1-1-2013

Bayesian Aspects of Classification Procedures

Igar Fuki
University of Pennsylvania, igarfuki@wharton.upenn.edu

Bayesian Aspects of Classification Procedures

Abstract
We consider several statistical approaches to binary classification and multiple hypothesis testing problems. Situations in which a binary choice must be made are common in science. Usually, there is uncertainty involved in making the choice and a great number of statistical techniques have been put forth to help researchers deal with this uncertainty in separating signal from noise in reasonable ways. For example, in genetic studies, one may want to identify genes that affect a certain biological process from among a larger set of genes. In such examples, costs are attached to making incorrect choices and many choices must be made at the same time. Reasonable ways of modeling the cost structure and choosing the appropriate criteria for evaluating the performance of statistical techniques are needed. The following three chapters have proposals of some Bayesian methods for these issues.

In the first chapter, we focus on an empirical Bayes approach to a popular binary classification problem formulation. In this framework, observations are treated as independent draws from a hierarchical model with a mixture prior distribution. The mixture prior combines prior distributions for the ``noise'' and for the ``signal'' observations. In the literature, parametric assumptions are usually made about the prior distribution from which the ``signal'' observations come. We suggest a Bayes classification rule which minimizes the expectation of a flexible and easily interpretable mixture loss function which brings together constant penalties for false positive misclassifications and L_2 penalties for false negative misclassifications. Due in part to the form of the loss function, empirical Bayes techniques can then be used to construct the Bayes classification rule without specifying the ``signal'' part of the mixture prior distribution. The proposed classification technique builds directly on the nonparametric mixture prior approach proposed by Raykar and Zhao (2010, 2011).

Many different criteria can be used to judge the success of a classification procedure. A very useful criterion called the False Discovery Rate (FDR) was introduced by Benjamini and Hochberg in a 1995 paper. For many applications, the FDR, which is defined as the expected proportion of false positive results among the observations declared to be ``signal'', is a reasonable criterion to target. Bayesian versions of the false discovery rate, the so-called positive false discovery rate (pFDR) and local false discovery rate, were proposed by Storey (2002, 2003) and Efron and coauthors (2001), respectively. There is an interesting connection between the local false discovery rate and the nonparametric mixture prior approach for binary classification problems. The second part of the dissertation is focused on this link and provides a comparison of various approaches for estimating Bayesian false discovery rates.

The third chapter is an account of a connection between the celebrated Neyman-Pearson lemma and the area (AUC) under the receiver operating characteristic (ROC) curve when the observations that need to be classified come from a pair of normal distributions. Using this connection, it is possible to derive a classification rule which maximizes the AUC for binormal data.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

BAYESIAN ASPECTS OF CLASSIFICATION PROCEDURES

Igar Fuki

A DISSERTATION
in
Statistics

For the Graduate Group in Managerial Science and Applied Economics
Presented to the Faculties of the University of Pennsylvania
in
Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy
2013

Supervisor of Dissertation

_____

Linda Zhao, Professor, Statistics

Graduate Group Chairperson

_____

Eric Bradlow, K.P. Chao Professor, Marketing, Statistics and Education

Dissertation Committee
Linda Zhao, Professor of Statistics
Lawrence D. Brown, Miers Busch Professor, Statistics
Fernando Ferreira, Associate Professor of Real Estate

BAYESIAN ASPECTS OF CLASSIFICATION PROCEDURES

To my family.

Acknowledgement

This dissertation and all of my projects in graduate school would not have been possible without Linda Zhao. I would like to thank her for her incredible help and support. Professor Zhao not only set out my research path, but was a source of strength during the most difficult times. Thank you for everything.

I would like to thank my committee members, Lawrence Brown and Fernando Ferreira, for their help and invaluable advice with my dissertation. I would also like to express my deepest gratitude to Marja Hoek-Smit, whose door was always open for me.

Thank you to Todd Sinai, Warren Ewens, Murray Gerstenhaber, Carol Reich, Keith Weigelt, Daniel Yekutieli, and W. Bruce Allen for their guidance and advice. My thanks also to Vikas Raykar and Xu Han.

This thesis is based on several joint working papers, and I would like to thank my senior co-authors for their contributions to the text, for their inspiration, and for teaching me about research and scientific writing.

Thank you also to the faculty, staff, and students of the Statistics and Real Estate departments at Wharton, and to many others at the University of Pennsylvania for all of their help.

Finally, I want to thank my friends and family for their patience and kind support.

ABSTRACT

BAYESIAN ASPECTS OF CLASSIFICATION PROCEDURES

Igar Fuki

Linda Zhao

We consider several statistical approaches to binary classification and multiple hypothesis testing problems. Situations in which a binary choice must be made are common in science. Usually, there is uncertainty involved in making the choice and a great number of statistical techniques have been put forth to help researchers deal with this uncertainty in separating signal from noise in reasonable ways. For example, in genetic studies, one may want to identify genes that affect a certain biological process from among a larger set of genes. In such examples, costs are attached to making incorrect choices and many choices must be made at the same time. Reasonable ways of modeling the cost structure and choosing the appropriate criteria for evaluating the performance of statistical techniques are needed. The following three chapters have proposals of some Bayesian methods for these issues.

In the first chapter, we focus on an empirical Bayes approach to a popular binary classification problem formulation. In this framework, observations are treated as independent draws from a hierarchical model with a mixture prior distribution.

The mixture prior combines prior distributions for the "noise" and for the "signal" observations. In the literature, parametric assumptions are usually made about the prior distribution from which the "signal" observations come. We suggest a Bayes classification rule which minimizes the expectation of a flexible and easily interpretable mixture loss function which brings together constant penalties for false positive misclassifications and L_2 penalties for false negative misclassifications. Due in part to the form of the loss function, empirical Bayes techniques can then be used to construct the Bayes classification rule without specifying the "signal" part of the mixture prior distribution. The proposed classification technique builds directly on the nonparametric mixture prior approach proposed by Raykar and Zhao (2010, 2011).

Many different criteria can be used to judge the success of a classification procedure. A very useful criterion called the False Discovery Rate (FDR) was introduced by Benjamini and Hochberg in a 1995 paper. For many applications, the FDR, which is defined as the expected proportion of false positive results among the observations declared to be "signal", is a reasonable criterion to target. Bayesian versions of the false discovery rate, the so-called positive false discovery rate (pFDR) and local false discovery rate, were proposed by Storey (2002, 2003) and Efron and coauthors (2001), respectively. There is an interesting connection between the local false discovery rate and the nonparametric mixture prior approach for binary classification problems. The second part of the dissertation is focused on this link and provides a comparison of various approaches for estimating Bayesian false discovery rates.

The third chapter is an account of a connection between the celebrated Neyman-

Pearson lemma and the area (AUC) under the receiver operating characteristic (ROC) curve when the observations that need to be classified come from a pair of normal distributions. Using this connection, it is possible to derive a classification rule which maximizes the AUC for binormal data.

Contents

List of Tables

List of Figures

Chapter 1

A Nonparametric Bayesian Classifier under a Mixture Loss Function

1.1 Introduction

The problem of separating "signal" from "noise" is fundamental to many scientific applications. In formulating a concise model for a natural phenomenon, one attempts to identify relevant features ("signal") and separate them from the less relevant ones ("noise"). In biology, for example, one is often interested in finding genes that are responsible for certain traits in an organism. In such an application, the researcher may begin by examining hundreds or thousands of candidate genes in an effort to identify a much smaller subset of genes that are the most relevant to

a particular biological mechanism. One might need to make hundreds or thousands of classification decisions simultaneously and a classification rule that deals with the large amount of data in a reasonable way can therefore be quite useful.

In this chapter, an empirical Bayes approach to classification problems is considered. In general, the empirical Bayes approach can be described using a hierarchical framework. In this framework, a sample of unseen values $\theta_1, ..., \theta_n$ is drawn from an unknown prior distribution $\gamma(\theta)$. A sample of observations $Z_1, ..., Z_n$ is then drawn, with each observation Z_j coming from the distribution $f_{\theta_j}(z)$, which belongs to the known probability family $f_\theta(z)$. As noted by Efron (2013), the empirical Bayes literature can loosely be divided into two parts. One part of the research has focused on results which rely on estimating the distribution $f_\theta(z)$, and the other part on estimating the prior distribution $\gamma(\theta)$. For example, work based on the classical James-Stein estimator is directly connected to empirical Bayes approaches (for a discussion of the connections, see, for example, Efron and Morris, 1975) and can be classified in the first category. Other work, such as Zhang (1997), has focused on problems that require better estimation of the prior distribution $\gamma(\theta)$. A very accessible review of the literature and of various empirical Bayes techniques is provided by Efron (2013).

Empirical Bayes techniques can be used in an intuitive way to attack classification problems. In this chapter, a nonparametric Bayes classification rule aimed at minimizing a highly interpretable risk function is proposed. Its performance is compared to that of a parametric classifier in simulations for various signal distributions, sparsity regimes, and signal strengths. When the prior distribution is misspecified

for the parametric classifier, the nonparametric classification rule performs better in terms of emprical risk. Reassuringly, even when the prior distribution assumptions are correct, the nonparametric classifier is seen to have comparable performance to its parametric counterpart.

In the next section, a commonly used model for the classification context is described. An intuitive loss function is then introduced and the problem is cast in a Bayesian framework.

1.2 The Model

In this section, a commonly used classification model is described. This is a model with n observations of the form

$$z_i = \theta_i + \epsilon_i, \tag{1.2.1}$$

where $i = 1, ..., n$ indexes the observations and the ϵ_i's are independent and normally distributed with mean 0 and constant variance σ^2 (using the notation $N(\epsilon_i|0, \sigma^2)$ to denote this). Without loss of generality, $\sigma^2 = 1$ is set for the remainder of the chapter.

In this setup, $\mathbf{z} = (z_1, z_2, ..., z_n)^T$ constitutes a vector of observations,

$$\theta = (\theta_1, \theta_2, ..., \theta_n)^T$$

is an unobserved vector of corresponding means, and $\epsilon = (\epsilon_1, \epsilon_2, ..., \epsilon_n)^T$ is an unob-

served additive random error vector. With the applications described in the Introduction in mind, it is assumed that a large proportion of the θ_i's may be equal to zero. For the classification problem, the goal is to decide which $\theta_i = 0$ (corresponding to "noise") and which $\theta_i \neq 0$ (corresponding to "signal"). More formally, given the data $\mathbf{z}$, one would like to provide an n-dimensional decision vector $\mathbf{a} = (a_1, a_2, ..., a_n)^T$, where

$$a_i = \begin{cases} 0, & \text{for deciding that } \theta_i = 0 \\ \\ 1, & \text{for deciding that } \theta_i \neq 0. \end{cases}$$

In other words, declaring $a_i = 0$ corresponds to a decision that $\theta_i = 0$ and declaring $a_i = 1$ corresponds to a decision that $\theta_i \neq 0$. As is usual for such models, it is assumed that with each incorrect classification decision, the researcher incurs some cost. Given a particular form for the cost structure, the goal is to select a decision vector $\mathbf{a}$ that makes the overall cost small. This is formalized using the standard decision-theoretic loss function framework. A highly interpretable loss function for the classifier is described in the next section.

