead to problems with using a normal prior for ranking. For our purposes, we are very concerned about the tail, in particular, the top $\alpha\%$ of associations with the highest true leverage.

3.4 Result of General Rank Methods

In this section, we provide the result of ranking applying different ranking methods, such as posterior mean method with t-prior and normal prior, r-value method with normal prior, MLE method and p-value method to our simulated market basket data.

It is convenient to use the same description as in usual ranking problems with equation (1.1). In the simulated market basket data, there are n=10,000 association rules. Assume the ith estimate of leverage $\hat{\theta}_i$ is calculated from the observations sampled from corresponding population with unknown real-valued parameter of interest θ_i, that is the n populations have density functions

$$f(\hat{\theta}_1;\theta_1,\sigma_1^2),\ f(\hat{\theta}_2;\theta_2,\sigma_2^2),\ \cdots,\ f(\hat{\theta}_n;\theta_n,\sigma_n^2), \tag{3.14}$$

where $\hat{\theta}_i$ is the estimates of θ_i calculated from observations, and σ_i^2 as the variance of the estimate $\hat{\theta}_i$. We denote $g(\theta)$ to be the prior density of θ_i, and assume the θ_i is i.i.d $g(\theta)$.

The best result should be given by, using a three-stage hierarchical mixture model based on the true prior:

$$\begin{aligned} f(\hat{\theta}_i;\theta_i,\sigma_i^2) &= p(\hat{\theta}_i|\theta_i,\sigma_i^2) = \mathcal{N}(\theta_i,\sigma_i^2) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(\hat{\theta}_i-\theta_i)^2}{2\sigma_i^2}} \\ g(\theta_i) &= \gamma g_1(\theta_i) + (1-\gamma) g_2(\theta) \end{aligned} \tag{3.15}$$

where $g_1(\theta_i)$ is 0 and $g_2(\theta_i)$ is $\mathcal{N}(0,\tau^2)$ with $\gamma = 0.8$. The τ is given by

$$\tau = 0.2ab\ \mathrm{I}(0.2ab < 0.0002) + 0.0002\ \mathrm{I}(0.2ab \geq 0.0002) \tag{3.16}$$

with $p(a)$ and $p(b)$ are Beta(1,50). Our task is to rank units by θ_i from large to small here.

Moreover, our discussion focuses on ranking based on the positive estimated leverage $\hat{\theta}_i$, which is a positive association rule. It is important to separate positive and negative associations in business applications, since positive association rules represents benefits we can earn from an association. Generally, we can still rank the negative association rules using same method. We will apply what we have learned in Chapter 2 and use a heavy-tailed prior to analyse the data.

3.4.1 Posterior Mean Ranking Method

Normal/Normal Model

The first result is applying the posterior mean ranking method to market basket data to rank the association rules. A simple two-stage hierarchical model can be established with likelihood density $f(\hat{\theta}_i; \theta_i, \sigma_i^2)$ being normal $\mathcal{N}(\theta_i, \sigma_i^2)$, and the prior of θ_i for $i = 1, 2, \cdots, n$ being normal $\mathcal{N}(\hat{\mu}, \hat{\tau}^2)$. That is,

$$\hat{g}(\theta) = p(\theta_i|\hat{\mu}, \hat{\tau}^2) = \frac{1}{\sqrt{2\pi}\hat{\tau}} e^{-\frac{(\theta_i - \hat{\mu})^2}{2\hat{\tau}_i^2}} \tag{3.17}$$

where $(\hat{\mu}, \hat{\tau})$ are estimates of prior hyperparameters (μ, τ) through maximum likelihood from our samples $\hat{\theta}_i$ and σ_i^2. Then the posterior mean of θ_i is easily given by equation (2.8), which is

$$\mathrm{E}[\theta_i|\hat{\theta}_i, \sigma_i^2] = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \sigma_i^2}\hat{\theta}_i + \frac{\sigma_i^2}{\hat{\tau}^2 + \sigma_i^2}\hat{\mu} \tag{3.18}$$

In Figure 3.4, the left panel is a plot of estimated leverage against its estimated standard error, while the right panel is the plot of estimated leverage against true leverage. Those points colored in red are the top 1% selected by posterior mean ranking method using normal/normal model here. As we can see, using normal prior, posterior mean ranking prefers estimates with small standard error. From the right panel, we are able to see that the method gives bad results in ranking the top 1% units by leverage.

