Hunting for Significance: Bayesian Classifiers under a Mixture Loss Function

Igar Fuki, Lawrence Brown, Xu Han, Linda Zhao

February 13, 2014

Abstract

Detecting significance in a high-dimensional sparse data structure has received a large amount of attention in modern statistics. In the current paper, we introduce a compound decision rule to simultaneously classify signals from noise. This procedure is a Bayes rule subject to a mixture loss function. The loss function minimizes the number of false discoveries while controlling the false non discoveries by incorporating the signal strength information. Based on our criterion, strong signals will be penalized more heavily for non discovery than weak signals. In constructing this classification rule, we assume a mixture prior for the parameter which adapts to the unknown sparsity. This Bayes rule can be viewed as thresholding the "local fdr" (Efron 2007) by adaptive thresholds. Both parametric and nonparametric methods will be discussed. The nonparametric procedure adapts to the unknown data structure well and outperforms the parametric one. Performance of the procedure is illustrated by various simulation studies and a real data application.

Keywords: High dimensional sparse inference, Bayes classification rule, Nonparametric estimation, False discoveries, False nondiscoveries

1 Introduction

Consider a normal mean model:

$$Z_i = \beta_i + \epsilon_i, \qquad i = 1, \cdots, p \tag{1}$$

where $\{Z_i\}_{i=1}^n$ are independent random variables, the random errors $(\epsilon_1, \cdots, \epsilon_p)^T$ follow a multivariate normal distribution $N_p(0, \sigma^2 \mathbf{I}_p)$, and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ is a p-dimensional unknown vector. For simplicity, in model (1), we assume σ^2 is known. Without loss of generality, let $\sigma^2 = 1$. (Our interest is in settings with large p and significant sparsity. In such settings σ^2 can be quite accurately estimated. So the assumption that σ^2 is known is not a restrictive limitation.) The central problem is to make inference about $\boldsymbol{\beta}$ based on the observations $\{z_i\}_{i=1}^n$. For example, much attention has focused on estimating the unknown $\boldsymbol{\beta}$. James & Stein (1961) showed that shrinkage type estimators performs better than the conventional maximum likelihood estimator under the squared error loss function.

In modern scientific applications, $\boldsymbol{\beta}$ is usually a high dimensional sparse vector with a large proportion of elements as 0. For example, in genome-wide association studies, scientists are interested in finding out which SNPs are associated with some observable traits (e.g. weight, blood pressure, some disease). The number of SNPs can be tens of thousands or even millions. However, a majority of the SNPs does not have any association with the response variable. This scientific problem can be formularized as the normal mean model in (1). Various applications of model (1) in high dimensional setting also occur in model selection in machine learning (George & Foster 2000), smoothing/de-noising in signal processing (Johnstone & Silverman 2004) and significance testing in genomics (Efron & Tibshirani 2007). The nonzero elements in $\boldsymbol{\beta}$ are called "signals", and the zero ones are called "noise". In practice, one does not know which elements are signals. Furthermore, the proportion of signals is also usually unknown. Currently a predominant statistical interest involves detecting which coordinates are signals and which are noise.

In the current paper, we will introduce a compound decision rule to simultaneously detect signals in the high dimensional sparse setting. In the decision making process, the mistakes can be classified as two types: false discoveries and false non discoveries. In practice, it is usually more severe that a strong signal is undetected than an weak signal. Our decision rule will aim to minimize the total number of false discoveries while controlling the false non discoveries with a criterion which incorporates the signal strength information. The classification procedure is a Bayes rule under a mixture loss function. It turns out that our Bayes rule can be viewed as thresholding the "local fdr" (Efron 2007) by adaptive thresholds. The mixture prior for $\boldsymbol{\beta}$ is further investigated both parametrically and non parametrically. We will show that the nonparametric Bayes procedure adapts well to the unknown sparsity and the underlying data structure.

The rest of the paper is organized as follows: In section 2, we introduce a mixture loss function, a mixture prior for $\boldsymbol{\beta}$ and the Bayes classification rule. Both parametric and nonparametric procedures will be discussed. Section 3 contains various simulation studies to illustrate the performance of our method. In section 4, the method is applied to a publicly available gene expression dataset.

2 Method

Classification between the signals and the noise in (1) can be modeled by the following hypothesis testing problem:

$$H_{i0} : \beta_i = 0 \qquad \text{vs} \qquad H_{i1} : \beta_i \neq 0, \qquad i = 1, \cdots, p.$$

We will introduce a compound decision rule. If we claim $\beta_i \neq 0$, let $a_i = 1$, otherwise let $a_i = 0$. The decision vector $\mathbf{a} = (a_1, \cdots, a_p)$ describes our classification rule to distinguish signals from noise. The oracle procedure is that $a_i = 1$ when $\beta_i \neq 0$ and $a_i = 0$ otherwise. In applications there may be false discoveries, that is, $a_i = 1$ while $\beta_i = 0$. In practice, we want to control the total number of false discoveries. On the other hand, one can also make false non-discoveries, that is, $a_i = 0$ while $\beta_i \neq 0$. Instead of directly controlling the total number of false non-discoveries, we formulate a loss function that only mildly penalizes false non discoveries when the true value of β_i is not far away from zero. Combining the intuition from the above argument, we consider the following mixed loss function:

$$L(\mathbf{a}, \boldsymbol{\beta}) = \sum_{i=1}^{p} \left[a_i \mathbf{I}_{\beta_i=0} + (1-a_i) c \beta_i^2 \right]. \tag{2}$$