As discussed in the next section, the selected loss function has two main appealing features. First, one can argue that it is well-motivated from the standpoint of typical applications, such as the biological microarray framework. In such applications, it seems reasonable to assume that false positives and false negatives do not carry equal weight, and should therefore be penalized differently. This loss function also allows the researcher to get an estimate of a Bayes rule without having to specify an explicit prior distribution $\gamma(\theta)$.

1.3 A Highly Interpretable Loss Function

For a particular value of θ_i, let $L(a_i, \theta_i)$ be the loss incurred from making the decision a_i. In what follows, we will assume that the total loss incurred is additive; that is, we assume that the total loss $TL(\mathbf{a}, \theta)$ from selecting a decision vector $\mathbf{a}$ for classifying the observations $\mathbf{z}$ is

$$TL(\mathbf{a}, \theta) = \sum_{i=1}^{n} L(a_i, \theta_i). \tag{1.3.1}$$

We use the following penalty structure for each classification decision:

$$L(1, \theta_i) = \begin{cases} 0 & \text{if} \quad \theta_i \neq 0 \\ 1 & \text{if} \quad \theta_i = 0 \end{cases}$$

and

$$L(0, \theta_i) = \begin{cases} 0 & \text{if} \quad \theta_i = 0 \\ c\theta_i^2 & \text{if} \quad \theta_i \neq 0 \end{cases} \tag{1.3.2}$$

That is, the cost of saying that $\theta_i \neq 0$ when it is in fact equal to 0 is constant (and normalized to be 1). On the other hand, the cost of saying that $\theta_i = 0$ when it is non-zero is proportional to the square of its magnitude. In the genetic array framework, this idealized cost structure can be interpreted as putting a fixed cost for each subsequent experiment performed to sequence genes that were called "differentially expressed" ($\theta_i \neq 0$) in the initial screening step and costs proportional to the magnitude of the differential expression for failing to make a discovery. The

total loss under this structure is given by

$$TL(\mathbf{a}, \theta) = \sum_{i=1}^{n} \left(a_i \, \mathbf{1}_{\theta_i = 0} + (1 - a_i) \, c \, \theta_i^2 \right), \tag{1.3.3}$$

where a_i corresponds to the classification decision for the i^{th} observation and $\mathbf{1}_{\theta_i}$ is an indicator variable which equals one when $\theta_i = 0$ and equals zero when $\theta_i \neq 0$. In Section 3, we use the total loss function (1.3.3) to evaluate the performance of two classifiers.

1.3.1 Bayes Rules

A vast literature covers various aspects of the model in expression (1.2.1) in the context of microarray analysis, signal processing, statistical model selection, machine learning, and other fields. In Bayesian approaches to this problem, one places prior distributions on parameters of interest in the model and computes various posterior distribution quantities based on the observed data. For a particular prior distribution structure, a classification rule which minimizes the expected loss is called a Bayes rule. In the next section, we focus on a Bayesian approach that relies on mixture prior distributions and formulate a Bayes rule for the loss structure in (1.3.2).

1.4 A Bayesian Apporach Based on Mixture Priors

1.4.1 The Mixture Prior Formulation

A sensible Bayesian approach for treating the model (1.2.1) is to place a mixture prior distribution of the form

$$p(\theta_i|\omega,\gamma) = \omega\delta(\theta_i) + (1-\omega)\gamma(\theta_i) \tag{1.4.1}$$

on the θ_i's and to compute posterior probabilities for $\theta_i = 0$ and $\theta_i \neq 0$. In this parametrization of the mixture, ω is the weight placed on an atom of probability at 0 and γ is a density function from which the non-zero θ_i's are thought to come. The prior distribution on θ_i is thus a weighted mixture of a delta function, which places an atom of mass at 0, and some density γ.

Because of independence, the likelihood function of the observations $\mathbf{z} = (z_1, z_2, ..., z_n)$ given the parameters $\theta = (\theta_1, \theta_2, ..., \theta_n)$ can be factored as

$$p(\mathbf{z}|\theta) = \prod_{i=1}^{n} p(z_i|\theta_i) = \prod_{i=1}^{n} N(z_i|\theta_i, 1). \tag{1.4.2}$$

The posterior distribution of θ given ω and γ is given by

$$p(\theta|\mathbf{z},\omega,\gamma) = \frac{\prod_{i=1}^{n} p(z_i|\theta_i)p(\theta_i|\omega,\gamma)}{m(\mathbf{z}|\omega,\gamma)}, \tag{1.4.3}$$

where

$$m(\mathbf{z}|\omega,\gamma) = \prod_{i=1}^{n} \int p(z_i|\theta_i)p(\theta_i|\omega,\gamma)d\theta_i \tag{1.4.4}$$

is the marginal distribution of the data given the hyperparameters. For the likelihood in (1.4.2) and the mixture prior in (1.4.1), the integral in (1.4.4) can be rewritten as

$$\begin{aligned} &\int p(z_i|\theta_i,1)p(\theta_i|\omega,\gamma)d\theta_i \\ &= \omega N(z_i|0,1) + (1-\omega)g(z_i), \end{aligned} \tag{1.4.5}$$

where

$$g(z_i) = \int N(\theta_i|z_i,1)\gamma(\theta_i)d\theta_i. \tag{1.4.6}$$

Here, g is the marginal density of z_i given that θ_i is non-zero. The posterior in (1.4.3) can then be factored as $p(\theta|\mathbf{z},\omega,\gamma) = \prod_{i=1}^{n} p(\theta_i|z_i,\omega,\gamma)$, with

$$\begin{aligned} & p(\theta_i|z_i,\omega,\gamma) \\ = \; & \frac{\omega\delta(\theta_i)N(z_i|0,1) + (1-\omega)\gamma(\theta_i)N(z_i|\theta_i,1)}{\omega N(z_i|0,1) + (1-\omega)g(z_i)} \\ = \; & p_i\delta(\theta_i) + (1-p_i)G(\theta_i), \end{aligned} \tag{1.4.7}$$

where

$$p_i = p(\theta_i = 0|z_i,\omega,\gamma) = \frac{\omega N(z_i|0,1)}{\omega N(z_i|0,1) + (1-\omega)g(z_i)} \tag{1.4.8}$$

is the posterior probability that $\theta_i = 0$ and

$$G(\theta_i) = \frac{N(\theta_i|z_i,1)\gamma(\theta_i)}{\int N(\theta_i|z_i,1)\gamma(\theta_i)d\theta_i} \tag{1.4.9}$$

is the posterior density of θ_i when $\theta_i \neq 0$.

1.4.2 A Bayes Rule

Under the mixture prior distribution in (1.4.1) and the loss structure in (1.3.2), it is easy to find a decision procedure that minimizes the expectation of the total loss (1.3.3) for classifying data from the model (1.2.1). The i^{th} component of the n-dimensional Bayes classification rule for this setup can be written in terms of the posterior probability p_i that $\theta_i = 0$ and the second moment of θ_i under the posterior distribution $G(\theta_i)$:

Proposition 1: Denoting the i^{th} component of the Bayes rule by a_i^{Bayes} and the second moment of θ_i under $G(\theta_i)$ by $E_G[\theta_i^2]$, the rule is

$$a_i^{Bayes} = \begin{cases} 1, & if\ p_i < \frac{cE_G[\theta_i^2]}{1+cE_G[\theta_i^2]} \\ \\ 0, & otherwise, \end{cases} \tag{1.4.10}$$

where, again, c is the cost constant from expression (1.3.2).

In other words, the Bayes rule is to decide that $\theta_i \neq 0$ if and only if the posterior probability p_i that $\theta_i = 0$ is below a certain threshold.

To see why (1.4.10) minimizes the expected loss, note that the expectation of the total loss (1.3.3) can be minimized component-wise. The i^{th} component of the Bayes rule is to decide that $\theta_i \neq 0$ precisely when

$$E(L(1, \theta_i)) < E(L(0, \theta_i)), \tag{1.4.11}$$

where $E(L(a_i, \theta_i))$ stands for the expectation of the loss from the decision a_i when the parameter is θ_i. These component-wise expected losses are given by

$$E(L(1, \theta_i)) = \int L(1, \theta_i)\pi(\theta_i|\ data)d\theta_i = p_i$$

and

$$E(L(0, \theta_i)) = \int L(0, \theta_i)\pi(\theta_i|\ data)d\theta_i$$

$$= c(1 - p_i) \int \theta_i^2 G(\theta_i)d\theta_i\ \ = c(1 - p_i)E_G[\theta_i^2].$$

Based on these expressions, we arrive at the Bayes rule in (1.4.10).

Under mild conditions (see (Brown, 1971)), the classification rule in (1.4.10) can be rewritten in a form that is particularly useful for estimation. The rule can be written in terms of the observations as

$$a_i^{Bayes} = \begin{cases} 1, & \text{if } p_i < \frac{c\left(\frac{g''(z_i)}{g(z_i)} + z_i^2 + 2z_i\frac{g'(z_i)}{g(z_i)} + 1\right)}{1 + c\left(\frac{g''(z_i)}{g(z_i)} + z_i^2 + 2z_i\frac{g'(z_i)}{g(z_i)} + 1\right)} \\ \\ 0, & \text{otherwise,} \end{cases} \tag{1.4.12}$$

where c is the cost constant from expression (1.3.2), z_i is the i^{th} observation, g is the marginal density function of z_i given that $\theta_i \neq 0$, and g' and g'' are its first two derivatives.

In expression (1.4.8), the posterior probability p_i is defined in terms of the marginal density g. The form of the function g is determined by the prior distribution γ. In the remainder of this section, we discuss the choice of γ.

1.4.3 Parametric Prior γ

Typically, in this context, γ is taken to be a parametric distribution. One common choice for the prior distribution $\gamma(\theta_i)$ on the non-zero θ_i's is the normal distribution $N(\theta_i|\theta, \tau^2)$. The marginal density g is then determined analytically and the threshold in (1.4.12) can be estimated using empirical Bayes techniques.

As shown in (1.4.12), the Bayes classification rule for the loss function in (1.3.2) may be written in terms of the marginal density g of z_i given that θ_i is non-zero. For the case of the normal prior $\gamma(\theta_i) = N(\theta_i|\theta, \tau^2)$, equation (1.4.6) for the marginal density g and equation (1.4.8) for the posterior probability p_i become, respectively,

$$g(z_i) = \int N(z_i|\theta_i, 1)N(\theta, \tau^2)d\theta_i = N(z_i|\theta, 1 + \tau^2) \tag{1.4.13}$$

and

$$p_i = \frac{\omega N(z_i|0, 1)}{\omega N(z_i|0, 1) + (1 - \omega)N(z_i|\theta, 1 + \tau^2)}. \tag{1.4.14}$$

A fully Bayesian treatment in which prior distributions are also placed on the posterior probability that θ_i is non-zero and on the proportion ω of non-zero θ_i's is preferable if the prior distribution γ is specified accurately (see (Scott and Berger, 2006)). In practice, the true shape of the distribution of θ_i is typically unknown and, often, the hyperparameters θ and τ^2 are instead estimated using empirical Bayes techniques. For our normal prior-based classifier, we iteratively maximize the marginal likelihood of the data in terms of each parameter while holding the other parameters fixed and repeat until the algorithm converges. The Bayes classifica-

tion rule from (1.4.12) is approximated using plug-in estimates and empirical risk calculations are provided in Section 5.