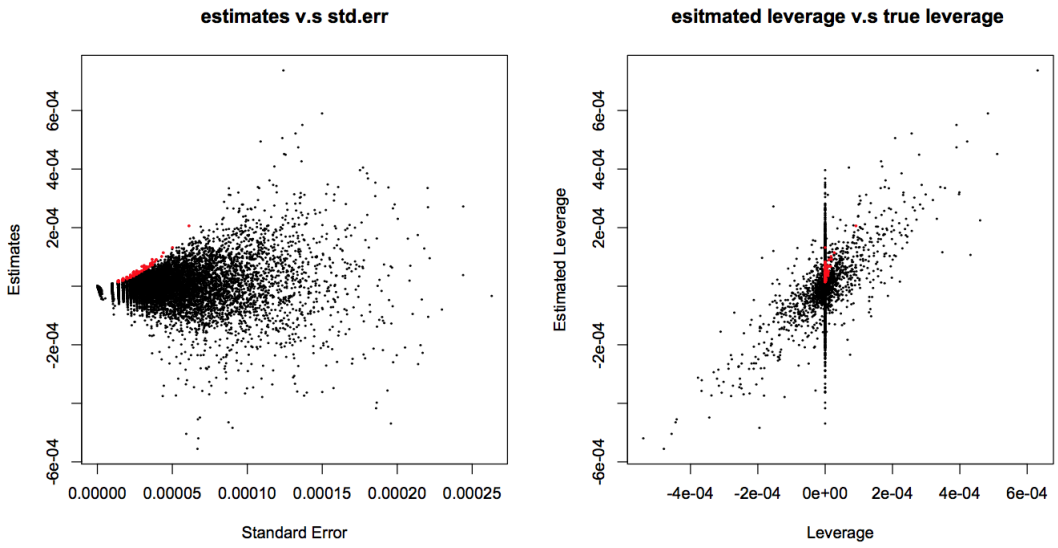

Figure 3.4: Posterior mean: The red points are the top 1% association rules ranked by posterior mean ranking using a normal distribution estimated from the data as prior.

T/Normal Model

Now, we change the choice of prior in the posterior mean ranking method from a normal distribution with light-tail to a Student's t-distribution with heavy-tail. A t/normal model is applied in the empirical Bayes method.

The posterior mean of a Student's t-distribution $\mathrm{E}_T[\theta_i|\hat{\theta}_i, \sigma_i^2]$ is given by equation (2.11). From the discussion of estimations of parameters for prior in Section 2.5, it is not clear what the most appropriate estimates are for the hyperparameters, but the estimates should not harm our ranking too much. Therefore, we use the first two moment estimators to estimates the hyperparameters of the prior for simplicity of computation. We choose the degrees of freedom of the t-prior to be 3, so that this will satisfy our intention of selecting a heavy-tail prior. That is,

- Location parameter $\hat{\eta} = \sum_{i=1}^{N} \hat{\theta}_i = 4.059472 \times 10^{-7}$;
- Inverse scale parameter $\hat{\lambda} = \frac{\frac{\nu}{\nu-2}}{\mathrm{Var}[\hat{\theta}_i] - \mathrm{E}[\sigma_i^2]} = 630500364$;
- Degrees of freedom $\hat{\nu} = 3$.

Figure 3.5 shows the results of top 1% leverage (red points) under posterior mean

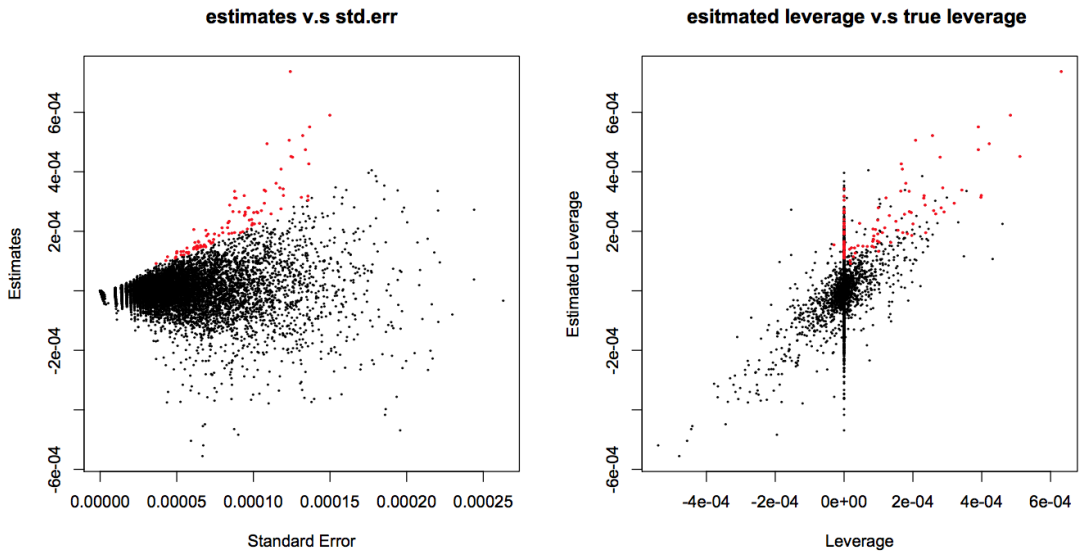

Figure 3.5: Posterior mean: The red points are the top 1% association rules ranked by posterior mean ranking method using student's t-distribution as prior.