The first part $L_1 = \sum_{i=1}^{p} a_i \mathbf{I}_{\beta_i=0}$ is the total number of false discoveries. The second part $L_2 = \sum_{i=1}^{p} (1-a_i)\beta_i^2$ is the squared error loss if $a_i = 0$ but $\beta_i \neq 0$. The tuning parameter c balances L_1 and L_2, and the effect of c will be explored through simulations studies in later sections. To minimize loss function $L(\mathbf{a}, \boldsymbol{\beta})$, it is equivalent to minimizing the total false discoveries L_1 subject to the constraint $L_2 \leq A$ where A is some predetermined constant. This mixed loss function is more likely to detect the strong signals than the weak signals since L_2 will be penalized more heavily when β_i^2 is large. It is worth mentioning that the loss function (2) is designed to find a classification rule $\mathbf{a}$. We are not proposing to estimate the signal strength of β by $\mathbf{a}$.

In large scale multiple testing, Sun & Cai (2007) also introduced a compound decision rule to detect signals from noise with the following loss function:

$$L^*(\mathbf{a}, \boldsymbol{\beta}) = \sum_{i=1}^{p} \left[c a_i \mathbf{I}_{\beta_i=0} + (1-a_i) \mathbf{I}_{\beta_i \neq 0} \right]. \tag{3}$$

The second part $L_2^* = \sum_{i=1}^{p} (1-a_i)\mathbf{I}_{\beta_i \neq 0}$ is the total number of false non-discoveries. To minimize $L^*(\mathbf{a}, \boldsymbol{\beta})$, it is equivalent to minimize L_2^* subject to controlling false discoveries at a particular level. This loss function treats the nonzero signals with equal weight while ignoring the actual signal strength. The loss function (2) is different from (3) by incorporating the signal strength information for the false non discoveries.

A related explanation for loss function (2) is based on different penalty structures for each of the classification mistakes. The cost of saying that $\beta_i \neq 0$ when it is in fact equal to 0 is constant (and normalized to be 1). On the other hand, the cost of saying that $\beta_i = 0$ when it is nonzero 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" ($\beta_i \neq 0$) in the initial screening step and costs proportional to the magnitude of the differential expression for failing to make a discovery. Scott & Berger (2006) briefly discussed a similar loss function based on the above argument from the cost perspective. Instead of considering the squared error loss for L_2 term in (2), they assume that the cost for a false non discovery is proportional to the absolute value of its magnitude. Comparing with Scott & Berger (2006), our loss function magnifies the effect of false non discovery when $|\beta_i| > 1$ and shrinks the effect when $|\beta_i| < 1$. Furthermore, based on loss function (2), the threshold of the Bayes classification procedure introduced in section 2.2 is an explicit expression involving the marginal densities $g(z_i)$.

Muller, Parmigiani & Rice (2006) also realized that false non-discoveries should not be treated equally. However, the class of loss functions L_m that they consider while related to our loss does not include ours as a special case. Further, they do not construct explicit nonparametric empirical Bayes procedures as we do later in our paper.

In Benjamini & Hochberg (1997), they considered a weighted multiple testing framework, where they maximized a weighted number of rejections with controlling a weighted number of false rejections. More specifically, maximize $E(\sum_{i=1}^p b_{i1}a_i)$ such that $E(\sum_{i=1}^p b_{i2}a_i\mathbf{I}_{\beta_i \neq 0}) \leq \alpha$, where $\{b_{i1}\}_{i=1}^p$ and $\{b_{i2}\}_{i=1}^p$ are two sets of positive weights satisfying $\sum_{i=1}^p b_{i1} = \sum_{i=1}^p b_{i2} = p$. This criterion assigned different weights for different hypothesis, but it did not give any detailed suggestions on how to assign these weights. Furthermore, maximizing number of rejections is one way to increase the power of the test, but it is still different from minimizing the number of false non discoveries.

In the classic decision theoretic framework, a Bayes rule is to find $\mathbf{a}$ which minimizes the expectation of the loss function (2) with respect to the posterior distribution of $\boldsymbol{\beta}$ conditional on the observation. Before we introduce our Bayes rule $\mathbf{a}$, we first propose a prior structure for the unknown high-dimensional sparse vector $\boldsymbol{\beta}$ and calculate the corresponding posterior distributions.

2.1 Mixed Prior and Related Results

Since a large proportion of $\{\beta_i\}$ are zero, but this proportion is unknown, we will assume a mixed prior for $\boldsymbol{\beta}$:

$$p(\beta_i|w,\gamma) = w\delta_i + (1-w)\gamma(\beta_i). \tag{4}$$

In (4), w is the unknown proportion of zero elements in $\boldsymbol{\beta}$, δ_i is the probability mass 1 at point 0, $\gamma(\beta_i)$ is the prior density for the nonzero β_i. With the mixed structure of (4), we assume that $\beta_i = 0$ with probability w and follows density $\gamma(\beta_i)$ with probability $1-w$. It is worth mentioning that we do not assume w to necessarily be very close to 1 here. Our estimate of w in sections 2.3 and 2.4 adapts to the unknown sparsity. In this section, we will not give any detailed expressions for $\gamma(\beta_i)$. In sections 2.3 and 2.4, for deriving the Bayes rule $\mathbf{a}$, both parametric and nonparametric methods will be considered with respect to the mixture prior (4).