1.4.4 Nonparametric Prior γ

In contrast to the rigid assumptions of the parametric prior distribution approach, no explicit functional form is assumed for γ in our nonparametric classification method. Instead, we estimate the components of the Bayes rule threshold in equation (1.4.12) nonparametrically through an iterative Expectation-Maximization (EM)-style procedure suggested by Raykar and Zhao, 2010. Our estimates for the marginal density g, as well as for its derivatives g' and g'', rely on a kernel density estimator function K with bandwidth h. For our simulations, we use K equal to the normal density function with mean zero and unit variance ($K(x) = N(x|0,1)$). The bandwidth for the kernel is set using the normal reference rule (Wand and Jones, 1995) to $h = O(n^{-1/5})$. The algorithm for constructing our nonparametric classification begins by iterating the following two steps until convergence:

1. Compute an estimate of the posterior probability $\hat{p_i}$ using the current estimate $\hat{\omega}$ of the proportion of non-zero θ_i's and the current estimate $\hat{g}(z_i)$ of the marginal density corresponding to non-zero θ_i's:

$$\hat{p_i} = \frac{\hat{\omega} N(z_i|0,1)}{\hat{\omega} N(z_i|0,1) + (1-\hat{\omega})\hat{g}(z_i)} \tag{1.4.15}$$

2. Re-estimate $\hat{\omega}$ and $\hat{g}(z_i)$, as well as $\hat{g}'(z_i)$ and $\hat{g''}(z_i)$, using the current estimates

$\hat{p_i}$:

$$\hat{\omega} = \frac{1}{n}\sum_{i=1}^{n}\hat{p}_i, \tag{1.4.16}$$

$$\hat{g}(z_i) = \frac{1}{\tilde{p}h}\sum_{j=1}^{n}(1-\hat{p}_j)\,K\left(\frac{z_i - z_j}{h}\right), \tag{1.4.17}$$

$$\hat{g}'(z_i) = \frac{1}{\tilde{p}h^2}\sum_{j=1}^{m}(1-\hat{p}_j)\left(-\frac{z_i - z_j}{h}\right)K\left(\frac{z_i - z_j}{h}\right), \tag{1.4.18}$$

$$\hat{g}''(z_i) = \frac{1}{\tilde{p}h^3}\sum_{j=1}^{m}(1-\hat{p}_j)\left(\left(\frac{z_i - z_j}{h}\right)^2 - 1\right)K\left(\frac{z_i - z_j}{h}\right), \tag{1.4.19}$$

where $\tilde{p} = \sum_{j=1}^{n}(1-\hat{p}_j)$, K is the kernel density function, and h is its bandwidth. Note that estimates for $\hat{g'}$ and $\hat{g''}$ do not play a role in re-estimating $\hat{p}_i$ in step 1 and may be computed once at the end.

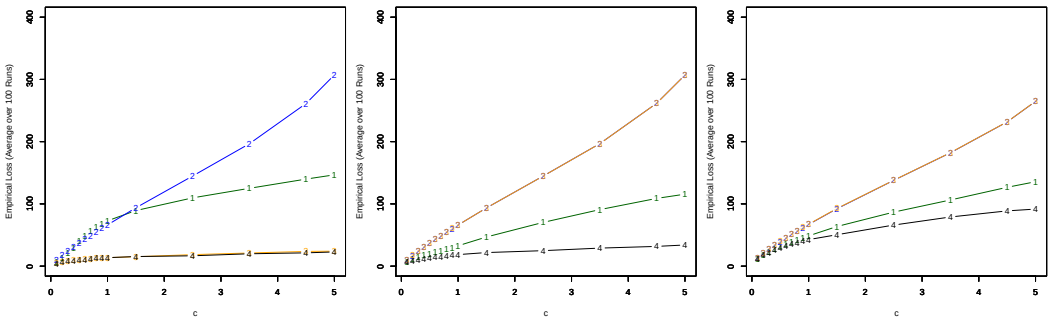
Once the algorithm converges, our nonparametric classifier for a particular value of the cost constant c from the loss function in (1.3.2) is constructed by plugging these estimates of p_i, ω, $g(z_i)$, $g'(z_i)$, and $g''(z_i)$ into the Bayes rule formulation of equation (1.4.12). In the next section, we compare the performance of the nonparametric and the normal prior-based classifiers in terms of average loss on simulated data for various values of c.

1.5 Simulations

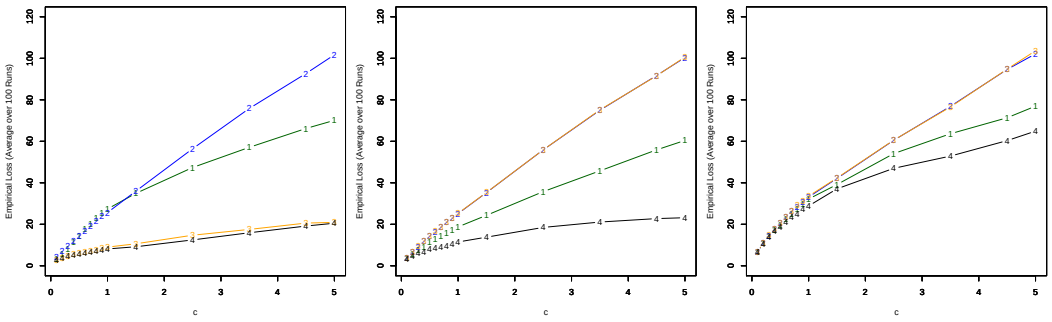
For each simulation run, we generate 100 samples of 500 observations each from a model of the form in equation (1.2.1) and come up with decision vectors $\mathbf{a}$ using different classification rules. We then compare the performance of several classifi-

cation methods in terms of the average of the loss in expression (1.3.3). To test the classification rules under various conditions, each simulated set of 100 samples comes from a model with varying sparsity, proportion, and generating distribution for the non-zero θ_i's.

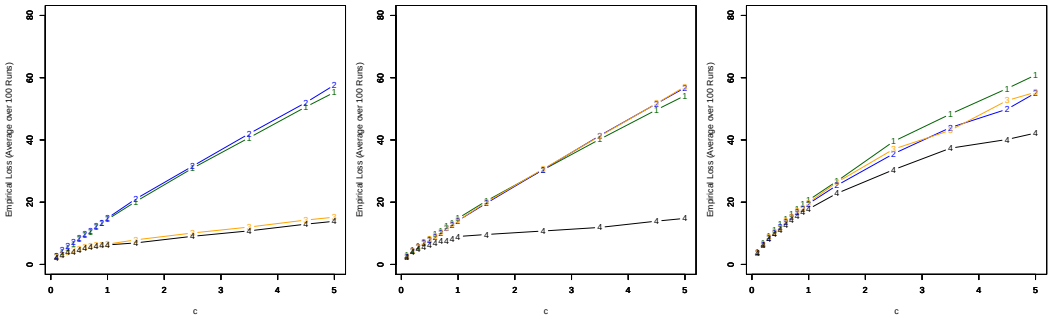
Figure 1.1 shows some representative plots which compare the average total loss of the normal prior competitors under different sparsity and signal distribution conditions when the signal is relatively strong. For this figure, the non-zero θ_i's are generated from the $N(5,1)$ distribution, from a mixture of $N(5,1)$ and $N(-5,1)$ distributions, or from a unit mass at the value 5. The proportion $1-\omega$ of non-zero θ_i's is set to $0.05, 0.10$, and 0.30. The classifiers are compared at various values of the cost constant c, which corresponds to the relative cost placed on false negative results when signal is mistaken for noise. Our nonparametric classifier typically outperforms the parametric competitors (i.e., has lower average loss) for broad ranges of c values when the prior distribution is not specified correctly. In the graph, all three classifiers were compared to a classification rule for which the correct prior distribution was used. Similar simulation setups with weaker signal and higher signal sparsities were also tried. For weaker signal, the performance of the parametric and nonparametric classifiers were typically closer.



(a) $w = 0.70$, $\theta = 5$ (b) $w = 0.70$, $\theta = 5$ or -5 (c) $w = 0.70$, θ from $N(5, 1)$ or $N(-5, 1)$



(d) $w = 0.90$, $\theta = 5$ (e) $w = 0.90$, $\theta = 5$ or -5 (f) $w = 0.90$, θ from $N(5, 1)$ or $N(-5, 1)$



(g) $w = 0.95$, $\theta = 5$ (h) $w = 0.95$, $\theta = 5$ or -5 (i) $w = 0.95$, θ from $N(5, 1)$ or $N(-5, 1)$

Figure 1.1: Comparison of average empirical loss over 100 simulation runs with varying signal sparsity and distribution for the parameter θ.

1.6 Classification of Microarray Experiment Output

Cellular organisms have internal biochemical mechanisms that help them to adjust to changes in the surrounding environment by activating or repressing the expression of certain parts of their genome in response to external changes. To better understand which areas of an organism's genome are involved in its response to outside factors, researchers can use microarrrays to compare gene expression levels under various conditions. In this section, we apply two of the classification rules described before, the nonparametric and parametric with estimated mean and variance, to a publicly available gene expression dataset.

Many observed differences in gene expression may indeed be due to the change in conditions under investigation. Given the large number of genes involved and the complexity of the genome, other changes in expression levels, however, may be due to other factors. In identifying the part of the genome actually involved in the organism's response, classification algorithms which balance the costs of false positives and false negatives can therefore be useful. As discussed above, the loss function in (1.3.2) is readily interpretable in this context.

1.6.1 Gene Expression Data

When yeast cells experience harsh changes in their surroundings, they activate internal mechanisms to mitigate the stress. In the dataset, expression levels of several thousand genes for yeast cells were compared before and after temperature and

chemical shocks to their environment. The data is collected using two-channel microarray techniques for multiple timepoints. The researchers measure changes in expression levels at several times after the environmental shock and use statistical techniques to cluster sets of genes with similar expression patterns to help identify the parts of the yeast genome which are involved in various stress-response mechanisms.

To illustrate the use of our classifiers, we focus on the data collected from just one timepoint after a yeast colony has been subjected to an increase in hydrogen peroxide concentration. We then work with the data as though all of the observations are independent, as specified in model (1.2.1). In future work, we hope to extend our nonparametric Bayes rule to deal with the richer time and dependence structure of multiple timepoint microarray data.

1.6.2 Classification Results

The classification rules were tried on relative gene expression levels for one timepoint in one of the microarray experiments (microarray y9-40, 10 minutes of exposure to hydrogen peroxide) from the publicly available data. The gene expression data is reported as "zero transformed" observations which summarize the gene expression at each post-environmental change timepoint relative to expression levels before the change. Positive (negative) values correspond to genes for which relative expression levels were seen to go up (down) after the change to the environment. The classifiers were then used to identify genes which are "sufficiently" over- or underexpressed given different values of the cost constant c.