ranking using student's t distribution as prior. The left panel shows the trade-off between the estimates and standard errors. Since it takes a heavy-tail distribution as prior, this ranking method puts more weight on estimated value, and less weight on standard error, compared to Figure 3.4, which overpopulates estimates with small standard error. Both these two figures, Figure 3.5 and Figure 3.4 support our conclusion that using a conjugate prior will overweight standard error, owing to estimating the prior as a too light-tailed distribution. This shows the ranking method chooses a high estimate as generated from a very large error instead of its large true leverage.

3.4.2 R-value Ranking Method with Normal Prior

Figure 3.6 gives the result of using the r-value ranking method using a normal prior with hyperparameters being estimated through maximum likelihood. The left panel of Figure 3.6 is plot of estimated leverage against its estimated standard error. The right panel shows estimated leverage against true leverage. Red points in both plots are the top 1% units ranked by r-value. As the right panel of this figure shows, r-value ranking chooses some units agreeing with the best selection. It selects the top units with largest leverage and large standard error. Although r-value tries to maximize the agreement between observation/estimates and prior, the method is hugely affected by

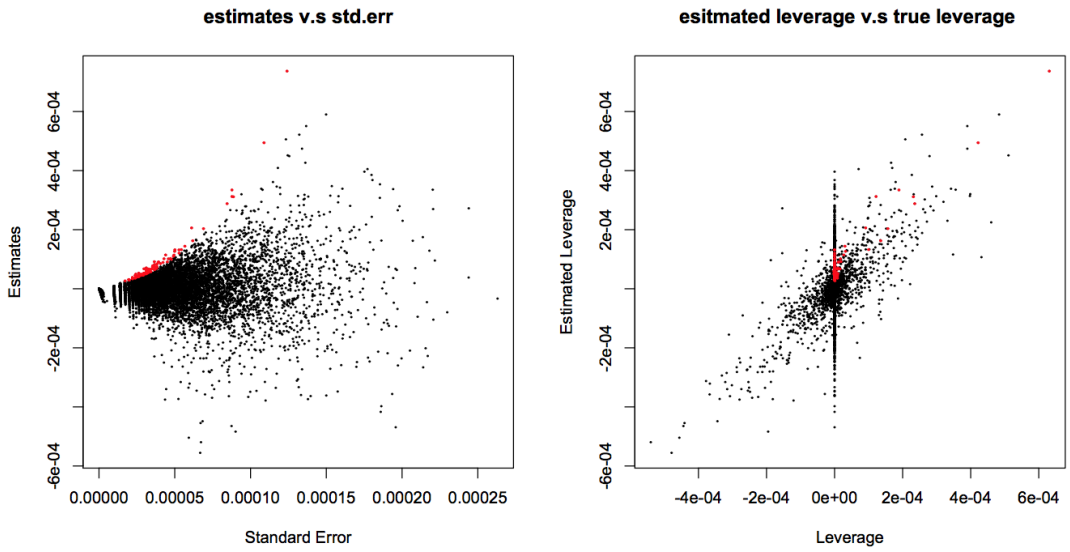

Figure 3.6: R-value: The red points are the top 1% association rules ranked by r-value ranking method using conjugate prior model.

its estimation of prior. Moreover using a normal prior has a problem that we stated in Section 2.4, that it prefers units with small standard error. We see that this problem has resulted in r-value ranking selecting many points with very small true leverage.

3.4.3 Local Maximum Likelihood (ML) Approach

Next we compare the local maximum likelihood (ML) approach. The local maximum likelihood approach estimates θ_i through the maximum likelihood estimate, which is $\hat{\theta}_i$ in our simulation. The MLE ranking method ranks units directly from estimates without considering standard error.

The left panel of Figure 3.7 is plot of estimated leverage against estimated standard error. Also the right panel shows estimated leverage against true leverage. Red points in both plot are top 1% units ranked by MLE. Even though the method prefers units with large error, it still gives an acceptable result in choosing the top 1% leverage values, as the left panel of Figure 3.7 shows. Note that our simulation gives ML ranking an advantage because ML ranking does not downweight observations with large standard error, and in our simulation, the observations with larger leverage have larger standard error.