The mixture structure (4) has often been favored by statisticians to model the sparsity in $\boldsymbol{\beta}$. Johnstone & Silverman (2004) first proposed this structure and assumed parametric priors for $\gamma(\beta_i)$, e.g. the normal prior or the Laplace prior. These will be discussed in section 2.3 of the current paper. Later, Raykar & Zhao (2010) considered an unspecified distribution for the nonzero part of (4) and proposed a nonparametric empirical Bayes method to estimate $\boldsymbol{\beta}$. Brown & Greenshtein (2009) applied a nonparametric prior for $\boldsymbol{\beta}$ directly but without the mixture structure in (4). See their papers for additional references.

The likelihood function of the observations $\boldsymbol{z}$ given $\boldsymbol{\beta}$ can be expressed as follows:

$$p(\boldsymbol{z}|\boldsymbol{\beta}) = \prod_{i=1}^{p} p(z_i|\beta_i). \tag{5}$$

The posterior distribution of $\boldsymbol{\beta}$ given the data $\boldsymbol{z}$, the hyper parameter w and the non zero prior γ is given by the Bayes formula:

$$p(\boldsymbol{\beta}|w,\gamma,\boldsymbol{z}) = \frac{\prod_{i=1}^{p} p(z_i|\beta_i)p(\beta_i|w,\gamma)}{m(\boldsymbol{z}|w,\gamma)}, \tag{6}$$

where

$$m(\boldsymbol{z}|w,\gamma) = \prod_{i=1}^{p} \int p(z_i|\beta_i)p(\beta_i|w,\gamma)d\beta_i = \prod_{i=1}^{p} m(z_i|w,\gamma) \tag{7}$$

is the marginal distribution of the data given the hyper parameter and γ. Let $N(z|\mu,\sigma^2)$ denote the normal density with indicated mean and variance. Then

$$m(z_i|w,\gamma) = wN(z_i|0,1) + (1-w)g(z_i), \tag{8}$$

where

$$g(z_i) = \int N(z_i|\beta_i,1)\gamma(\beta_i)d\beta_i \tag{9}$$

is the marginal density of z_i given that β_i is nonzero. The posterior in (6) can be factored as $p(\boldsymbol{\beta}|\boldsymbol{z},w,\gamma) = \prod_{i=1}^{p} p(\beta_i|z_i,w,\gamma)$ with

$$p(\beta_i|z_i,w,\gamma) = p_i\delta(\beta_i) + (1-p_i)G(\beta_i), \tag{10}$$

where

$$p_i = p(\beta_i = 0|z_i,w,\gamma) = \frac{wN(z_i|0,1)}{wN(z_i|0,1) + (1-w)g(z_i)} \tag{11}$$

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

$$G(\beta_i) = \frac{N(z_i|\beta_i,1)\gamma(\beta_i)}{\int N(z_i|\beta_i,1)\gamma(\beta_i)d\beta_i} \tag{12}$$

is the posterior density of β_i when $\beta_i \neq 0$. Note that the posterior p_i is exactly the "local fdr" defined in Efron (2007). Our classification rule in section 2.2 will be constructed based on p_i, and both parametric and nonparametric estimation methods will be applied to estimate G in (10)-(12).

2.2 Bayes Classification Procedure

The above expressions yield the Bayes rule under the mixture loss function (2) with respect to the mixture prior $p(\beta_i|w,\gamma)$. Note that the Bayes rule is to minimize $E\big[L(\mathbf{a},\boldsymbol{\beta})|\mathbf{Z}\big]$ where the expectation is for the posterior distribution of $\boldsymbol{\beta}$ given $\mathbf{Z}$. Here is the formal statement.

Theorem 1. *Denoting the ith component of the Bayes rule by a_i^{Bayes} and the second moment of β_i under $G(\beta_i)$ in (12) by $E_G(\beta_i^2)$, the rule is $a_i^{Bayes} = 1$ if $p_i < \frac{cE_G(\beta_i^2)}{1+cE_G(\beta_i^2)}$ and equals zero otherwise, where p_i and $G(\beta_i)$ are given in (11) and (12) respectively.*

Proof of Theorem 1: Note that the conditional expectation of the mixture loss (2) can be minimized component-wise. For the ith component of the decision vector, if $a_i = 1$, then

$$E[L(1,\beta_i)|z_i] = \int L(1,\beta_i)p(\beta_i|z_i)d\beta_i = p_i, \tag{13}$$

and if $a_i = 0$, then

$$E[L(0,\beta_i)|z_i] = \int L(0,\beta_i)p(\beta_i|z_i)d\beta_i = c(1-p_i)E_G(\beta_i^2). \tag{14}$$

The Bayes rule is $a_i^{Bayes} = 1$ when $E[L(1,\beta_i)|z_i] \leq E[L(0,\beta_i)|z_i]$. This is exactly the condition given in Theorem 1. The proof is now complete.

To better understand the effect of the tuning parameter (cost constant) c, we consider some extreme cases. When $c \to \infty$, we want to minimize the number of false non discoveries without consideration of false discoveries. To achieve this, we claim all β_i as signals. This is consistent with our Bayes rule where the threshold goes to 1 and each β_i will be detected as nonzero element. When $c \to 0$, we just want to minimize the number of false discoveries, so we claim all β_i as noise. On the other hand, our Bayes rule does not detect any signals as the threshold goes to 0.

Recall that in Scott & Berger (2006), they considered the absolute value $|\beta_i|$ in the loss function (2) instead of our β_i^2. Therefore, their Bayes rule involves $E_G(|\beta_i|)$ instead of our $E_G(\beta_i^2)$. There is no closed form expression for $E_G(|\beta_i|)$ and this term has to be evaluated numerically through some integration involving the marginal densities $g(z_i)$. Compared with Scott & Berger (2006), the following Theorem 2 further shows that the threshold in our Bayes rule can be conveniently and explicitly expressed in terms of the marginal densities $g(z_i)$.