Observations are characterized as signal by a classifier if and only if they are in a region where the $\hat{p}_i$ curve for the classifier is below the corresponding Bayes rule threshold curve. Thus the decision rule for each classifier is characterized by the relative gene expression levels at which its $\hat{p}_i$ curve crosses its Bayes rule threshold curve. It was found that, for example, for $c = 4$, the nonparametric prior Bayes procedure classifies all observations with relative expression levels outside the region $[-3.10, 2.38]$ as signal. For $c = 4$, the corresponding region for the parametric prior rule is very slightly more conservative for the underexpressed genes and very slightly less conservative for the overexpressed genes; it classifies the observations outside of $[-3.20, 2.35]$ as signal. For $c = 4$, the classification results are extremely close, with the nonparametric prior rule classifying 80 genes as differentially expressed as compared to 79 for the parametric prior rule.

Results for other values of c were also obtained. For $c = 10$, for example, the difference between the two classifiers becomes much more noticeable, with 153 signal genes for the nonparametric prior classifier and 214 for the parametric prior rule. It can be seen that, for reasonable values of c, the classification decisions provided by the two rules in a particular dataset can vary greatly or overlap almost exactly depending on the cost constant. At the same time, it should not be surprising that there is less overlap in the decision rules for higher values of c if, as is the case for this dataset, most of the values are concentrated closer to 0, so that even small changes in the threshold boundaries can produce large changes in the decision vector $\mathbf{a}$.

1.7 Conclusion

In this chapter, we propose a Bayesian classifier in the context of a highly interpretable loss function. While parametric Bayes classifiers may be conceptually simpler, the nonparametric rule outperforms them in terms of the risk function when the prior is not specified correctly. In particular, when the prior distribution is misspecified for the parametric classifier, the nonparametric technique dominates over the range of c values. This is reassuring because the particular choice of c is a measure of the relative cost of false negatives to a researcher and, in practice, may be difficult to specify precisely for some classification problems.

We illustrate the performance of two procedures using a publicly available gene expression data set. It is seen that, while the decisions produced by the rules can be similar, they can also vary greatly for reasonable values of the cost constant c. For the gene expression application in this chapter, we focus on a single time point from a multi-timepoint microarray experiment and treat the observations as if they were independent. In future work, we hope to extend the nonparametric classification procedure to capture time and observation dependence structure.

Chapter 2

Classification Procedures based on False Discovery Rates

2.1 Introduction

The simultaneous testing of multiple statistical hypotheses has been an active area of research for many decades. The need to make many decisions at the same time arises in the most diverse applications. One of the principal concerns of the multiple testing literature is the search for useful criteria for evaluating statistical decision-making techniques; given a criterion, techniques which satistfy it are necessary. In this chapter, we focus on one such criterion, the False Discovery Rate, and propose new ideas aimed at bounding it using nonparametric Bayesian techniques.

The False Discovery Rate (FDR) was introduced by Benjamini and Hochberg in 1995. Since then, a large literature has grown around it. Here, we briefly touch on

the papers that set up the main ideas that will be necessary for the sequel.

The FDR of a hypothesis testing procedure is defined as the expected proportion of falsely rejected hyptoheses under the procedure given that at least one hypothesis is rejected by it multiplied by the probability of rejecting at least one hypothesis. Benjamini and Hochberg (1995) provided a so-called "linear step-up" procedure for controlling the FDR at a desired level based on p-values for the case where the test statistics from the hypotheses are independent. In subsequent work, Yekutieli and Benjamini (2001) showed that the same p-value step-up procedure controls the FDR for a broad class of dependence structures for the test statistics. Storey (2002; 2003) focused on another criterion which had been highlighted by Benjamini and Hochberg (1995) : the expected proportion of falsely rejected hypotheses given that at least one hypothesis is rejected. Benjamini and Hochberg (1995) had ultimately rejected this criterion, called positive FDR or pFDR by Storey (2002), in favor of the FDR because the pFDR cannot be controlled in cases when all of the null hypotheses are true. Yet, as Storey (2002) showed, if the test statistics are independent and if it is assumed that whether each test statistic truly comes from the null hypothesis or from the alternative can be thought of as binomial trials with some constant probability of success, then the pFDR corresponding to the hypothesis rejection region equals the probability that a hypothesis is null given that its test statistic falls in the rejection region. This powerful connection makes the pFDR a natural quantity to study in the Bayesian framework, and Storey (2002, 2003) proposes a procedure that estimates this quantity for preselected rejection regions.

In related work, Efron et al. (2001) connected the FDR criterion to the empirical

Bayes approach. The quantity of interest in their work was the probability that a null hypothesis is true given the value of the associated test statistic. This posterior probability was termed the *local* FDR since it was shown to be equivalent to the FDR if the rejection region were restricted to a small ("local") region around this realized value of the test statistic. Efron et al. (2001) proposed one method for estimating this posterior probability without parametric assumptions about the alternative hypothesis distribution; other methods are, of course, also available, and much of our later discussion focuses on studying the connection to FDR of the nonparametric Bayes approach suggested by Raykar and Zhao (2010) .

The layout for this Chapter is as follows: in Section 2 of this Chapter, we begin by discussing the FDR criterion as it was introduced by Benjamini and Hochberg (1995). We then examine later work which built on the original formulation. We look at the nonparametric prior procedure from the previous chapter and fit it within this framework.

2.2 Multiple Hypothesis Testing and the False Discovery Rate Criterion

In this section, we lay out a framework for studying binary classification and introduce the notation used in this chapter (most of our notation follows the standard notation in Benjamini and Hochberg (1995)). To discuss this problem, it will be convenient to refer to a so-called "confusion matrix," shown in Table 2.1, which summarizes the counts of correctly and incorrectly classified observations. In this

Table 2.1: A "confusion matrix" for multiple hypothesis testing

	Declared non-significant	Declared significant	Total
H_0 is true	U	V	m_0
H_1 is true	T	S	$m - m_0$
	$m - R$	R	m

table, m represents the total number of hypotheses being tested and m_0 stands for the number of truly null hypotheses among them. Given the data and a classification procedure, the null hypothesis is rejected in R of the m hypothesis tests. Of the R rejections, V are incorrect because they come from the m_0 hypothesis tests in which the null hypothesis is in fact true. Similarly, there are T incorrect declarations of non-significance for which the alternative hypothesis is actually true. Obviously, it is desirable to have classificiation procedures for which both V and T are small, but, usually, trade-offs are necessary. Traditionally, in formulating multiple hypothesis testing procedures, the focus has been on controlling the quantity $Prob(V > 0)$, which is called the family-wise error rate (FWER). For example, one well-known approach which controls the FWER is the Bonferroni procedure (for extensive references, see for example, Lehmann's *Testing Statistical Hypotheses* (2005)). Quite often, however, approaches which control the FWER are much more conservative than needed for specific applications. For such cases, other criteria for evaluating the performance of procedures for multiple hypothesis testing are available. One especially popular criterion is the control of the so-called false discovery rate, or FDR, as suggested by Benjamini and Hochberg (1995). We presently discuss the details of the FDR criterion and a procedure to control it in the next section.

2.2.1 The False Discovery Rate

The false discovery rate (FDR) was introduced by Benjamini and Hochberg (1995) as a quantity to control for multiple hypothesis testing. The FDR is defined as the expected proportion of incorrectly rejected hypotheses among all the rejected hypotheses given that at least one ejection is made times the probability that at least one rejection is made. That is

Definition:

$$FDR = E(\frac{V}{R}|R > 0)Prob(R > 0) \tag{2.2.1}$$

For many situations, this much less stringent error rate is more reasonable than the FWER.

A procedure to control the FDR at a preset level α was also introduced in Benjamini and Hochberg (1995). The procedure consists of ranking the p-values from the m hypothesis tests and then comparing each of them, in order from smallest to largest, to a constant that depends on the rank of the p-value, on m, and on α. The first time a p-value exceeds the corresponding constant, the procedure is stopped, with that p-value and all of the smaller p-values declared to belong to tests in which the null hypothesis should be rejected. In other words, the procedure goes as follows:

The Benjamini-Hochberg (1995) step-up procedure:

1. Rank from smallest to largest the p-values from the m tests. Denote the resulting ordered list as $p_{(1)}, p_{(2)}, ..., p_{(m)}$ and denote the corresponding hypothesis tests by $H_{(1)}, H_{(2)}, ..., H_{(m)}$.

2. Let $\hat{k} = max\{i : p_{(i)} \leq \frac{i}{m}\alpha\}$.

3. Reject the hypotheses $H_{(1)}, ..., H_{(k)}$ and accept the others.

Benjamini and Hochberg show that this procedure results in $FDR \leq \frac{m_0}{m}\alpha \leq \alpha$. It is important to emphasize that this procedure only provides "control" of an expected quantity and not of the proportion of falsely rejected null hypotheses in a particular sample, since the FDR is an expected value. Also, note that FDR is actually controlled at a level which is typically more conservative than the stated level α. If m_0 were known, then the procedure could be made less conservative. Of course, m_0 is unknown, but estimates of m_0 or of the fraction m_0/m can be made, and later work by Storey (2002) and by Benjamini, Krieger, and Yekutieli (2006) shows that less conservative procedures can be obtained by using estimates of these quanitites.

2.2.2 The pFDR criterion

A Bayesian framework for FDR was studied by Storey (2002, 2003). He reexamined a quantity that had originally been considered and rejected in favor of FDR in the work of Benjamini and Hochberg (1995) and showed that, under broadly applicable assumptions, it is equivalent to the posterior probability that the null hypothesis is true given that the associated test statistic falls in the rejection region. This quantity is defined as the expected value of the fraction of false discoveries among all the discoveries, given that at least one discovery has occurred; when no discoveries have occurred, this quantity is set to zero. Using the notation from above, this can

be written as

$$pFDR = E\left(\frac{V}{R}|R > 0\right). \tag{2.2.2}$$

Storey (2002, 2003) shows that when the sample consists of indpendent observations, then the pFDR corresponds to the probability that a null hypothesis is true given that it was declared false. Using Storey's notation, this can be written as

$$pFDR(\Gamma) = Prob(H = 0|X \in \Gamma), \tag{2.2.3}$$

where H is a binary indicator for whether the null hypothesis is true, X is some test statistic, and Γ stands for the rejection region. The formula can also be written in terms of rejection regions for p-values, with hypotheses with test statistics that have p-values in some interval $[0, \gamma]$ being rejected. We can then rewrite the formula above as

$$pFDR(\gamma) = \frac{\pi_0 Prob(p - value \leq \gamma|H = 0)}{Prob(p - value \leq \gamma)}. \tag{2.2.4}$$

Storey (2002) proposes a technique for estimating the proportion π_0 of true null hypotheses, the probability $Pr(R > 0)$ that at least one hypothesis is rejected, and the pFDR associated with a particular rejection region $[0, \gamma]$:

$$\hat{\pi}_{0,Storey}(\lambda) = \frac{W(\lambda)}{(1 - \lambda)m}, \tag{2.2.5}$$

$$\hat{Pr}(R > 0) = \frac{R(\gamma)}{m} \tag{2.2.6}$$

and

$$pFDR_{\lambda}(\gamma) = \frac{\hat{\pi}_{0,Storey}(\lambda)\gamma}{\hat{Pr}(R > 0)(1 - (1 - \gamma)^m)} \tag{2.2.7}$$

where $W(\lambda)$ is the number of p-values which exceed a tuning parameter λ, $R(\gamma)$ is the number of p-values that fall in the rejection interval $[0, \gamma]$, and m is the total number of hypotheses being tested.