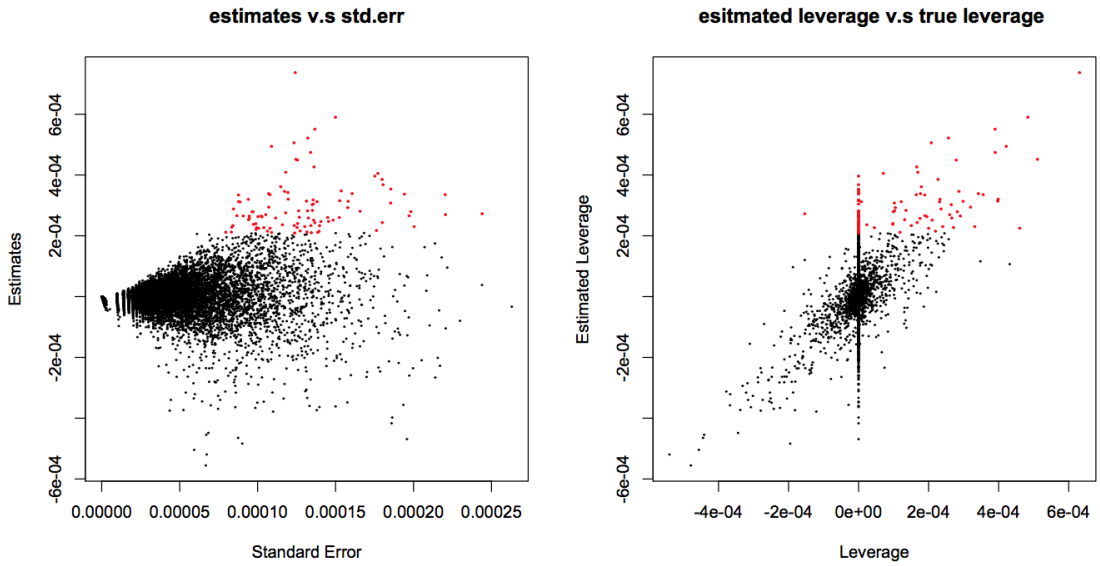

Figure 3.7: MLE: The red points are the top 1% association rules ranked by maximum likelihood (ML) ranking method.

3.4.4 Testing Approach

The testing approach is mentioned in Section 1.1. Here, we take the null hypothesis as $H_0 : \theta_i = 0$ for $i = 1, 2, \cdots, n$. Then our p-value with a normal likelihood $\mathcal{N}(\theta_i, \sigma_i^2)$ for each unit θ_i is

$$\begin{aligned} \text{p-value} =& \Pr(\text{observation} \geq \hat{\theta}_i | H_0 \text{ is true}) \\ =& \Pr(\text{estimates} \geq \hat{\theta}_i | \theta_i = 0) \\ =& 1 - \Phi\left(\frac{\hat{\theta}_i - 0}{\sigma_i}\right) \end{aligned} \tag{3.19}$$

Note that since the p-value is a decreasing function of $\frac{\hat{\theta}_i}{\sigma_i}$, p-value ranking in this case is equivalent to ranking by $\frac{\hat{\theta}_i}{\sigma_i}$. The result of the testing approach is presented in Figure 3.8. The left panel of this figure is a plot of estimated leverage against estimated standard error, while the right panel shows estimated leverage against true leverage. Red points in both plots are top 1% units ranked by p-value. Compared to MLE ranking, this method selects fewer associations with true leverage 0, but also misses some of the top associations because they have high standard errors.