Theorem 2. *Under mild conditions in Brown (1971), the $E_G(\beta_i^2)$ in the Bayes rule thresh-*

olding condition can be written as

$$E_G(\beta_i^2) = z_i^2 + 1 + \frac{g''(z_i)}{g(z_i)} + 2z_i\frac{g'(z_i)}{g(z_i)}. \tag{15}$$

The Bayes classification rule will be defined correspondingly.

Proof of Theorem 2: Background for the following can be found in Brown (1986). To show (15), we note that the derivative in

$$\frac{\partial^2}{\partial z_i^2}\, g(z_i) \;\;=\;\; \frac{\partial^2}{\partial z_i^2}\int N(\beta_i|z_i,1)\gamma(\beta_i)d\beta_i$$

can be taken inside the integral because the density $N(\beta_i|z_i,1)$ comes from an exponential family. Then

$$\begin{aligned}\frac{\partial^2}{\partial z_i^2}\, g(z_i) \;\;&=\;\; \int \frac{\partial^2}{\partial z_i^2}[\frac{1}{\sqrt{2\pi}}e^{-\frac{(\beta_i-z_i)^2}{2}}]\gamma(\beta_i)d\beta_i \\ &=\;\; \int \frac{\partial}{\partial z_i}(\beta_i - z_i)[\frac{1}{\sqrt{2\pi}}e^{-\frac{(\beta_i-z_i)^2}{2}}]\gamma(\beta_i)d\beta_i \\ &=\;\; \int \frac{\partial}{\partial z_i}(\beta_i)[\frac{1}{\sqrt{2\pi}}e^{-\frac{(\beta_i-z_i)^2}{2}}]\gamma(\beta_i)d\beta_i \\ &\quad\; -\int \frac{\partial}{\partial z_i}[(z_i)[\frac{1}{\sqrt{2\pi}}e^{-\frac{(\beta_i-z_i)^2}{2}}]]\gamma(\beta_i)d\beta_i.\end{aligned}$$

Iterating the integration above to compute the derivative in the first term and using the product rule for the derivative in the second term, we see that the expression above is

$$\begin{aligned}&= \;\int(\beta_i)(\beta_i - z_i)[\frac{1}{\sqrt{2\pi}}e^{-\frac{(\beta_i-z_i)^2}{2}}]\gamma(\beta_i)d\beta_i \\ &\quad -\Big(\int[\frac{1}{\sqrt{2\pi}}e^{-\frac{(\beta_i-z_i)^2}{2}}]\gamma(\beta_i)d\beta_i + \int z_i(\beta_i - z_i)[\frac{1}{\sqrt{2\pi}}e^{-\frac{(\beta_i-z_i)^2}{2}}]\gamma(\beta_i)d\beta_i.\Big)\end{aligned}$$

Collecting terms,

$$\begin{aligned}\frac{\partial^2}{\partial z_i^2}\, g(z_i) \;\;&=\;\; \int \beta_i^2 N(\beta_i|z_i,1)\gamma(\beta_i)d\beta_i - z_i\int \beta_i N(\beta_i|z_i,1)\gamma(\beta_i)d\beta_i \\ &-\;\; g(z_i) - z_i\int \beta_i N(\beta_i|z_i,1)\gamma(\beta_i)d\beta_i + z_i^2 g(z_i).\end{aligned}$$

Dividing both sides by $g(z_i)$,

$$\frac{g''(z_i)}{g(z_i)} = \frac{\int \beta_i^2 N(\beta_i|z_i,1)\gamma(\beta_i)d\beta_i}{g(z_i)} - 2z_i \frac{\int \beta_i N(\beta_i|z_i,1)\gamma(\beta_i)d\beta_i}{g(z_i)} - 1 + z_i^2.$$

Note that

$$\frac{\int \beta_i N(\beta_i|z_i,1)\gamma(\beta_i)d\beta_i}{g(z_i)} = z_i + \frac{g'(z_i)}{g(z_i)}$$

as in Brown (1971) and Robbins (1956).

Plug in and rearrange terms to get

$$\frac{\int \beta_i^2 N(\beta_i|z_i,1)\gamma(\beta_i)d\beta_i}{g(z_i)} = \frac{g''(z_i)}{g(z_i)} + z_i^2 + 2z_i \frac{g'(z_i)}{g(z_i)} + 1.$$

The proof is now complete.

The Bayes classification procedure defined in Theorem 1 is based on the hyper parameter w and the nonzero prior γ. The following sections discuss estimation of w and γ in various data settings.