Note that the last formula provides an *estimator* of pFDR for a rejection region $[0, \gamma]$ of p-values selected by the researcher. Storey (2002) shows that this estimator has an upward bias for estimating the true pFDR of a rejection region. The estimation approach can be contrasted against the more traditional aim of providing procedures which control the error rate at a desired level α, such as the linear step-up procedure of Benjamini and Hochberg (1995). In the next section, we discuss a connection between FDR, pFDR, and empirical Bayes procedures. We then look at ways of bounding pFDR using a nonparametric Bayesian approach.

2.3 Nonparametric Bayesian Classification and FDR

The FDR-controlling procedures described above rely on the explicit use of observed p-values. We now change focus to a different approach which uses the posterior probabilty that a null hypothesis is true given the value of the associated test statistic. This quantity can be estimated using empirical Bayes techniques and the average such posterior probability in a rejection region turns out to equal the pFDR of the region, as discussed by Efron et al. (2001) and Efron (2005). The goal of this section is to provide a procedure for multiple hypothesis testing using the connection

between FDR and nonparametric Bayes techniques. A useful model for this setting can be expressed as the mixture f of a null density f_0 and an alternative density g

$$f(z) = \pi_0 f_0(z) + (1 - \pi_0) g(z), \tag{2.3.1}$$

Here, the symbol z stands for one-dimensional scores which are used for making the binary classification decisions. For example, they may be transformed gene expression values from a large microarray experiment in which the goal is to determine which genes change their expression levels (that is, they become overexpressed or underexpressed) in response to a biologically interesting treatment, such as ioninzing radiation. As discussed by Efron et al. (2001) , it is often the case that an appropriate data reduction technique must first be found to form the one-dimensional statistics Z, since data on multiple characteristics for each unit of observation is often available. Ways of reducing the data to single-dimenisonal summary statistics are discussed in the paper, but these are not integral to our discussion in this section. Here, we focus instead on the alternative methods for estimating posterior probabilities.

Using Bayes' Rule, the posterior probability of interest for the $i'th$ observation z_i can be written as

$$1 - p_i(z_i) = 1 - \pi_0 f_0 / f(z_i) \tag{2.3.2}$$

and

$$p_i(z_i) = \pi_0 f_0(z_i) / f(z_i), \tag{2.3.3}$$

where $p_i(z_i)$ is the a posteori probability that the null hypothesis is true for an obseration with summary score z_i. In Efron et al. (2001) , this is called the local FDR because, asymptotically, it is equivalent to the proportion of falsely rejected hypotheses if the rejection region consists of test statistics close to z_i. Note that the expression on the right side of equation 2.3 has the same general form as the quantity computed in 1.4.8 of Chapter 1. We look at this connection next.

Note that the mixture density f can be estimated from the data, but this estimate is not directly useful by itself if the density f_0 is unknown. To remedy this, one can make distributional assumptions or use permutations of the density f_0, as is done, for example, in Efron (2001), Efron (2005), Raykar and Zhao (2010), and other work where some prior distribution is assumed for the null density f_0.

2.4 Simulation Results

Our work compares the empirical false discovery proportion V/R for various classification rules. We report the results of several simulations here. For each simulation, 500 observations were generated from two distributions, the null hypothesis distribution N(0,1) and some alternative hypothesis distribution. The false discovery rates associated with several different classification rules were then computed. For each combination of null and alternative hypothesis distributions, these computations were repeated for 100 samples of 500 observations. We performed computations for each of the significance levels (α) reported in the first column of the table. In the second column, we report the average of the ratios V/R realized in the 100 runs

when using the nonparametric Bayes rule at each significance level α with true values of g and π_0 plugged in. In the third column, we use the same rule, but with estimates of g and π_0 plugged in. The averages in this third column are computed in the following way:

Algorithm 3.1

The non-parametric prior empirical Bayes rule for significance level α.

1. For each observation z_i, compute the estimate $\hat{p}_{iNB}(z_i)$ for the posterior probability $Prob(\beta_i = 0|z_i, \mathbf{z})$ using the formula

$$\hat{p}_{iNB}(z_i) = \frac{\hat{\pi_0}\phi(z_i)}{\hat{\pi_0}\phi(z_i) + (1 - \hat{\pi_0})\hat{g}(z_i)}, \tag{2.4.1}$$

where $\hat{\pi_0}$ stands for the estimate of $\pi_0 = Prob(\beta_i = 0)$, $\hat{g}(z_i)$ is an estimate of the marginal density of z_i given that $\beta_i \neq 0$, and $\phi(z_i)$ is the value of the $N(0, 1)$ density at z_i.

2. Order the values $\hat{p}_{NB}(z_i)$ computed in Step 1 from smallest to largest and denote the ordered list as $\{\hat{p}_{(1)}, ..., \hat{p}_{(m)}\}$ and let $x_(j)$ stand for the observation from Step 1 that is associated with the j'th largest $\hat{p}$ (and NOT for the j'th largest z_i). Let

$$K = max\{k\, s.t.\, \sum_{j=1}^{k} \hat{p}_{(k)}/k \leq \alpha\}. \tag{2.4.2}$$

Classify the observations $x_(j)$ as having come from the alternative distribution for $j \leq K$ and from the null distribution for all other j.

In the fourth column of each table, we report the average proportion of false

rejections when using the original linear step-up Benjamini and Hochberg (1995) procedure. The fifth column of each table shows the results for the two-stage version of the step-up procedure introduced in Benjamini, Krieger, and Yekutieli (2006), in which the ratio m_0/m of true null hypotheses is estimated. The sixth column has results for an ad hoc version of the two-stage step up procedure of the fifth column in which $\hat{\pi}_{0,NB}$ is used to estimate the ratio m_0/m. Finally, the seventh column presents results based on the estimates of pFDR from Storey (2002). As noted above, the procedure in Storey (2002) is unlike the other approaches in the sense that it provides estimates of an error rate for predetermined rejection intervals instead of providing rejection intervals for desired error levels of an error rate. To make the procedures comparable, we compute the Storey (2002) estimate of pFDR for each observation and reject the null hypothesis for all observations for this estimate falls below the desired level α.

2.5 Discussion of Simulation Results

The aim for methods that focus on the FDR and pFDR is usually a slight conservative bias in expectation. In other words, the goal is typically to come up with procedures for which the expected value of the error rate in question falls below the nominal significance level α. For example, the linear step-up procedure proposed by Benjamini and Hochberg (1995) controls the FDR, which is defined as the expected value in (3.3.6), below the desired nominal rate α. In fact, the control for this procedure is at the more conservative rate $\frac{m_0}{m}\alpha$ and the later work by Benjamini, Krieger,

Table 2.2: Average empirical V/R, testing $\beta_i = 0$ against $\beta_i = 2$, π_0=0.9

α	true NB	NB	BH	YKB	YKB using $\hat{\pi}_{NB0}$	Storey
0.01	0	0	0.003	0.003	0.003	0
	(0)	(0)	(0.033)	(0.033)	(0.033)	(0)
0.05	0.026	0.038	0.037	0.037	0.039	0
	(0.067)	(0.078)	(0.091)	(0.09)	(0.092)	(0)
0.1	0.086	0.081	0.066	0.066	0.085	0.007
	(0.078)	(0.082)	(0.097)	(0.097)	(0.099)	(0.052)
0.15	0.133	0.117	0.118	0.12	0.139	0.104
	(0.072)	(0.085)	(0.108)	(0.11)	(0.122)	(0.192)
0.2	0.186	0.146	0.174	0.178	0.205	0.192
	(0.074)	(0.083)	(0.123)	(0.125)	(0.125)	(0.206)
0.25	0.238	0.184	0.222	0.224	0.248	0.281
	(0.072)	(0.087)	(0.115)	(0.116)	(0.113)	(0.174)
0.3	0.29	0.23	0.269	0.276	0.307	0.324
	(0.072)	(0.088)	(0.117)	(0.122)	(0.115)	(0.137)

Table 2.3: Average empirical V/R, testing $\beta_i = 0$ against $\beta_i = 2$ or -2, π_0=0.9

α	true NB	NB	BH	YKB	YKB using $\hat{\pi}_{NB0}$	Storey
0.01	0	0	0	0	0	0
	(0)	(0)	(0)	(0)	(0)	(0)
0.05	0.031	0.047	0.028	0.028	0.037	0
	(0.13)	(0.112)	(0.083)	(0.083)	(0.092)	(0)
0.1	0.064	0.07	0.065	0.066	0.077	0.003
	(0.111)	(0.118)	(0.101)	(0.103)	(0.105)	(0.033)
0.15	0.126	0.107	0.115	0.117	0.132	0.092
	(0.112)	(0.108)	(0.115)	(0.115)	(0.119)	(0.171)
0.2	0.193	0.165	0.166	0.159	0.197	0.191
	(0.102)	(0.148)	(0.124)	(0.119)	(0.127)	(0.187)
0.25	0.248	0.193	0.23	0.235	0.256	0.294
	(0.093)	(0.115)	(0.12)	(0.122)	(0.114)	(0.201)
0.3	0.292	0.236	0.286	0.287	0.302	0.329
	(0.084)	(0.137)	(0.105)	(0.105)	(0.11)	(0.131)

Table 2.4: Average empirical V/R, testing $\beta_i = 0$ against β_i from $0.5N(2,1) + 0.5N(-2,1)$, π_0=0.9

α	true NB	NB	BH	YKB	YKB using $\hat{\pi}_{NB0}$	Storey
0.01	0.002	0.006	0.006	0.006	0.008	0
	(0.017)	(0.034)	(0.034)	(0.034)	(0.036)	(0)
0.05	0.042	0.041	0.038	0.04	0.04	0
	(0.078)	(0.078)	(0.072)	(0.073)	(0.072)	(0)
0.1	0.088	0.08	0.085	0.087	0.093	0.052
	(0.076)	(0.087)	(0.084)	(0.086)	(0.085)	(0.126)
0.15	0.144	0.126	0.137	0.142	0.153	0.167
	(0.079)	(0.085)	(0.085)	(0.086)	(0.087)	(0.136)
0.2	0.194	0.169	0.183	0.188	0.2	0.224
	(0.07)	(0.084)	(0.086)	(0.088)	(0.087)	(0.107)
0.25	0.249	0.208	0.232	0.238	0.247	0.277
	(0.076)	(0.09)	(0.095)	(0.096)	(0.1)	(0.108)
0.3	0.307	0.258	0.271	0.29	0.312	0.33
	(0.067)	(0.085)	(0.1)	(0.099)	(0.095)	(0.108)

Table 2.5: Average empirical V/R, testing $\beta_i = 0$ against $\beta_i = 5$, π_0=0.9