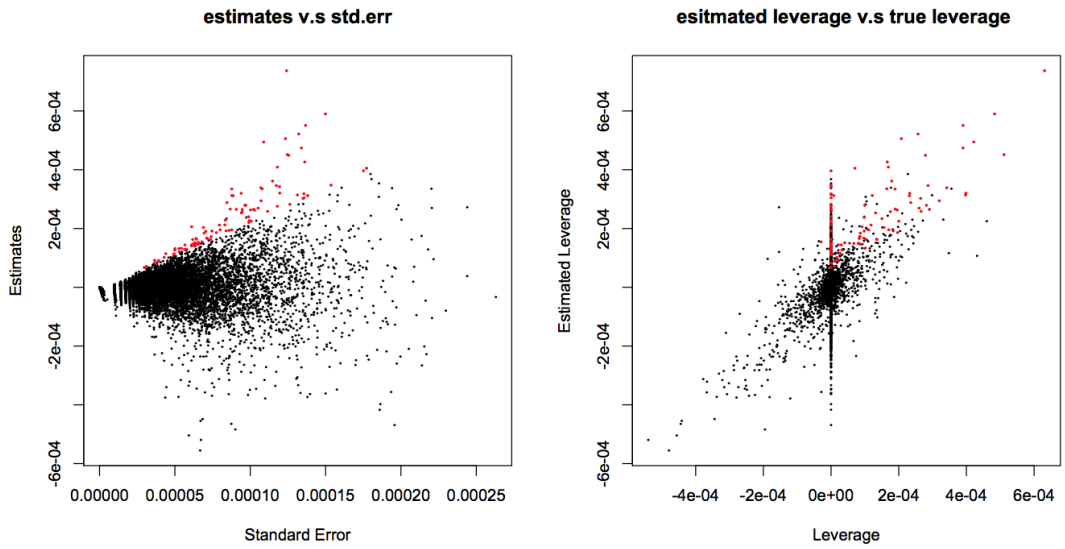

Figure 3.8: P-value: The red points are the top 1% association rules ranked by test approach (p-value) ranking method using normal model, and $H_0 : \theta_i = 0$.

3.4.5 Comparison of Different Methods

In the last part of this chapter, I compare the result of the different methods applied above.

Figure 3.9 provides us more details of how various methods rank the top units. Each panel of this figure corresponds to a different ranking method. The highest ranks of units are colored red. As the color changes from red to purple, the ranks of units vary from top 1% to top 30%. It is again shown in this figure that posterior mean with normal prior selects unit with relatively small standard errors, while MLE gives no penalty to observation variance in ranking units. Posterior mean with t prior gives a more desirable trade-off between standard error and estimates due to the fact that its choice of a heavy-tailed prior.

Figure 3.10 displays a plot of cumulative average of true leverage ($\bar{\theta}_\alpha$) against the percentage (α), with the cumulative average defined by equation (2.16). We explained in Section 2.4 why this is a good measure of performance. That is

$$\bar{\theta}_\alpha = \frac{1}{k} \sum_{j=m_{[1]}}^{m_{[k]}} \theta_j, \tag{3.20}$$

where $m_{[1]}, m_{[2]}, \cdots, m_{[k]}$ are the indices of units (θ_i's) ranked by different ranking

methods.