2.3 Parametric Prior for γ

Practical Bayesian modeling often assumes some certain parametric density functions for γ. Popular choices are as follows:

1 A normal prior $N(\theta, \tau^2)$

$$\gamma(\beta_i) = N(\beta_i|\theta, \tau^2) = (2\pi\tau^2)^{-1/2} \exp\{-\frac{(\beta_i - \theta)^2}{2\tau^2}\}.$$

2 A double exponential (Laplace) with scale parameter a

$$\gamma(\beta_i) = 0.5a \exp\{-a|\beta_i|\}.$$

The Laplace prior has heavier tails than the normal prior. Details are referred to Johnstone & Silverman (2004). Laplace prior with shifted location is also a possible choice, which can also be implemented by the following procedure. To simplify the discussion, we only consider the Laplace prior in Johnstone & Silverman (2004). In both cases, the marginal density of z_i conditional on the hyper parameters can be given analytically. For the normal prior, since $Z_i|\beta_i \sim N(\beta_i, 1)$ and $\beta_i|\theta, \tau^2 \sim N(\theta, \tau^2)$, we have $Z_i|\theta, \tau^2 \sim N(\theta, 1+\tau^2)$. For the double exponential prior,

$$\begin{aligned} g(z_i|a) &= \int N(z_i|\beta_i, 1)\gamma(\beta_i|a)d\beta_i \\ &= 0.5a\exp(\frac{a^2}{2})\Big[\exp(-az_i)\Phi(z_i-a)+\exp(az_i)\Phi(-z_i-a)\Big]. \end{aligned}$$

Suppose the marginal of the data given the hyper parameters is $g(z_i|b)$, where for the normal prior $b = (\theta, \tau^2)$ and for the double exponential prior $b = a$. Then the log-marginal likelihood is

$$\log m(\boldsymbol{z}|w, b) = \sum_{i=1}^{p} \log\Big[wN(z_i|0,1) + (1-w)g(z_i|b)\Big]. \tag{16}$$

We can find suitable estimates of w and b by maximizing the log-marginal likelihood

$$(\widehat{w}, \widehat{b}) = argmax_{w,b} \log m(\boldsymbol{z}|w, b).$$

Feasible computation proceeds in stages: for a fixed w, we find the b which maximizes $\log m(\boldsymbol{z}|w, b)$; then given the best a, we maximize with respect to w. We repeat this process until convergence. The Bayes classification procedure in Theorem 1 is given by substituting these final estimates of the hyper parameters into the expressions in Theorem 1 and Section 2.1.

2.4 Nonparametric Estimation of the Marginal

If we misspecified a density function for the true prior γ, then the Bayes classification procedure in Theorem 1 usually does not perform well in practice. This will be illustrated in the simulation studies. To eliminate the misspecification effect from the parametric prior, we will consider a nonparametric approach which adapts well to the data. Note that in the Bayesian classification procedure, the prior γ plays a role only through the marginal $g(z_i)$. Instead of assuming any prior for γ, we will estimate the marginal directly by a nonparametric method.

If p_i were known we would proceed as follows. We first introduce independent Bernoulli random variables $\{\Delta_i\}_{i=1}^p$ where $\Delta_i = 1$ if $\beta_i = 0$ and $\Delta_i = 0$ if $\beta_i \neq 0$. Our nonparametric Bayes estimate of $g(z)$ is given as

$$\widehat{g}(z) = \frac{1}{Nh}\sum_{j=1}^{p}(1-\Delta_j)K(\frac{z-z_j}{h}) \tag{17}$$

where K is a prespecified kernel function which satisfies $\int K(x)dx = 1$, h is the bandwidth of the kernel and N is the total number of non zero β_is. In our procedure, we will use a standard normal density for the kernel function and consider $h = O(p^{-1/5})$ for the bandwidth. More details about kernel density estimation are referred to Wand & Jones (1995).

Since $\{\Delta_i\}_{i=1}^p$ are unknown, we will apply EM algorithm (Dempster, Laird & Rubin 1977) to iteratively estimate w and Δ_is. Consider the complete log-marginal likelihood if we know the missing data $\boldsymbol{\Delta} = (\Delta_1, \cdots, \Delta_p)$, then we have

$$\log m(\boldsymbol{z}, \boldsymbol{\Delta}|w, g) = \sum_{i=1}^{p}\log\Big[\Delta_i w N(z_i|0,1) + (1-\Delta_i)(1-w)g(z_i)\Big]. \tag{18}$$

Our EM algorithm consists of the following two major steps:

Step 1 Given initial values for w and g, calculate the conditional expectation of the log-marginal (18) with respect to the posterior distribution of $\boldsymbol{\Delta}$ given z, w and g. The result is given as follows:

$$E[\log m(\boldsymbol{z}, \boldsymbol{\Delta}|w, g)] = \sum_{i=1}^{p}[p_i \log wN(z_i|0,1) + (1-p_i)\log(1-w)g(z_i)] \tag{19}$$

where p_i is defined in (11). Find w which maximizes (19) by taking the partial derivative of (19) with respect to w and setting it equal to zero. It is easy to obtain the estimate of this hyper parameter as

$$\widehat{w} = \frac{\sum_{i=1}^{p} p_i}{p}. \tag{20}$$

Step 2 Correspondingly the nonparametric estimate of g in (17) will be updated as

$$\widehat{g}(z_i) = \frac{1}{\widetilde{p}h}\sum_{j=1}^{p}(1-\widehat{p}_j)K(\frac{z_i - z_j}{h}) \tag{21}$$

where $\widetilde{p} = \sum_{j=1}^{p}(1-\widehat{p}_j)$.