α	true NB	NB	BH	YKB	YKB using $\hat{\pi}_{NB0}$	Storey
0.01	0.006	0.012	0.008	0.008	0.008	0
	(0.011)	(0.014)	(0.013)	(0.013)	(0.014)	(0)
0.05	0.039	0.073	0.044	0.05	0.052	0.069
	(0.017)	(0.023)	(0.027)	(0.029)	(0.031)	(0.041)
0.1	0.09	0.14	0.088	0.101	0.104	0.117
	(0.018)	(0.025)	(0.04)	(0.044)	(0.046)	(0.05)
0.15	0.141	0.199	0.137	0.155	0.158	0.169
	(0.018)	(0.028)	(0.048)	(0.054)	(0.054)	(0.057)
0.2	0.193	0.256	0.179	0.205	0.207	0.214
	(0.017)	(0.028)	(0.057)	(0.061)	(0.061)	(0.067)
0.25	0.244	0.309	0.225	0.258	0.26	0.264
	(0.016)	(0.028)	(0.063)	(0.065)	(0.067)	(0.071)
0.3	0.294	0.362	0.272	0.308	0.31	0.31
	(0.015)	(0.028)	(0.065)	(0.068)	(0.068)	(0.071)

Table 2.6: Average empirical V/R, testing $\beta_i = 0$ against $\beta_i = 5$ or -5, π_0=0.9

α	true NB	NB	BH	YKB	YKB using $\hat{\pi}_{NB0}$	Storey
0.01	0.006	0.01	0.008	0.008	0.009	0
	(0.011)	(0.014)	(0.013)	(0.014)	(0.015)	(0)
0.05	0.042	0.073	0.044	0.05	0.052	0.069
	(0.02)	(0.025)	(0.028)	(0.03)	(0.03)	(0.041)
0.1	0.091	0.141	0.088	0.101	0.104	0.118
	(0.022)	(0.027)	(0.04)	(0.045)	(0.046)	(0.051)
0.15	0.142	0.203	0.137	0.155	0.16	0.169
	(0.022)	(0.029)	(0.048)	(0.053)	(0.055)	(0.057)
0.2	0.193	0.26	0.179	0.205	0.208	0.214
	(0.022)	(0.03)	(0.057)	(0.061)	(0.062)	(0.067)
0.25	0.244	0.314	0.226	0.257	0.261	0.264
	(0.02)	(0.029)	(0.063)	(0.065)	(0.066)	(0.07)
0.3	0.294	0.368	0.272	0.308	0.31	0.31
	(0.019)	(0.029)	(0.065)	(0.067)	(0.067)	(0.071)

and Yekutieli (2006) refines the original approach to make the control in expectation less conservative by estimating m_0/m. Similarly, Storey (2002, 2003) shows that, in expectation, his estimator overshoots the true pFDR of a fixed rejection region.

On the other hand, as far as we know, there are no procedures that control the Bayes posterior probabilities p_i or their averages in expectation. The current absence of such procedures may be seen as a weakness of using posterior probabilitites estimated by nonparametric Bayes approaches for working with false discovery rates. And yet, if the estimates of p_i are good, this approach seems reasonable and offers greater flexibility for interpretation, as argued by Efron et al. (2001), because estimates of posterior probabilities are provided for each tested hypothesis.

In our simulations, we focused on the realized proportion V/R of hypotheses which

were falsely rejected among all the declared rejections. In tables 2.2-2.6, we report the empirical average and empirical standard deviation of this proportion for several different classification procedures. The second column in these tables (labeled as "true NB") reports results for the nonparametric Bayes procedure with true marginal density g and true population proportion π_0. Not surprisingly, this column gives excellent results in the sense that the averages fall just below the stated significance levels α and the empirical standard deviations are the lowest in the tables. Of course, the true g and π_0 are unknown and must be estimated. This is done in the next column, labeled NB, using the estimates of g and π_0 provided by Raykar and Zhao (2010a). Reassuringly, this column gives results which are quite close to the results for the classification procedure which uses the true g and π_0 for much of the time. Interestingly, the simulation results for this procedure are better for the harder classification problems described in tables 2.2-2.4 than in the relatively easier ones in tables 2.5 and 2.6; for the latter set-ups, the nonparametric Bayes procedure rejects too many hypotheses. A similar pattern is seen for the classification procedures described in the fifth and sixth columns, which correspond to two different versions of the linear step-up procedure for which the population proportion π_0 of null hypotheses is estimated. The results in the fourth column of each table are for the original Benjamini and Hochberg (1995) procedure. The empirical average results in this column always fall below the stated level α, but the nonparametric Bayes results are often better for small values of α.

The results in the last column were produced using the estimator of pFDR that is described in Storey (2002, 2003) (with $\lambda = 1/2$, as in the first section of Storey

(2002), but using two-sided p-values). Because the procedure described in Storey's work is an estimator of pFDR for fixed rejection regions and the other procedures used in the simultations are instead ways of limiting pFDR given the desired significance level, Storey's approach is not directly comparable to the others. To make it comparable, we first use it to compute estimates of pFDR for every hypothesis and then reject the hypotheses for which these estimates fall below the deisred level α. The results using this procedure seem to be too conservative for smaller values of α and (slightly) too liberal for larger signifcance levels.

Chapter 3

A Recalibration Procedure which maximizes the AUC: A Use-Case for Binormal Assumtions

3.1 Introduction and Related Work

Most binary classifiers make their final decision as to whether an instance is positive or negative based on a scalar score, which is computed as a function of the features corresponding to that instance. The popular and widely used procedure chooses a single threshold value on the score scale and assigns the positive label (1) to observations with scores that fall above this value and the negative label (0) to observations with scores that fall below it. The general form of the classification

threshold can then be written as

$$\delta(\boldsymbol{x}) = \begin{cases} 1 & \text{if } f(\boldsymbol{x}) \geq \theta \\ 0 & \text{otherwise} \end{cases}, \tag{3.1.1}$$

where $f(\boldsymbol{x})$ is the raw score computed as a function f for an instance $\boldsymbol{x} \in \mathbf{R}^d$ (the d-dimensional feature vector) and θ is an appropriately chosen *threshold parameter*. This thresholding rule is essentially built on the assumption that a larger score $f(\boldsymbol{x})$ provides a larger chance of $y = 1$.

One popular way of evaluating the performance of such binary classification rules is to use the Receiver Operating Characteristic (ROC) curve. There are connections between binary regression generalized linear models and ROC curves (Pepe, 2000). The ROC curve essentially is a plot of the *sensitivity* on the y-axis and *1-specificity* on the x-axis. Each threshold θ corresponds to a point on the ROC plot and the ROC curve is obtained as θ is swept from $-\infty$ to ∞. Classifiers that simultaneously have higher sensitivity and higher specificity are more desirable and dominate their competitors. In practice, however, one usually finds several classifiers with intersecting ROC curves. One popular procedure selects the classifier with the highest area under its ROC curve (AUC).

In this chapter we focus on a raw score recalibration procedure that can maximize the AUC for thresholding rules under certain assumptions. We do not dwell on the particular classifier used, as the recalibration we propose can be used with any general black-box classifier which uses scores to make a final decision. Area under the receiver operating characteristic curve is a popular measure for evaluating

the quality of binary classification rules. Commonly used score-based classifiers label an outcome as a positive if the score is greater than a certain threshold. We show that this may not be optimal in terms of maximizing the AUC. Under certain assumptions the optimal thresholding rule is derived using the Neyman-Pearson lemma. Specifically, we show that a thresholding rule that is quadratic in the score dominates the commonly used linear thresholding rule. We discuss the following facts:

1. We show that the commonly used linear thresholding rule (3.1.1) is not optimal in terms of maximizing the AUC.

2. We show that a simple quadratic transformation of the scores is optimal in terms of maximizing the AUC. Specifically, using the Neyman-Pearson Lemma (Section 3.3.3) we show that the following quadratic thresholding rule

$$\delta(\boldsymbol{x}) = \begin{cases} 1 & \text{if } a(f(\boldsymbol{x}))^2 + bf(\boldsymbol{x}) + c \geq \theta \\ 0 & \text{otherwise} \end{cases}, \tag{3.1.2}$$

   maximizes the AUC under certain assumptions, where a, b, and c are constants chosen based on the training set (see Section 3.3.3).

3. Our results are based on the assumption that the scores for the positive and negative populations are normally distributed with different parameters. As discussed below, this is a reasonable assumption for scores produced by many classifiers. We further show that the commonly used linear classification rule

(3.1.1) and the proposed quadratic rule (3.1.2) agree when the standard deviations of the two normal distributions are equal.

Essentially, a thresholding rule that is quadratic in the score dominates the commonly used linear rule when the variance of the score for the positive population is different from the variance of the score for the negative population in the bi-normal case. Hence a very simple method to improve the score from any general classifier is to recalibrate the scores using the quadratic transformation $s' \leftarrow as^2 + bs + c$, where s and s' denote, respectively, a raw score and its quadratic transformation.

The Neyman-Pearson lemma guarantees that the AUC for a procedure which thresholds the transformed scores s' will be greater than or equal to the AUC of thresholding the raw scores s. In fact, the Neyman-Pearson guarantee is stronger: under the bi-normality assumption, thresholding the quadratically-transformed scores dominates any other thresholding rule in the sense that it is guaranteed to produce an ROC curve that is uniformly above those of the other procedures. In Section 4, we illustrate our procedure with data where a 25 % increase in AUC is observed. Our experimental results with other datasets show more modest, but positive gains in the AUC obtained by simply applying the proposed quadratic transformation without resorting to any sophisticated AUC-maximizing classifiers proposed in the literature, as, for example, in the work of Herschtal and Raskutti (2004).

The proposed quadratic recalibration procedure is, of course, not appropriate for all datasets. Indeed, it is well documented (see, for example, (Bennett, 2003)) that mode-symmetry assumptions about the distributions of scores for the positive and negative populations in binary classification are difficult to justify in many setups.

In (Platt, 1999), there are examples of score-based classification in the context of clearly non-Gaussian class distributions. On the other hand, the bi-normal framework is reasonable for a range of frequently-used applications; a typical use-case to motivate the Gaussian assumptions with non-equal class-conditional variances may come about in datasets for which linear and logistic regression techniques are often used to construct scores in practice. For example, scores which are linear combinations of a multitude of covariates (or their transformations) are often, at least approximately, normal by Central Limit Theorem considerations; for illustration purposes, we provide an application of our recalibration procedure to a restaurant patron tipping dataset from in which we use tipping percentages to classify patrons as smokers or non-smokers.

The remainder of the chapter is organized as follows. The problem of binary classification based on scores and the commonly used raw score-based thresholding is presented in Section 3.2; we also review the ROC curve (Section 3.2.3) and area under the ROC Curve (Section 3.2.4) framework for classifier evaluation. The proposed AUC-maximizing raw score recalibration is presented in Section 3.3 by invoking the Neyman-Pearson lemma (Section 3.3.1) and the bi-normality assumption for the classifier scores (Section 3.3.2). The proposed quadratic score based thresholding rule is presented in Section 3.3.3. The improvement obtained is clearly illustrated in Section 3.4 on the restaurant tipping dataset, a real dataset in which the class-conditional variances vary by a factor of 2.7.