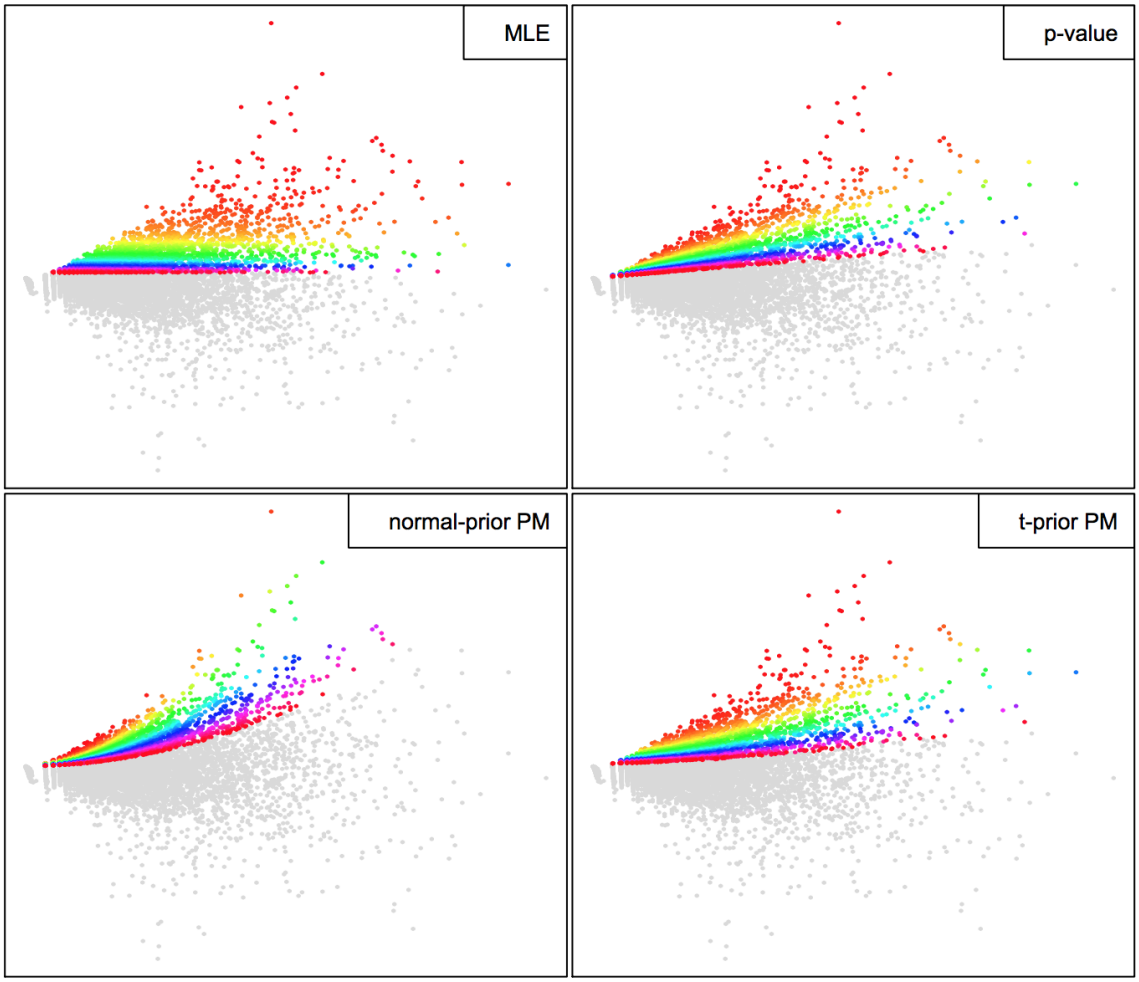

Figure 3.9: Comparisons of ranking via various methods. These plots show that the top 30% units ranked by MLE, p-value, posterior mean with normal prior and posterior mean with t-prior colored from red to purple.

Here the "Best choice" in the black dashed line is the cumulative average of the top θ ranked by true values. "unconditional PM" in the purple line is the cumulative average of top θ_i ranked by theoretical unconditional posterior mean, which is given by the hierarchical model, in equation (3.15) and equation (3.16). Consequently, the

unconditional posterior mean is calculated by a numerical integration of

$$\begin{aligned}\mathrm{E}[\theta_i|\hat{\theta}_i,\sigma_i^2] &= \frac{\iiint \theta_i\ p(\hat{\theta}_i|\theta_i,\sigma_i^2)\ p(\theta_i|\tau_i^2(a,b))\ p_1(a)\ p_2(b)\,da\,db\,d\theta_i}{\iiint\ p(\hat{\theta}_i|\theta_i,\sigma_i^2)\ p(\theta_i|\tau_i^2(a,b))\ p_1(a)\ p_2(b)\,da\,db\,d\theta_i} \\ &= \frac{\int_0^1\int_0^1\int_{-1}^1 \theta_i \frac{(1-a)^{49}(1-b)^{49}}{ab}\, e^{-\frac{(\hat{\theta}_i-\theta_i)^2}{2\sigma_i^2}-\frac{\theta_i^2}{2(0.2ab)^2}}\,da\,db\,d\theta_i}{\int_0^1\int_0^1\int_{-1}^1 \frac{(1-a)^{49}(1-b)^{49}}{ab}\, e^{-\frac{(\hat{\theta}_i-\theta_i)^2}{2\sigma_i^2}-\frac{\theta_i^2}{2(0.2ab)^2}}\,da\,db\,d\theta_i}\end{aligned} \tag{3.21}$$

This should give the best estimation of θ_i by posterior mean, that is the theoretically best ranks we can achieve by using posterior mean methods, since the prior is known here. It should give an upper bond on how well we could do using posterior mean. The gap between "unconditional PM" and "Best choice" is the irreducible error, that is the part of the error in ranking due to noise in the data.

"t-distribution prior PM" in the red line is the cumulative average of top θ_i ranked by posterior mean using a Student's t-distribution as prior, while "conjugate prior PM" in the orange line is the cumulative average of top θ_i ranked by posterior mean using normal prior. "r-value" in the green line is the cumulative average of top θ_i ranked by r-value with a conjugate normal prior. "MLE" and "p-value" are cumulative average of top θ_i ranked by MLE and p-value.

Figure 3.10 gives several results of comparing different ranking and selection methods. It is very close between the red line, green line and blue line in Figure 3.10, which means the result of ranking using MLE, p-value and t/normal posterior mean are similar. Still the red line lies above the others. We see that there is still a small gap between the posterior mean with t-prior and the unconditional posterior mean. This indicates that further work on selecting a prior and estimating the hyperparameters could improve the results. The other empirical Bayes methods with normal prior, including both r-value and posterior mean give an inadequate ranking.