Final Step In (20) and (21), p_i will be given from (11) based on the estimate of w and $g(z)$ from the last iteration. We will repeat Step 1 and Step 2 until convergence. Suppose the final estimates of p_i is given as $\widehat{p_i}$, then in the Bayes classification procedure from Theorem 1, $g'(z_i)$ and $g''(z_i)$ are estimated as

$$\begin{aligned}
\widehat{g}'(z_i) &= \frac{1}{\widetilde{p}h^2}\sum_{j=1}^{p}(1-\widehat{p}_j)(-\frac{z_i - z_j}{h})K(\frac{z_i - z_j}{h}), \\
\widehat{g}''(z_i) &= \frac{1}{\widetilde{p}h^3}\sum_{j=1}^{p}(1-\widehat{p}_j)((\frac{z_i - z_j}{h})^2 - 1)K(\frac{z_i - z_j}{h}).
\end{aligned}$$

Nonparametric density estimation through kernel functions has a long history in statistics, and the density derivative estimation has also been theoretically studied. Consider a kernel estimator

$$\widehat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x_i - x}{h}).$$

Hansen (2009) has shown the bias and the variance of the derivative of $\widehat{f}(x)$. Let r be either 1 or 2, then the bias is

$$Bias\big(\widehat{f^{(r)}}(x)\big) = \frac{1}{2}f^{(r+2)}(x)h^2\kappa_2(K) + o(h^2),$$

where $\kappa_2(K) = \int_{-\infty}^{\infty} u^2K(u)du$ is the second moment of the kernel, and $f^{(r)}$ is the rth derivative of a function f. The variance is

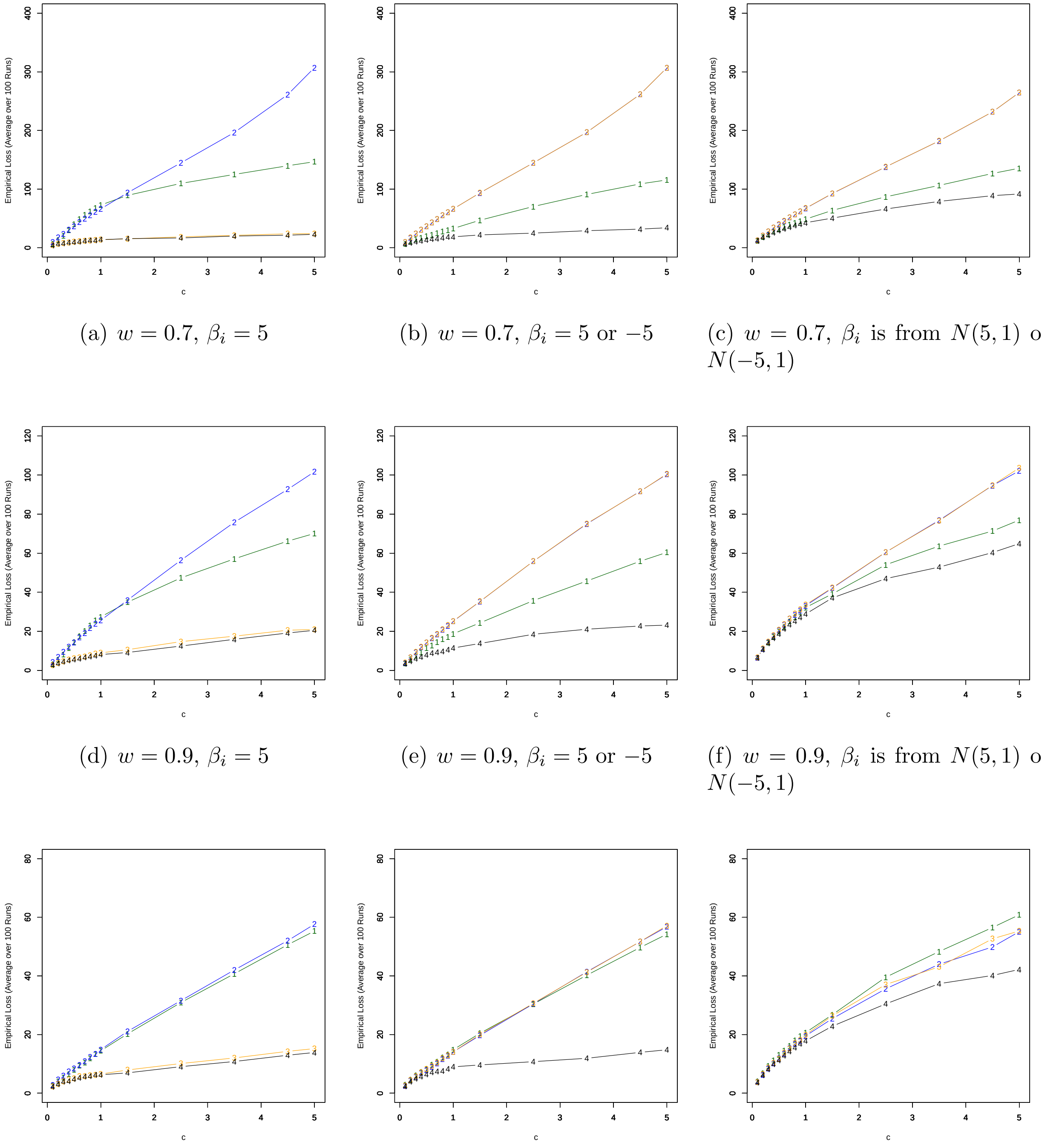
$$Var\big(\widehat{f}^{(r)}(x)\big) = \frac{f(x)}{nh^{1+2r}} \int_{-\infty}^{\infty} K^{(r)}(u)^2 du + O\big(\frac{1}{n}\big).$$

Our proposed Bayesian classifier may be influenced by the accuracy of the density derivative estimation. More theoretical properties is beyond the scope of the current paper, and the performance of our classifier will be evaluated through simulation studies.