3.2 Binary classification based on scores

In a typical binary classification scenario we are given a training set $\mathcal{D} = \{(\boldsymbol{x}_j, y_j)\}_{j=1}^n$ containing n instances, where $\boldsymbol{x}_j \in \mathbf{R}^d$ is an instance (the d-dimensional feature vector) and $y_j \in \mathcal{Y} = \{0, 1\}$ is the corresponding known label. The task is to learn a *classification function* $\delta : \mathbf{R}^d \to \mathcal{Y}$, which minimizes the error on the training set and generalizes well on unseen data.

3.2.1 Discriminant function and classifier score

Instead of learning δ directly, very often it is convenient to learn a real valued *discriminant function* $f : \mathbf{R}^d \to R$. The discriminant function can take different forms depending on the specific classifier. For example, for linear classifiers like logistic regression, linear discriminant analysis (LDA), linear support vector machine (SVM), etc., the discriminant function f is a linear function of the feature vector, that is, $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$ where $\boldsymbol{w} \in \mathbf{R}^d$ is the weight vector and the scalar b is the bias term. For non-linear kernel machines like SVM the discriminant function is of the form $f(\boldsymbol{x}) = \sum_{j=1}^n \alpha_j k\left((\boldsymbol{x} - \boldsymbol{x}_j)/h\right)$ where k is the kernel function and h is the bandwidth of the kernel. For a neural network f is essentially the final output of the neural net obtained via forward propagation. We will refer to this value of the discriminant function as the score for an instance, that is, $s = f(\boldsymbol{x})$. We can now rewrite equation (3.1.1) inserting s for $f(\boldsymbol{x})$.

3.2.2 Score-based thresholding

Irrespective of the classifier the final classification function is usually written as

$$\delta(\boldsymbol{x}) = \begin{cases} 1 & \text{if } s = f(\boldsymbol{x}) \geq \theta \\ 0 & \text{otherwise} \end{cases}, \tag{3.2.1}$$

where θ is an appropriately chosen *threshold parameter*. This thresholding rule is monotonic in s and uses a single threshold value θ to decide between $y = 1$ and $y = 0$. It is built on the assumption that a larger score $s = f(\boldsymbol{x})$ provides a larger chance of $y = 1$.

3.2.3 Receiver Operating Characteristic curve

One popular way of evaluating the performance of such binary classification rules is to use the Receiver Operating Characteristic curves (ROC curves). These curves give a convenient graphical representation of two-by-two contingency tables or *confusion matrices*that can be used to formally evaluate classifiers. The ROC curve is a plot of the *sensitivity* on the y-axis and *1-specificity* on the x-axis. The *true positive rate* (TPR) (or sensitivity) is defined as the probability of correctly classifying an instance whose true label is 1; that is,

$$\text{TPR}(\delta) := \Pr[\delta(\boldsymbol{x}) = 1 \mid y = 1].$$

The *false positive rate* (FPR) (or 1-specificity) is defined as the probability of incorrectly classifying an instance as 1 when the true label is 0, that is,

$$\mathrm{FPR}(\delta) := \Pr[\delta(\boldsymbol{x}) = 1 \mid y = 0].$$

For the threshold based classification rule (3.2.1) the parameter θ determines the operating point of the classifier and corresponds to a point on the ROC plot with a specific $\mathrm{TPR}(\theta)$ and $\mathrm{FPR}(\theta)$. The ROC curve is obtained as θ is swept from $-\infty$ to ∞.

3.2.4 Area under the ROC curve

Naturally, classifiers that simultaneously have higher sensitivity and higher specificity are more desirable and dominate their competitors. In practice, however, one usually finds several classifiers with intersecting ROC curves. As a result, an additional criterion is often needed to decide among competing classifiers. One popular procedure selects the classifier with the highest area under its ROC curve ("area under the curve," or AUC). An excellent introduction to the topic is provided by Pepe (2003). The AUC is obtained by integrating the ROC curve; that is,

$$\mathrm{AUC} = \int_0^1 \mathrm{TPR}(\mathrm{FPR}^{-1}(t))dt.$$

Good classifiers have an AUC close to 1 while a random classifier has an AUC close to 0.5.

3.3 An AUC-maximizing recalibration

Many different classification procedures can be constructed based on the training data. The commonly used procedure in (3.2.1) is one such example which thresholds the raw scores; it has a corresponding AUC. It is intuitive and easy to use, but it may not be optimal in terms of maximizing AUC. Using the Neyman-Pearson Lemma, in this section we will show a recalibration procedure which maximizes the AUC under bi-normal population assumptions (this recalibration contains (3.2.1) as a special case).

3.3.1 Neyman-Pearson lemma and AUC

Let p_0 and p_1 be the class-conditional densities of the score $s = f(\boldsymbol{x})$ in class 0 and 1 respectively, that is,

$$\begin{aligned} p_0(s) &= \Pr[s|y=0] \quad \text{and} \\ p_1(s) &= \Pr[s|y=1]. \end{aligned}$$

The binary classification problem can be viewed in the framework of statistical hypothesis testing. Assigning a label $\{0, 1\}$ based on a score $s = f(\boldsymbol{x})$ is equivalent to deciding whether the score in question arose from the distribution p_0 (the *null hypothesis*) or the distribution p_1 (the *alternative hypothesis*). Clearly, it is desirable to have a decision procedure that combines a low rate of incorrectly rejecting the null hypothesis when it is in fact true (a low false positive rate) with a high rate of

accepting the alternative hypothesis when the null hypothesis is false (a high true positive rate).

The Neyman-Pearson lemma provides a way to properly balance these two competing goals. An especially thorough treatment of the Neyman-Pearson lemma and its many extensions appears in several chapters of Lehmann and Romano (2005). In the context of the classification problem above, the lemma states that for a fixed false positive rate α, a decision procedure maximizes the true positive rate if and only if it rejects the null hypothesis in favor of the alternative hypothesis for observed scores in the region (defined using the likelihood ratio)

$$C = \left\{ s \ : \ \frac{p_1(s)}{p_0(s)} \geq K_\alpha \right\}, \tag{3.3.1}$$

and does not reject the null hypothesis otherwise. Here K_α is the (1-α) quantile of the p_0 distribution.

The Neyman-Pearson lemma provides a way to construct procedures that maximize the true positive rate for each false positive rate. As a result, ROC curves of classifiers constructed using the Neyman-Pearson lemma must be above the ROC curves constructed using other methods; *classifiers constructed in this way therefore also have the highest AUC values*. In the next section, we use the Neyman-Pearson lemma to find a classification rule under the commonly used bi-normality assumption.

3.3.2 Bi-normality assumption for the scores

We will assume that the scores for the positive and negative populations are normally distributed. As discussed in the next section, this is a reasonable assumption for scores constructed as the linear combination of many covariates (features); the non-equal variance assumption is illustrated in the restaurant tips data of the next section as well. The score s has a separate normal distribution corresponding to $y = 1$ and $y = 0$, that is,

$$\begin{aligned} p_0(s) &= \Pr[s|y=0] = \mathcal{N}(s \mid \mu_0, \sigma_0^2) \quad \text{and} \\ p_1(s) &= \Pr[s|y=1] = \mathcal{N}(s \mid \mu_1, \sigma_1^2), \end{aligned}$$

where $\mathcal{N}(s \mid \mu, \sigma^2)$ is the normal distribution with mean μ and variance σ^2. Without loss of generality we further assume that $\mu_1 > \mu_0$.

3.3.3 Quadratic score based thresholding

Under the bi-normality assumption the rejection region as specified in (3.3.1) by the Neyman-Pearson lemma can be written as

$$C = \left\{ s \ : \ \frac{\mathcal{N}(s \mid \mu_0, \sigma_0^2)}{\mathcal{N}(s \mid \mu_1, \sigma_1^2)} \geq K_\alpha \right\}. \tag{3.3.2}$$

Since logarithm is a monotonic transformation, this expression can be rewritten in terms of the log-likelihood ratio.

$$C = \left\{ s \ : \ \log \frac{\mathcal{N}(s \mid \mu_0, \sigma_0^2)}{\mathcal{N}(s \mid \mu_1, \sigma_1^2)} \geq \log K_\alpha \right\}. \tag{3.3.3}$$

Simplifying (3.3.3) yields the following decision rule

$$\delta(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \ as^2 + bs + c \geq \theta \\ 0 & \text{otherwise} \end{cases}, \tag{3.3.4}$$

where a, b, and c are defined as,

$$\begin{aligned} a &= \frac{1}{2}\left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}\right), \\ b &= \frac{\mu_1}{\sigma_1^2} - \frac{\mu_0}{\sigma_0^2}, \quad \text{and} \\ c &= \log\left(\frac{\sigma_0}{\sigma_1}\right) + \frac{1}{2}\left(\frac{\mu_0}{\sigma_0^2} - \frac{\mu_1}{\sigma_1^2}\right), \end{aligned} \tag{3.3.5}$$

θ is a threshold determined by the desired false positive rate, and $s = f(\boldsymbol{x})$ is the classifier score. As described in Section 3.1, under the bi-normality assumptions, the Neyman-Pearson lemma guarantees that the ROC for thresholding the recalibrated scores is above all other ROC curves, since the true positive rate is maximized for each false positive rate. In other words, under the bi-normal assumptions, the quadratic classification rule (3.3.4) attains the maximum AUC.

3.3.4 Discussion

1. ***The case of equal variances*** When $\sigma_0 = \sigma_1 = \sigma$ the constants evaluate to

$$\begin{aligned} a &= 0, \\ b &= \frac{\mu_1 - \mu_0}{\sigma^2}, \quad \text{and} \\ c &= \frac{1}{2}\left(\frac{\mu_0 - \mu_1}{\sigma^2}\right). \end{aligned}$$

   The decision rule simplifies to

$$\delta(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \; bs + c \geq \theta \\ 0 & \text{otherwise} \end{cases},$$

   which is equivalent to the linear classification rule (3.2.1). Hence the commonly used linear classification rule (3.2.1) and the proposed quadratic rule (3.3.4) agree when the standard deviation of the two normal distributions are equal, that is, $\sigma_0 = \sigma_1 = \sigma$.

2. ***Rule after raw score recalibration uses second moments***

   The linear thresholding rule (3.2.1) is monotonic in s and implicitly means that a larger score $s = f(\boldsymbol{x})$ provides a larger chance of $y = 1$. In contrast, the quadratic thresholding rule decides that an instance is positive if the score is quite high (greater than a certain threshold) or quite low (less than a certain threshold). This may appear counterintuitive, yet the Neyman-Pearson lemma guarantees that this is indeed the best choice under the bi-normality

assumption.

The linear classification rule uses only the first moments of the scores; that is, it uses the fact that the means are different. The quadratic rule also captures the second moments by assuming that both the means and the variances are different. For example, if we know that the positive class has a much larger variance than the negative class, then a very small score, though intuitively negative, is very unlikely to have come from the negative distribution since its variance is much less than that of the positive population distribution (we show a real-world use-case of this phenomenon in the next section). In the same spirit, classification rules using higher order moments could be designed. The quadratic recalibration is optimal under the bi-normality assumption, while the linear rule is optimal when the variances of both the distributions are equal.