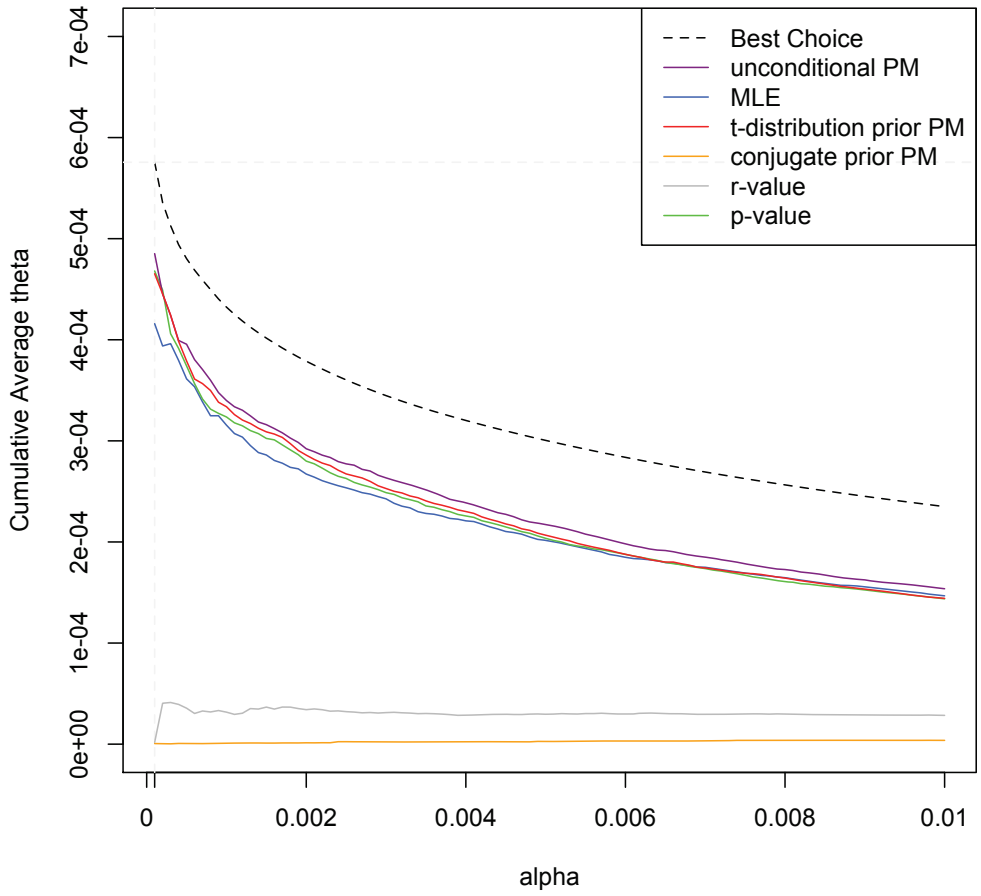

Figure 3.10: Cumulative average plot. This plot shows the cumulative average of leverage, $\bar{\theta}_\alpha$ over 20 sets of simulated market basket data using different ranking methods.

Chapter 4

Conclusion and Future Work

In this thesis, we have studied the effect of choice of priors on empirical Bayes posterior mean ranking methods. In a simulation study, we compared the performances of a light-tailed normal prior, a medium-tailed gamma prior and a heavy-tailed Student's t-prior. As expected, the true prior performs best in each case. However, we found that the ranking method using the normal prior when the true prior has a heavy tail performs far worse than using the t-prior when the true prior has a lighter tail. We conclude that using a heavier tailed prior is more robust to misspecification of the prior.

In Chapter 3, we applied different ranking approaches to the simulated market basket data. As was shown in ranking the leverage, empirical Bayes approaches with normal prior produce very poor results, while posterior mean ranking method with a Student's t prior outperforms the other methods.

Although, we have mainly demonstrated the benefits of Student's t-prior in posterior mean ranking method, we expect the use of a heavy-tailed prior to be beneficial for other empirical Bayes approaches. We have only studied one example each of light-tailed, medium-tailed and heavy-tailed distributions. Further work is needed to determine whether another heavy-tailed distribution might perform better. Since there isn't an explicit form of posterior distribution by using such prior, it requires more computation to calculate the posterior distribution. Consequently, to choose another heavy-tail prior is worthwhile to be done in the future. Also, use of a heavy-tailed prior for empirical Bayes methods should be applied to a real data sets, to confirm that the benefits observed in simulations are actually achieved in practice.

Further work is also needed to determine the extent to which a heavy-tailed prior gains robustness to prior misspecification at the expense of efficiency. This will allows us to find detailed recommendations regarding how heavy-tailed the prior distribution should be.