3 Simulation Studies

For each simulation run, we generate 100 samples of 500 observations each from a model of the form in equation (1) and come up with decision vectors **a** using different classification rules. We then compare the performance of the two proposed classification methods in terms of the average of the loss in expression (2). To test the classification rules under various conditions, each simulated set of 100 samples comes from a model with varying sparsity, and generating distribution for the non-zero β_i's.

Figure 1 shows some representative plots which compare the average total loss for our nonparametric classifier and the normal prior competitor under different sparsity and signal distribution conditions when the signal is relatively strong. For this figure, the non-zero β_i's are generated from a mixture of $N(5,1)$ and $N(-5,1)$ distributions, from a unit mass at the value 5, or from a mixture of a unit mass at the value 5 and the value -5. The proportion $1-\omega$ of non-zero β_is is set to $0.05, 0.1$, and 0.3. The classifiers are compared at various values of the cost constant c, which corresponds to the relative cost placed on false negative results when a signal is mistakenly classified as noise. Our nonparametric classifier outperforms the parametric competitor (i.e., has lower average loss) for broad ranges of c values in each case. 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) $w = 0.7$, $\beta_i = 5$

(b) $w = 0.7$, $\beta_i = 5$ or -5

(c) $w = 0.7$, β_i is from $N(5,1)$ or $N(-5,1)$

(d) $w = 0.9$, $\beta_i = 5$

(e) $w = 0.9$, $\beta_i = 5$ or -5

(f) $w = 0.9$, β_i is from $N(5,1)$ or $N(-5,1)$

(g) $w = 0.95$, $\beta_i = 5$

(h) $w = 0.95$, $\beta_i = 5$ or -5

(i) $w = 0.95$, β_i is from $N(5,1)$ or $N(-5,1)$

Figure 1: Average empirical total loss over 100 samples of 500 observations. Line 1: non-parametric Bayes classifier; Line 2: parametric classifier with prior mean=0 and variance=1; Line 3: parametric classifier with estimated prior mean and variance; Line 4: classifier using true prior function. The samples are drawn from a model of the form in equation (1) with varying signal sparsity (w) and varying underlying distributions of non-zero β_i's. The two classification rules are compared at different values of the cost constant c.

4 Real Data Analysis

In this section, the nonparametric prior Bayesian classifier is compared with two parametric competitors using a publicly available gene expression data set described in Efron (2009). This microarray data set consists of gene expression measurements for 6,033 genes from 102 people, 52 of whom have been diagnosed with prostate cancer. The goal is to identify genes which are linked to the disease. Following Efron (2009), prior to applying the classification procedures, the 6,033 by 102 raw data matrix is reduced to 6,033 summary statistics. For each of the 6,033 genes, a two-sample t test statistic is constructed using observations on the two groups of study participants. They have the form

$$t_i = c_0 \frac{\bar{x}_{1i} - \bar{x}_{2i}}{\sigma_i}, \tag{22}$$

where $c_0 = \sqrt{n_1 n_2 / n}$, $n_1 = 50$, $n_2 = 52$, x_{1i} and x_{2i} are the average expression levels for gene i among the people in the two groups, and σ_i is the pooled standard deviation of the expression levels of gene i. The t-scores are transformed into standard normal quantiles z_i using the transformation

$$z_i = \Phi^{-1}(F_{n-2}(t_i)), \tag{23}$$

where F_{n-2} is the cumulative distribution function of the t distribution with $n-2$ degrees of freedom, and Φ is the standard normal cumuluative distribution function. A histogram of the data is presented in Figure 2.

The nonparametric prior procedure and two parametric classification procedures were applied to the z_i values to classify the 6,033 genes into two classes: “not associated” and “associated” with disease. Using each procedure, three sets of estimates for the probability that each gene is not associated with the disease were obtained. These are presented in Figure 3. The decision rule for each procedure is obtained by comparing these probability estimates to estimates of the threshold from Theorem 1 & 2. Genes for which the probability estimates fall below the threshold are classified as associated with the disease, and the other genes are classified as not associated. The probability estimates from the nonparametric approach are represented by the top, dark green curve. The thin bottom curves correspond to the estimates from the parametric approaches. Note that the results from the parametric procedures are almost identical. This is not surprising since the only difference between the two procedures is that the mean of the normal prior is set to zero for one of them and estimated from the data for the other, and the data is roughly centered around zero. The nonparametric approach gives much higher estimates of the probabilities that the genes with expression levels close to zero are not associated with the disease. These higher values seem reasonable since one probably does not expect to find many genes associated with disease for which the level of differential expression is very low. The classification results for the nonparametric procedure can also be quite different from those from the parametric approaches. For example, when the cost constant $c = 5$, while 1095 genes are classified as associated with disease using the nonparametric procedure, the parametric procedure with prior mean set equal to zero classifies 983 genes as associated with disease; the other parametric procedure gives almost identical results and classifies 984 genes in this way. At the same time, while the absolute numbers of genes differ greatly, the decision rule cut-offs are not far apart. With the nonparametric approach, genes with z_i values below -1.470 and above 1.449 are classified as signal. For the parametric approach with prior mean set equal to zero, the cut-offs are -1.53 and 1.53. For the other parametric approach, the cut-offs are -1.520 and 1.546. This shows that even small shifts in the estimated cut-off points can produce large changes in the number of genes classified as signal when there are many observations with close values.