3. ***Quadratic transformation to improve AUC*** We have shown that a classification rule that is quadratic in s dominates the commonly used classifier in (3.2.1) when the variance of s for the positive population is different from the variance of s for the negative population. Hence a very simple method to improve the score from any general classifier is to transform the scores using the quadratic transformation $s' \leftarrow as^2 + bs + c$. The Neyman-Pearson lemma guarantees that the AUC for scores s' will be greater than or equal to the AUC of s and, in fact, that the ROC for the quadratic rule dominates the ROC of the linear thresholding rule.

4. ***Estimating a, b, and c from training data*** In practice the constants a, b, and c can be estimated from the training data by plugging in into (3.3.5) the empirical estimates for the population means $\hat{\mu}_1$ and $\hat{\mu}_0$ and the standard deviation $\hat{\sigma}_1$ and $\hat{\sigma}_0$ using the positive and negative class examples respectively.

5. ***Parametric ROC and AUC*** For the bi-normal assumption it is possible to derive an analytical expression for the ROC and the AUC of the quadratic scoring rule. For any threshold θ the TPR and the FPR can be written as

$$\text{TPR}(\theta) = \Pr[s \geq \theta \mid y = 1] = \Phi\left(\frac{\mu_1 - \theta}{\sigma_1}\right),$$

$$\text{FPR}(\theta) = \Pr[s \geq \theta \mid y = 0] = \Phi\left(\frac{\mu_0 - \theta}{\sigma_0}\right).$$

where $\Phi(x) = (1/\sqrt{2\pi}) \int_{-\infty}^{x} \exp(-t^2/2)dt$ is the cdf of the standard normal distribution. Hence for a particular $FPR(\theta) = t$, we can write $\theta = \mu_0 - \sigma_0\Phi^{-1}(t)$ and hence

$$\text{ROC}(t) = \text{TPR}(t) = \Phi(A + B\Phi^{-1}(t)),$$

where we define $A = (\mu_1 - \mu_0)/\sigma_1$ and $B = \sigma_0/\sigma_1$. The term A is called the *intercept* and B the *slope* of the binormal ROC curve. By integrating this expression the AUC for the binormal ROC curve is given by

$$\text{AUC} = \Phi\left(\frac{\mu_1 - \mu_0}{\sqrt{\sigma_1^2 + \sigma_0^2}}\right).$$

6. ***Equivalence to the optimal Bayes rule*** The optimal classifier is the Bayes rule given by

$$\delta(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \log \frac{\Pr[y=1|s]}{\Pr[y=0|s]} \geq 0 \\ 0 & \text{otherwise} \end{cases},$$

Let π_0 and π_1 be the prior probability of class 0 and 1 respectively, that is, $\pi_0 = \Pr[y=0]$ and $\pi_1 = \Pr[y=1]$. From Bayes theorem we have the following posterior for the positive class

$$\Pr[y=1|s] = \frac{p_1(s)\pi_1}{p_0(s)\pi_0 + p_1(s)\pi_1}.$$

Using this under the earlier bi-normal assumptions, the Bayes rule simplifies to the rule with quadratic recalibration obtained earlier; that is,

$$\delta(\boldsymbol{x}) = \begin{cases} 1 & \text{if } as^2 + bs + c \geq 0 \\ 0 & \text{otherwise} \end{cases},$$

where the constants a and b remain the same as defined earlier (3.3.5) but the parameter c gets modified to include the prior class probabilities.

$$c \;=\; \log\left(\frac{\sigma_0\pi_1}{\sigma_1\pi_0}\right) + \frac{1}{2}\left(\frac{\mu_0}{\sigma_0^2} - \frac{\mu_1}{\sigma_1^2}\right),$$

While the Bayes rule implicitly defines the optimal threshold, in principle, we can vary the threshold to get the ROC curve.

3.4 Illustrations and Empirical Evaluation

As discussed in the introduction, the literature is full of examples in which bi-normal assumptions are not appropriate at all. The idea that cases with the most negative scores may be classified as positive to increase the AUC seems rather counterintuitive; the result comes from the fact that the variances of the classes are unequal, so that, for example, the most negative scores are actually more likely to come from the positive class distribution. Simulated, bi-normal data with different class variances shows how the recalibration method can be applied, but a natural question is whether the underlying assumptions (bi-normality and unequal variances) which make the recalibration work are ever appropriate in practice. Here, we present a real-world classification example using data in which the proposed recalibration increases the AUC by a full 25%.

Among other variables, this data set contains the total bill amounts and tip amounts paid at an American restaurant by 244 patrons (or groups of patrons) and on whether there were smokers among the patrons. Interestingly, it turns out that the variability in the tip percentage (defined as tip amount divided by total bill amount without the tip) is much higher for the 93 patron groups with smokers than for the 151 patron groups without smokers. The average tip percentage is close to 15 % for both groups (not surprisingly, as the customary tipping rate at American restaurants is around 15%), with more patrons tipping below 15 % than above for both smokers and non-smokers. Our goal is to try to classify the patrons as smokers or non-smokers based on their tip percentage.

To construct the raw scores, we take the logarithm of the tip percentages. This initial (and monotonic) transformation helps to alleviate the skew due to the "undertipping" behavior of most of the patrons and makes the normality assumptions more appropriate. The mode symmetry assumptions are not unreasonable; the variance of the scores for smokers is 0.203, while the variance for the non-smokers is only 0.073, or roughly 1/3 that of the smokers' scores. The score distributions overlap, making classification difficult. In using the threshold classifiers, however, it is seen that classification using the recalibrated (quadratic) raw scores gives an AUC that is 25% ($= \frac{0.65-0.52}{0.52}$) higher than classification with the raw scores In other words, the using the recalibrated gives a much better classifier in terms of AUC than using the raw scores, where the AUC of 0.52 is almost as bad as random guessing.

3.5 Conclusions and Proposed Extensions

We used the Neyman-Pearson lemma to show that a popular classification procedure based on scoring can be made better in terms of the AUC criterion when the underlying populations have different variances. We proposed a quadratic recalibration which maximizes the AUC and contains the usual procedure based on raw scores as a special case when the population variances are equal. Our results are based on the bi-normal population assumption for the scores, which can be appropriate in many real-world settings. The increase in AUC grows as the difference in the variances of the two populations increases, with an increase of 25 % recorded for the Restaurant Patron Tipping data in our illustration and modest improvements in AUC for other

common reference datasets. We hope to extend our work by investigating the procedure for data sets in which the scores are sample averages from samples of various sizes, as this is a natural setting for normal scores with unequal variances.

Chapter 4

Conclusion

In this dissertation, we have explored three specific areas of Bayesian classification procedures. The first chapter focused on a new classification procedure using a nonparametric mixture prior distribution and empirical Bayes techniques to minimize a loss function that applies to many scientific settings. The second chapter turns to a popular criterion for evaluating classifiers, the false discovery rate, and gives a way of estimating Bayesian versions, the pFDR and local false discovery rate, using a nonparametric mixture prior. In the last chapter, we look at the AUC criterion in classification problems with normal observations, which can arise frequently when many covariates are combined into summary classification scores through averaging or regression techniques.

There are many interesting questions in the field of our work that can be explored further. For example, better ways of controlling local false discovery rates, and not just the FDR, can be useful. The sense in which an error rate is controlled is also

open for additional work because current techniques focus on providing bounds on expected error rate values, while in applications, more attention to sample-specific statements may also be needed. Work by Jin and Cai (2007) suggests that it may be possible to make the techniques proposed in Chapter 2 of this dissertation more general by providing estimates of the noise distribution because misspecification error can lead to inaccurate estimates of local false discovery rates. It would also be interesting to extend local FDR techniques to interaction effects in model selection in ways similar to the hierarchical FDR model proposed by Yekutieli (2008).

Bibliography

[1] Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1): 289–300, 1995.

[2] Benjamini, Y., Krieger, A.M., and Yekutieli, D. Adaptive linear step-up False Discovery Rate controlling procedures. *Biometrika*, 93(3): 491–507, Sep 2006.

[3] Benjamini Y. and Yekutieli, D. The control of the False Discovery Rate in multiple testing under dependency. *The Annals of Statistics*, 29(4): 1165–1188, 2001.

[4] Brown, L.D. Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *The Annals of Mathematical Statistics*, 42(3): 855-903, 1971.

[5] Cai, T. and Jin, J. Optimal rates of convergence for estimating the null density and proportion of non-null effects in large-scale multiple testing. *The Annals of Statistics*, 38 : 100–145, 2010.

[6] Duncan, D.B. A Bayesian approach to multiple comparisons. *Technometrics*, 7: 171–222, 1965.

[7] Efron B., Tibshirani R., Storey J.D., and Tusher V. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96 (456): 1151–1160, 2001.

[8] Efron, B. Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypohtesis. *Journal of the American Statistical Association*, 99, 96-104, 2004.

[9] Efron, B. Local false discovery rates. Technical report. Division of Biostatistics, Stanford University, 2005.

[10] Efron, B. Empirical Bayes modeling, computation, and accuracy. Technical report. Stanford University, 2013.

[11] Efron, B. and Morris, C. Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70: 311-319, 1975.

[12] Gasch A. P, Spellman P.T., Kao C.M, Carmel-Harel O., Eisen M.B., Storz G., and Botstein D. Genomic expression progams in the response of yeast cells to environmental changes. *Molecular Biology of the Cell*, 11: 4241–4257, 2000.

[13] George, E. I. and Foster, D. P. Calibration and empirical Bayes variable selection. *Biometrika*, 87(4): 731–747, 2000.

[14] Herschtal, A. and Raskutti, B. Optimizing area under the ROC curve using gra-

dient descent. *Proceedings of the twenty-first international conference on machine learning*, 49–, 2004.

[15] Jin, J. and Cai, T. Estimating the null and the proportion of non-null effects in large-scale multiple comparisons. *Journal of the American Statistical Association*, 102 495-506, 2007.

[16] Johnstone I.M. and Silverman B.W. Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. *The Annals of Statistics*, 32(4): 1594–1649, 2004.

[17] Lehmann, E. and Romano, J. *Testing Statistical Hypotheses*. Springer Texts in Statistics, 2005.

[18] Pepe, M. An interpretation for the ROC curve and inference using GLM procedures. *Biometrics*, 56: 352–359, 2000.

[19] Pepe, M. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, 2003.

[20] Raykar, V.C. and Zhao, L.H. Nonparametric prior for adaptive sparsity. *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*. 629–636, 2010.

[21] Raykar, V.C. and Zhao, L.H. Empirical Bayesian thresholding for sparse signals using mixture loss functions. *Statistica Sinica*, 21 449–474, 2011.

[22] Scott, J.G. and Berger, J.O. An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference*, 136: 2144-2162, 2006.

[23] Storey, J.D. A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B*, 64(3): 479–498, 2002.

[24] Storey, J.D. The positive false discovery rate: a Bayesian interpretation and the q-value. *The Annals of Statistics*, 31(6): 2013–2035, 2003.

[25] Wand, M.P. and Jones, M.C. *Kernel Smoothing*. Chapman and Hall/CRC, 1995.

[26] Yekutieli, D. Hierarchical False Discovery Rate controlling methodology. *Journal of the American Statistical Association*, 103 (481): 309–316, 2008.

[27] Zhang, C.-H. Empirical Bayes and compound estimation of normal means. *Statistica Sinica*, 7: 181–193, 1997.