We also briefly examined the effect of prior parameter estimates on posterior mean ranking. We found that the true values do not always give the best result, particularly in cases where the prior is misspecified. In these cases, we found that overestimating the variance can be advantageous. More work is needed on this topic to determine how best to estimate parameters in the prior distribution for empirical Bayes ranking methods.

Bibliography

[1] Robert E Bechhofer. A single-sample multiple decision procedure for ranking means of normal populations with known variances. *The Annals of Mathematical Statistics*, pages 16–39, 1954.

[2] Shanti Swarup Gupta. *On a decision rule for a problem in ranking means*. PhD thesis, University of North Carolina at Chapel Hill, 1956.

[3] Robert Eric Bechhofer, Jack Kiefer, and Milton Sobel. *Sequential identification and ranking procedures: with special reference to Koopman-Darmois populations*, volume 3. University of Chicago Press, 1968.

[4] Jean Dickinson Gibbons and Subhabrata Chakraborti. *Nonparametric statistical inference*. Springer, 2011.

[5] Jean D Gibbons, Ingram Olkin, and Milton Sobel. An introduction to ranking and selection. *The American Statistician*, 33(4):185–195, 1979.

[6] Murray Aitkin and Nicholas Longford. Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society. Series A (General)*, pages 1–43, 1986.

[7] N. C. Henderson and M. A. Newton. Making the cut: improved ranking and selection for large-scale inference. *ArXiv e-prints*, December 2013.

[8] Nan M Laird and Thomas A Louis. Empirical bayes ranking methods. *Journal of Educational and Behavioral Statistics*, 14(1):29–46, 1989.

[9] James O Berger and John Deely. A bayesian approach to ranking and selection of related means with alternatives to analysis-of-variance methodology. *Journal of the American Statistical Association*, 83(402):364–373, 1988.

[10] Shanti S Gupta and Ping Hsiao. Empirical bayes rules for selecting good populations. *Journal of Statistical Planning and Inference*, 8(1):87–101, 1983.

[11] Shanti S Gupta and TaChen Liang. Empirical bayes rules for selecting good binomial populations. *Lecture Notes-Monograph Series*, pages 110–128, 1986.

[12] Rongheng Lin, Thomas A Louis, Susan M Paddock, and Greg Ridgeway. Loss function based ranking in two-stage, hierarchical models. *Bayesian Analysis (Online)*, 1(4):915, 2006.

[13] Hisashi Noma, Shigeyuki Matsui, Takashi Omori, and Tosiya Sato. Bayesian ranking and selection methods using hierarchical mixture models in microarray studies. *Biostatistics*, page kxp047, 2009.

[14] Wei Shen and Thomas A Louis. Triple-goal estimates in two-stage hierarchical models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2):455–471, 1998.

[15] James O Berger and Thomas Sellke. Testing a point null hypothesis: the irreconcilability of p values and evidence. *Journal of the American statistical Association*, 82(397):112–122, 1987.

[16] Tom Brijs, Dimitris Karlis, Filip Van den Bossche, and Geert Wets. A bayesian model for ranking hazardous road sites. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(4):1001–1017, 2007.

[17] Peter Hall and Hugh Miller. Modeling the variability of rankings. *The Annals of Statistics*, pages 2652–2677, 2010.

[18] Gustavo de los Campos, John M Hickey, Ricardo Pong-Wong, Hans D Daetwyler, and Mario PL Calus. Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics*, 193(2):327–345, 2013.

[19] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. Introduction to data mining, (first edition). chapter 6 Association Analysis: Basic Concepts and Algorithms, pages 327–414. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2005.

[20] Snowplow Analytics Limited. Market basket analysis: identifying products and content that go well together.

[21] Gregory Piatetsky-Shapiro. Discovery, analysis, and presentation of strong rules. *Knowledge discovery in databases*, pages 229–238, 1991.

[22] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC Texts in Statistical Science. Chapman & Hall/CRC, 2003.

[23] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. The elements of statistical learning. volume 1, chapter 14 Unsupervised Learning, pages 485–586. Springer series in statistics Springer, Berlin, 2001.

[24] Rakesh Agrawal, Heikki Mannila, Ramakrishnan Srikant, Hannu Toivonen, A Inkeri Verkamo, et al. Fast discovery of association rules. *Advances in knowledge discovery and data mining*, 12(1):307–328, 1996.

[25] Rupert G Miller Jr. Developments in multiple comparisons 1966–1976. *Journal of the American Statistical Association*, 72(360a):779–788, 1977.