5 Conclusion

In this paper, we propose two types of Bayesian classifiers in the context of a highly interpretable loss function. While the parametric Bayes classifier is conceptually simpler, the nonparametric rule outperforms it in terms of the risk function. 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

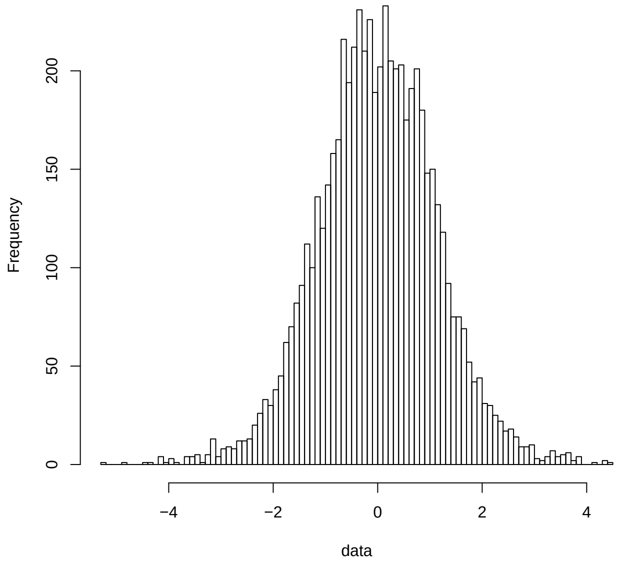


Figure 2: Histogram of 6,033 summary statistics z_i based on the 6,033 by 102 matrix of gene expression data

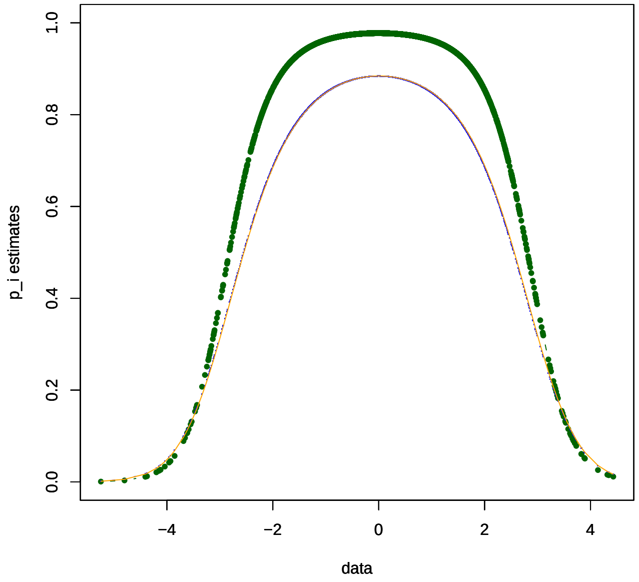


Figure 3: Estimates of probability p_i that the ith gene is not associated with disease. Top curve: nonparametric approach. Overlapping bottom curves: parametric approaches.

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 the two procedures using a publicly available gene expression data set. It is seen that, while the decisions produced by the two rules can be similar, they can also vary greatly for reasonable values of the cost constant c. For the gene expression application in this paper, 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.

REFERENCES

Brown, L.D. (1971). Admissible Estimators, Recurrent Diffusions, and Insoluble Boundary Value Problems. *Annals of Mathematical Statistics*, 42(3), 855-903.

Benjamini, Y. and Hochberg, Y. (1997). Multiple Hypotheses Testing with Weights. *Scandinavian Journal of Statistics*, 24, 407-418.

Brown, L.D. (1986). Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory. *Lecture Notes-Monograph Series*, Vol 9, Institute of Mathematical Statistics, Hayward, California.

Brown, L.D. and Greenshtein, E. (2009). Nonparametric Empirical Bayes and Compound Decision Approaches to Estimation of a High Dimensional Vector of Normal Means. *Annals of Statistics*, 37: 1685-1704.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society: Series B*, 39(1): 1-38.

Efron, B. (2007). Size, Power and False Discovery Rates. *Annals of Statistics*, 35(4): 1351-1377.

Efron, B. (2009). Empirical Bayes Estimates for Large-Scale Prediction Problems. *Journal of the American Statistical Association*, 104(487), 1015-1028.

Efron, B. and Tibshirani, R. (2007). On testing the significance of sets of genes. *Annals of Applied Statistics*, 1(1): 107-129.

George, E. I. and Foster, D. P. (2000). Calibration and Empirical Bayes Variable Selection. *Biometrika*, 87, 731-747.

Hansen, B. (2009). Lecture Notes on Nonparametrics. *Online Manuscript.*

James, W. and Stein, C. (1961). Estimation with Quadratic Loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 361-379.

Johnstone, I. M. and Silverman, B. W. (2004). Needles and Straw in Haystacks: Empirical Bayes Estimates of Possibly Sparse Sequences. *Annals of Statistics*, 32(4): 1594-1649.

Muller, P., Parmigiani, G. and Rice, K. (2006). FDR and Bayesian Multiple Comparisons Rules. *Proc. Valencia / ISBA 8TH World Meeting on Bayesian Statistics.*

Raykar, V. and Zhao, L. (2010). Nonparametric Prior for Adaptive Sparsity. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, JMLR: 629-636.

Robbins, H. (1956). An Empirical Bayes Approach to Statistics. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol 1, 157-163.

Scott, J. and Berger, J. (2006). An Exploration of Aspects of Bayesian Multiple Testing. *Journal of Statistical Planning and Inference*, 136:7, 2144-2162.

Sun, W. and Cai, T. (2007). Oracle and Adaptive Compound Decision Rules for False Discovery Rate Control. *Journal of the American Statistical Association*, 102:479, 901-912.

Wand, M. and Jones, M. (1995). Kernel Smoothing. *Chapman and Hall.*